De novo transcriptome analysis of white teak (*Gmelina arborea* Roxb) wood reveals critical genes involved in xylem development and secondary metabolism.

Mary Luz Yaya Lancheros[1], Krishan Mohan Rai[2], Vimal Kumar Balasubramanian[2], Lavanya Dampanaboina[2], Venugopal Mendu[2], Wilson Terán[1#]

[1]Department of Biology, Pontificia Universidad Javeriana Bogotá, Colombia

[2]Fiber and Biopolymer Research Institute, Department of Plant and soil sciences, Texas Tech University, TX, 79409, USA

[#]Corresponding author e- mail: wteran@javeriana.edu.co

Abstract

Background: *Gmelina arborea* Roxb is a fast-growing tree species of commercial importance for tropical countries due to multiple industrial uses of its wood. Wood is primarily composed of thick secondary cell walls of xylem cells which imparts the strength to the wood. Identification of the genes involved in the secondary cell wall biosynthesis as well as their cognate regulators is crucial to understand how the production of wood occurs and serves as a starting point for developing breeding strategies to produce varieties with improved wood quality, better paper pulping or new potential uses such as biofuel production.

In order to gain knowledge on the molecular mechanisms and gene regulation related with wood development in white teak, a *de novo* sequencing and transcriptome assembly approach was used employing secondary cell wall synthesizing cells from young white teak trees.

Results: For generation of transcriptome**,** RNAseq reads were assembled into 110992 transcripts and 49364 genes were functionally annotated using plant databases; 5071 GO terms and 25460 SSR markers were identified within xylem transcripts and 10256 unigenes were assigned to KEGG database in 130 pathways. Among transcription factor families, C2H2, C3H, bLHLH and MYB were the most represented in xylem. Differential gene expression analysis using leaves as a reference was carried out and a total of 20954 differentially expressed genes were identified including monolignol biosynthetic pathway genes. The differential expression of selected genes (*4CL*, *COMT*, *CCoAOMT*, *CCR* and *NST1*) was validated using qPCR.

Conclusions: We report the very first *de novo* transcriptome of xylem-related genes in this tropical timber species of commercial importance and constitutes a valuable extension of the publicly available transcriptomic resource aimed at fostering both basic and breeding studies.

Keywords: RNA-seq, xylem, differential gene expression, wood development.

BACKGROUND

Tree wood is considered as a sustainable alternative source for biofuel production [1] in addition to its current use in paper and pulp industries. Manipulation of woody biomass for various applications requires extensive knowledge of the pathways involved in the wood production [2, 3]. In rice, for instance, edition of a CAD (cinnamyl alcohol dehydrogenase) encoding gene using CRISPR-CAS (Clustered Regularly Interspaced Short Palindromic Repeats- CRISPR Associated). technology, altered cell wall compostion, reducing lignin content and increasing both cellulose and hemicellulose, which enchanced significantly the saccharification process [4]. A similar result was achieved in poplar, finding that a reduction in lignin biosynthesis led to an improvement of the biomass quality with higher saccharification efficiency[5]. *Gmelina arborea* Roxb. (white teak, Malay beechwood, Kashmir tree, gamari or yemane) is a fast-growing tree species belonging to the lamiaceae family, with tremendous economic importance in several tropical and subtropical areas of southeastern Asia, Africa and America. Its introduction and excellent adaptation to the American tropics (Costa Rica, Venezuela, Colombia and Guatemala) is due to the traits like fast growth, high biomass production (20-25 m^3/ha/year), less susceptibility to the local pests and high yields in addition to the versatility of its wood use which allow a faster investment

return [6]. Therefore, it is considered as a species of choice for both reforestation programs and agroforestry systems in these areas [6,7].

White teak has also shown natural tolerance to water stress and resistance to fire, both characteristics of high interest in the context of climate change. This species has been considered as a tree with higher bioenergetics production, generating an average of 265 m^3 of biomass/hectare/year [8]. White teak fruits and seed present interesting potential as sources of oil for biodiesel production whereas its lignocellulosic wastes serves as a source of bioethanol [810]. Wood is primarily composed of vascular cambium in the woody plants and is composed mainly by secondary xylem. Xylem allows water transport through the stem as well as the tree branches in addition to providing structural support [11].

Formation of wood xylem cells involves two basic processes occurring simultaneously i.e. formation of the secondary cell-wall and programmed cell death [11]. The secondary cell wall is mainly composed of cellulose, hemicellulose and lignin polymers in various proportions [12]. Cellulose is a linear polymer of beta 1-4 linked glucan units that forms microfibers structures which interacts with complex polymers collectively called hemicelluloses in order to form a reticulated matrix [13]. Lignin is a polyphenolic compound which is hydrophobic in nature filling the spaces between celluloses and hemicelluloses fibers and conferring additional mechanical support, rigidity and hydrophobicity [14 15] . After cellulose, lignin is the second most abundant polymer produced by plants, representing approximately 30% of the organic carbon in the biosphere [16]. Lignin polymers are produced from the hydroxycinnamyl alcohol (monolignol) pathway derived from phenylpropanoid pathway, which is also a source of other compounds such as flavonoids,

coumarins, phytoalexins and lignans that are important for plant defense against biotic stressors and commercial biomolecule production [17, 18]. Lignin plays a significant role in the growth and development of woody species which adds the required strength to grow upright and withstand against the mechanical pressure [15].

Lignin biosynthetic pathway involves eleven enzymes in order to produce three monolignols; *p*-coumaryl alcohol, sinapyl alcohol and coniferyl alcohol [19]. Polymerization of these monolignols produces the three types of lignin units, Hydroxyphenyl lignin (H-lignin), Syringyl lignin (S-lignin) and Guaiacyl lignin (G-lignin) and the type of lignin varies based on the species, tissue type and stage of development [12]. The gymnosperm lignin is mainly composed of H and G units, while angiosperms lignin from monocots is composed of H, G and S units whereas in dicots it is composed of G and S units [20, 21].

Various transcription factors have been identified and characterized as key players of wood development, primarily members of NAC and MYB families involved in the regulation of monolignol pathway and lignin polymerization [19, 22, 23]. The NAC family, the transcription factors SND1, NST1, VND6 and VND7 have been recognized as master switches involved in activation of cascade of transcription factors, converging ultimately into secondary xylem formation and lignification [23, 24], . The MYB family transcription factors appears to directly regulate the lignin biosynthetic as well as other cell wall biosynthetic genes. These MYB transcription factors recognize specific DNA sequence motifs on the promoter or regulatory regions of target genes and thereby activating or repressing transcriptional expression [23-25].

The monolignol pathway has been mainly studied in model plant species such as Arabidopsis and poplar [26, 27]. The knowledge generated from these species, has been used to modify tree species such as poplar and eucalyptus in order to reduce the lignin content [28-30]. Although white teak woody biomass presents a high potential for novel uses, lack of knowledge on metabolic and regulatory genes involved in wood development and lignin biosynthesis impairs it use for biofuel applications. A comprehensive knowledge on lignification pathways and its regulation is essential for the improvement of commercially important traits such as wood quality, paper pulping or biofuel production. Therefore, in the present study we have generated *de novo* xylem transcriptome and analyzed and identified xylem specific metabolic and regulatory genes which serves as target genes for future breeding developments in this species.

RESULTS

Generation and annotation of *de novo* reference xylem transcriptome

RNAseq of *G. arborea* xylem library resulted in approximately 165 million paired reads. Quality filtration for the low-quality reads (Q<20) and contaminants such as reads of ribosomal and organellar origin resulted in the removal of total of 18968 paired sequences. The cleaned reads were assembled using Trinity software to obtain the reference transcriptome with 110992 transcripts. The assembled transcripts showed a considerably higher N50 value of 1466 bases with the average transcript length of 864 bases (Table 1). Various publicly available tools and databases were used to annotation these assembled *G. arborea* transcripts. A more popular and conventional homology-based annotation with

NCBI NR database resulted in 49364 hits whereas using model plant *Arabidopsis thaliana* TAIR10 protein database resulted in 45377 hits representing 15445 unigenes. A higher percentage of transcripts with functional annotation was obtained with HMMER analysis: 64186 transcripts presented hits with PFAM database. Fig. 1 represents the main Gene ontology (GO) categories assigned for 14155 unigenes. At the level of cellular component, most of the transcripts were located in the category of organelle whereas at the level of molecular function, the binding and catalytic function categories were the most representative. Cellular and metabolic process were among the most significant biological processes, as well as some categories probably related with dynamic activity in xylem tissue like cell biogenesis, and development processes.

Table 1. Summary of assembly and annotation metrics of the reference transcriptome obtained from *G. arborea* secondary xylem.

Assembly	
Total number of sequences obtained	164,737,322
Number of sequences used for the assembly	164,718,354
Number of transcripts obtained post assembly	110,992
N50 value (in bp)	1466
Average contig length (in bp)	864
Putative gene number	81,269
Number of bases assembled	~95 M
Annotation	
Full length ORFs	17,809 (16%)
Quasi full length ORFs	14,017 (12.6%)
Transcripts with hits in the NCBI NR database (BLASTX)	49,364
Transcripts with hits in TAIR10 (BLASTX)	45,377
Transcripts with hits in *Populus trichocarpa* database	46,795

Transcripts with hits in the NCBI NR base (BLASTX)	45,708
Transcripts with PFAM domains	64,186
Transcripts classified in gene families	48,322
Transcripts with GO terms	39,465
Number of GO terms	5701
Number of KEGG pathways identified	130
Number of genes associated to KEGG pathways	10,256

142

143

144

145 Using the KEGG (kyoto encyclopedia of genes and genomes) database, 10256 genes were
146 assigned to 130 metabolic pathways (Table 1 and 2). Biosynthesis of secondary
147 metabolites, ribosomes and transduction of hormonal signals were the pathways with
148 highest number of associated genes. Phenylpropanoid biosynthesis was also in the top 20 of
149 the most representative pathways (Table 2).

150

151 **Table 2**. Top 20 KEGG pathways identified in the *G. arborea* xylem transcriptome.

Pathways identified	Number of genes
Metabolic pathways	1841
Biosynthesis of secondary metabolites	1020
Ribosome	346
Transduction of signals of plant hormones	262
Carbon metabolism	256
Aminoacid biosynthesis	251
Protein processing in endoplasmic reticulum	217
starch and sucrose metabolism	194

Spliceosome	189
RNA transport	165
Purine metabolism	156
Plant-pathogen interaction	154
Phenylpropanoid biosynthesis	152
Oxidative phosphorylation	149
Ubiquitin mediated proteolysis	148
Endocytosis	140
Amino sugar and nucleotide sugar metabolism	135
Glycolysis / Gluconeogenesis	113
Pyrimidine metabolism	112
Cysteine and methionine metabolism	112

152

153 **Identification of transcription factors, metabolic and regulatory genes involved in the**
154 **monolignol pathway**

155 The main families of transcription factors identified in the reference transcriptome are
156 presented in Fig 2. 101 unigenes were assigned to *C2H2*, 92 to *C3H*, 79 to *bHLH* and 72 to
157 *MYB* TF families; whereas 240 genes were assigned to the AP2-EREBP (56 genes),
158 Homeobox (54 genes), *NAC* (45 genes), *WRKY* (43 genes) and *bZIP* (42 genes) TF families.
159 Nine biosynthetic genes of the monolignol pathway and transcription factors of different
160 levels of regulation were identified from the reference transcriptome.

161 Among the NAC transcription factors, putative orthologs of Arabidopsis *VND7*, *SND2* and
162 *NST1*, reported as "master" regulators, were identified. In the case of MYB transcription
163 factors, *MYB46* and *MYB83*, which were classified as regulators of second level, and *MYB20*,
164 *MYB69* and *MYB85* which are directly related with the activation of biosynthetic genes, were

identified. Other important transcription factor encoding genes were found like *MYB7*, *MYB4*, *MYB32* and *KNAT7*, all reported as negative regulators, or *BES1* a specific activator of the synthesis of celluloses (Table 3). In order to clarify the relation and identity of NAC transcription factors identified as *VND7*, *SND2* and *NST1*, a phylogenetic analysis using possible orthologs from other species was performed (Fig 3).

Table 3. Genes related with lignin biosynthesis and its regulation, identified in the reference transcriptome.

Group	Identified genes
Monolignol pathway genes	Phenylalanine ammonia-lyase (*PAL)* [EC:4.3.1.24]
	Cinnamyl alcohol dehydrogenase *(CAD)* [EC:1.1.1.195]
	Ferulate 5-hydroxylase *(F5H)* [EC:1.14.-.-]
	Hydroxycinnamoyl-CoA reductase *(CCR)* [EC:1.2.1.44]
	Caffeic acid O-methyltransferase *(COMT)* [EC:2.1.1.68]
	4-coumarate-CoA ligase *(4CL)* [EC:6.2.1.12]
	p-hydro-xycinnamoyl-CoA *(HCT)* [EC:2.3.1.133]
	caffeoyl-CoA O-methyltransferase (*CCoAOMT)* [EC:2.1.1.104]
	p-coumarate 3-hydroxylase *(C3´H)* [EC:1.14.13.36]
MYB transcription factors	*MYB46*
	MYB61
	MYB83
	MYB103
	MYB4
	MYB7
	MYB32
	MYB52
	MYB20
	MYB63

	MYB69
	MYB85
NAC transcription factors	*SND2*
	VND7
	NST1
BES1/BZR1 transcription	*BES1*
KNOX transcription	*KNAT7*

In dendrogram (Fig 3), the transcription factor SND2 of white teak was related orthologs from other plant species, while NST1 presented a closer phylogenetic relationship with the NST1 transcription factor of Arabidopsis. In the case of white teak VND7 transcription factor, it was least related with the corresponding orthologs from other species.

Identification of Single Sequence Repeats (SSRs) markers

A total of 25460 SSR markers were identified with 2-5 nucleotides repeat motifs. Among them, the most predominant repetitions were dinucleotides (DNRs, 20634) and trinucleotides (TNRs, 4463) (Supplementary Table 2). In case of the DNRs, AT/AG and TC/TC were the most abundant motifs (33% and 29% respectively). Among TNRs, GAA/AGG (9.9%) and TTC/CCG (7.7%) were the most abundant motifs.

Differential expression analysis

With the goal to perform the differential expression analysis between xylem and leaves, we first generated a unique combined transcriptome using the leaves and xylem reads, since reference genome sequence is not available for *G. arborea*. A total of 196,317,195 sequences

were obtained from leaves; after removal of contaminants and low-quality sequences (about 50 millions of reads), 147,130,884 sequences were obtained. For generation of combined transcriptome, the sequences obtained from leaves were fused with the sequences obtained from xylem. The mapping of reads against this transcriptome indicated an average alignment percentage of 95%, which is indicative of a good representability of expressed transcripts in the transcriptome. Metrics related with the assembly and annotation of this transcriptome are shown in supplementary table 1.

Using this unique transcriptome as reference, the differential expression analysis between leaves and secondary xylem (stem) was performed using leaf tissue as a control. Principal component analysis (PCA) of transcript expression levels revealed a clear differentiation of the samples according to the tissue type (supplementary Fig 1). Results, also indicated that 38,350 transcripts were differentially expressed (adjusted p value < 0.05), out of which 20,964 showed log 2 fold change (Log_2FC) absolute values higher than 2 as a threshold: 9011 transcripts showed an induction pattern whereas 11953 were repressed in xylem compared to leaf tissue (Fig 4). Main functional categories of DEGs are shown in supplementary Fig 2.

To identify overall changes in xylem metabolic pathways encoded by these DEGs, the Mapman tool was used, using the same Log_2FC thresholds values ($|Log_2FC| \geq 2$). Fig 5 presents a general outlook of induction and repression patterns of transcripts involved in main primary and secondary cell metabolism.

As expected, genes involved in photosynthetic light reactions were clearly repressed in xylem compared to leaves, whereas those related to respiration were induced. Accordingly, genes related to cell wall synthesis tend to show an induction pattern in stem compared to leaves. Analysis of nine genes of the monolignol pathway showed a clear differential expression

between leaves and xylem (Fig 6). A general pattern of higher expression was identified for the *PAL*, *C4H*, *COMT* and *CCoAOMT* genes in xylem, while the *HCT* gene was repressed compared to leaves. In the case of *4CL*, *F5H*, *CCR* and *CAD* different transcripts (associated in various cases with possible splicing isoforms) of the same gene presented a higher expression in one or other tissue.

Additionally, transcripts encoding transcription factors belonging to MYB, NAC and homeobox families, were differentially expressed (Fig 7). A clear induction of transcripts annotated as members of MYB family was observed in xylem. In the case of NAC family, several transcripts encoding *NST1* transcription factor, were induced in xylem whereas one *VND7* homolog showed a repression pattern in xylem. Finally, *KNAT7*, a member of the homeobox family, was also induced in xylem tissue. Other genes involved in the development of secondary cell wall also showed differences between leaves and xylem (Fig 8). These genes were further classified into five groups based on their function: cellulose synthesis, hemicellulose synthesis, laccases, programed cell death and others.

Identification of paralogues and their respective splice variants of genes of monolignol pathway

Genes of monolignol pathway contain several variants or paralogues, which may be involved in the same function or have different functions. The reference transcriptome and the differential expression analysis allowed the identification of these paralogues and their possible splicing isoforms for some of the monolignol pathway genes. In the case of *PAL*, two possible paralogues *PAL1* and *PAL4* were identified and both generated different splicing isoforms, all of them upregulated in stem. In the case of *CAD*, possible orthologs of *CAD9* and *CAD3* were identified; the putative *CAD9* paralogue was expressed in both tissues,

239 whereas the *CAD3* was expressed only in stem. Additionally, other two genes, previously
240 not reported, showed a contrasting pattern of expression between tissue: for *4CL*, two
241 transcripts *4CL1* and *4CL2* were identified as possible variants; the last one was induced in
242 leaf, while *4CL1* was mainly induced in stem. Similarly, *CCoAOMT* presented two possible
243 variants *CCoAOMT1* and *CCoAOMT2*. No gene or transcript variants were detected for *C4H*,
244 *COMT*, *F5H*, *CCR*, and *HCT*, as a single transcript was identified.
245
246 **Phylogenetic analysis**
247 In order to determine the phylogenetic relations of some genes of the monolignol pathway
248 identified in white teak with homologous sequences reported for different species, a
249 dendrogram was generated using the protein sequences obtained from *G. arborea PAL* and
250 *CAD* genes with full length ORFs. These genes are the first (*PAL*) and the last (*CAD*) ones
251 to be involved in the monolignol pathway and are key players for the lignin biosynthesis. In
252 the case of *PAL*, one variant induced in stem (putative *PAL1*) was selected, while for *CAD*,
253 two variants were included: one upregulated in stem (called *CADS* and identified as putative
254 *CAD3*) and another one upregulated in leaf (called *CADL*) (Fig 9A and B).
255 In the case of PAL, white teak protein formed a single cluster with another possible ortholog
256 of a Lamiaceae family member, S*cutellaria baicalensi*, and also with *PAL1* of *Coffea*
257 *arabica* (Rubiaceae). For CAD protein, the two evaluated white teak members appeared in
258 different but closely located clusters where CADS was most related with CAD of *Salvia*
259 *miltiorrhiza* and *Sinopodophyllum hexandrum*, while CADL was most related with CAD of
260 *Sesamum indicum* and CAD4 of *Tectona grandis* (teak), two species belonging to lamiales
261 order . CAD1 and CAD4 from *Tectona grandis* (Lamiaceae), a species closely related to *G.*

arborea, were located in distant clusters, indicating a high degree of divergence amongst homologous members of this protein family.

Diiferential expression validation using quantitative reverse transcription PCR (RT-qPCR)

In order to validate the patterns of differential expression observed, a total of 12 genes (10 upregulated and 2 downregulated) were selected for qPCR validation: seven from metabolic genes of the monolignol pathway, two from regulatory genes (transcription factors) and three genes related with synthesis of celluloses and hemicelluloses. For each case, the genes were selected based on the Log_2FC values obtained previously. Comparing the values between the fold change observed in RT-qPCR and the fold change of gene expression obtained by RNAseq, a concordance was found between the values for the *COMT*, *CCR* and *NST1* genes. A similar trend in the expression pattern was found for *CCoAOMT*, *4CL*, *HCT* and *CAD* genes (induced in leaf) (Fig 10) however, no concordance between Log2FC values was found for the *MYB85*, *PAL*, *CESA*, *FRA8* and *PGSIP3* genes. Correlation analysis between the values of Log_2FC of genes with concordant patterns indicated a moderate general correlation coefficient of 0.50.

DISCUSSION

In Colombia, white teak plantations are located mainly in the dry tropical Caribbean zone area, characterized by the presence of a bimodal rainfall pattern, in which the plants are frequently subjected to drought periods that can affect the establishment of new plantations and yields [31]. During water stress conditions, it is common to find that the wood lignification patterns are also modified; these modifications have been related with morphological changes in structures like vessels, necessary to an adequate hydraulic conductivity [32] . However, the molecular mechanisms involved in this type of responses are not very clear yet; therefore, it is important to bring the knowledge about this type of mechanisms, especially in timber species of high importance whose plantations are frequently under stress conditions like white teak**.** There are only a few species such as *Eucalyptus sp* [33], *Populus sp* [34] and *Pinus radiata* [35] with reported transcriptomic data from xylem, probably due to the difficulty in tissue collection. Further, for tropical timber non-model species, genomic information is still scarce except for some species like Acacia [36] and teak [37-39],. Hence, this pioneering study provides information at genomic level associated with development of wood in non-model tropical species like *Gmelina arborea* Roxb.

The xylem transcriptome contains 110,992 transcripts, up to 60% of these could be annotated using different annotation methods (GO, protein domains, BLASTX, KEGG), and also generated a high percentage of transcripts with full length ORFs (16%) and quasi full length ORFs (12,6%). GO annotation revealedbinding and catalysis as main enriched molecular functions In the binding category, genes related to transcription factors predominated,

303 indicating that this function is critical for the development of white teak's xylem., while in
304 the catalysis GO category, the importance of different metabolic processes is reflected in this
305 tissue. One of the most represented category in KEGG pathway was the phenylpropanoid
306 pathway which gives rise to secondary metabolites that are important for different biological
307 processes like pigmentation, UV protection, or responses to pathogens [17]. Additionally,
308 this pathway also produces the monolignols, which are the components of the lignin
309 polymers. Therefore, the results obtained indicate, as expected, a high activity for the
310 pathways involved in the formation of lignin in the developing wood. The *de novo*
311 transcriptome assembly approach used allowed to identify and annotate nine of the ten
312 metabolic genes of the monolignol pathway, which are involved not only in lignin formation
313 but also in other biological processes [40]. Further functional characterization of these
314 individual genes and their variants will provide more information on their biological
315 importance.
316
317 ***Analysis and identification of exonic SSR markers.***
318 Identification of genetic polymorphisms from transcriptomic data, like SSRs markers, , is
319 also relevant for a non-model species as it can be used in future studies for associating
320 genotype/phenotype oriented towards germplasm bank characterization and breeding
321 processes. The analysis of SSRs markers in the white teak xylem-transcriptome indicated a
322 predominance of the dinucleotides AT and AG, which is in accordance with studies in
323 different dicot and gymnosperms species[41]. The xylem transcriptome of white teak showed
324 the GAA/AGG (9.9%) repetitions and TTC/CCG (7.7%) as the most common SSRs. The
325 AGG motif has been reported as highly frequent in monocot species [42], while GAA has
326 been identified mainly in regulatory regions in Arabidopsis [43].. It has been reported that

trinucleotides are less common than dinucleotides; however, their presence in coding regions, may be related to functional polymorphisms while maintaining intact open reading frames.

Analysis of wood and secondary cell wall developmental genes

In order to identify genes more specifically related with the wood development in white teak, the transcriptional profiles of growing trunks (secondary xylem) and leaves from young trees were contrasted. Differential expression analysis evidenced that, in the case of leaves, various transcription factors, predominantly upregulated, were related to leaf development and photomorphogenesis processes such as *KAN* family members that have been related to the abaxial identity [44], *MYB-like* related to foliar senescence [45] and *ELF3* related to development and flowering [46]. In the case of xylem, the significant activation of genes related to development of secondary cell wall was evidenced, which is in accordance with the developmental stage or maturity of the sampled trees. Analysis of the transcription factors involved in the regulation of secondary cell wall biosynthesis showed that *C2H2* and *C3H*, which are involved in the hormonal signal transduction process and different processes of development and response to stress in plants were the most abundant [47, 48] . Further, the *MYB* and *NAC* families, which are involved in different biological processes like response to biotic and abiotic stress, cell cycle control, amongst others [49, 50] were highly represented. These transcription factor families act like "master" regulators at different levels in the secondary cell wall development. Particularly, members of the *NAC* family of transcription factors such as *SND2*, *VND7* and *NST1* act as activators in the third and second level of the regulatory network [51]. The MYB transcription factors act as activators and repressors of

secondary cell wall biosynthetic genes [52, 53]. Interestingly, members of all the above families were represented and upregulated the stem xylem of white teak.

The secondary cell wall master regulator NAC transcription factors showed a general significant pattern of induction in stems was observed for *NST1* and *SND2* genes, whereas the transcript annotated as *VND7* was downregulated. *NST1* is involved mainly in the regulation of development of xylem fibers as has been reported for different species like Arabidopsis and Poplar [15, 54]. In case of *VND7*, although, it has been mainly related to regulation processes in the secondary cell wall formation in vessels [53], its low expression in stem could indicate that its role may be dynamic. This is in agreement with the observation by Mitsuda et al. [54], who affirm that although *NST* and *VND* are similar in their functions, there are some differences in the way in which they act during the formation of the secondary cell wall, being the *NSTs* factors more constant in its expression and *VNDs* more variable. However, it is necessary to validate the identity of this transcription factor, because the phylogenetic analysis was inconclusive. The direct downstream targets of NST1, MYB family of transcription factors such as MYB46, MYB61, MYB83 and MYB103 were significantly induced in stem. These transcription factors are involved in regulating other factors such as *MYB52* and *SND2*, related with the direct regulation of biosynthetic genes of the secondary cell wall [52], as well as MYB family belonging repressors like *MYB4*, *MYB7* and *MYB32*. Other downstream acting MYB factors, directly related with the regulation of the lignin synthesis, were upregulated in stems, such as MYB*20*, *MYB63*, *MYB69* and *MYB85*. Interestingly, the repressor genes *KNAT7* and *MYB4* were also found to be significantly induced in stem, which suggest the presence of negative control feedback loops induced along the development processes of *G. arborea* secondary cell wall.

Analysis of lignin biosynthetic genes

Specifically, the phenylpropanoids pathway showed a clear pattern of upregulation in xylem compared to leaves, as exemplified by *PAL*, *C4H*, *COMT* and *CCoAOMT* genes (Fig 6)However, some variants of biosynthetic genes behave differently. Homologous genes or transcript variants contribute to functional redundancy as well as phenotypic plasticity, where specialization may take place, giving rise to organ or environmental dependent expression. In the case of the *PAL* gene, four variants have been reported in Arabidopsis (*PAL1*, *PAL2*, *PAL3* and *PAL4*) [55], all of them with high importance in the process of lignin biosynthesis. Whereas in tobacco, it has been reported that *PAL2* is more related to processes of development of leaves and flowers as well as pollen viability [40]. In our transcriptomic profiling, unique white teak´s variants for *COMT* and *C4H* were identified and both were significantly upregulated in stem, whereas *F5H* and *4CL* were expressed in both tissues, which does not exclude the possible presence of other variants or multi-functionality of a same variant in other tissues or developmental process.

In the case of CAD enzyme, which catalyzes the last step of the biosynthesis of monolignols for the formation of the alcoholic forms, 9 different members have been reported in Arabidopsis and 12 in rice, some of them with different patterns of expression among different types of tissues [40, 56]. In the xylem of white teak 4 possible variants of *CAD* gene were identified, among which *CAD3* showed a predominant expression in stem and *CAD9* was equally expressed in both tissues, which could indicate a multifunctional role for this gene. *CAD9* has been related mainly to the lignification processes [57], with a gradual induction pattern during stem developmental stage succession [58], although its expression

has also been evidenced in leaves and as part of stress response mechanisms [58-60]. The identity of the other two possible *CAD* members was not determined, however both of them were predominantly expressed in leaves of white teak. Besides, some *Arabidopsis* variants of *CAD* (i.e. *CAD2* and *CAD3*) are poorly or not expressed during lignification processes, thereby indicating probable different roles in other biological processes [40]. . Phylogenetic analysis showed the relationship of two variants of CAD proteins found in white teak,with other possible homologs; the CADs variant (putative CAD3) was tightly related with *Salvia milthiorrhiza* CAD, whereas CADL grouped together with CAD4 of *Tectona grandis* and CAD of *Sesamum indicum*, indicating its possible relation with other members of thelamiales order. However, a more in-depth analysis is necessary to determine the specific identity, ortholog relationship, and biological function of all these members found in white teak.

Differential expression analysis showed that a unique *HCT* gene was significantly upregulated in white teak leaves. According to Besseau et al. [61], under certain conditions, *HCT* may have a key role in the synthesis of flavonoids which may be the case in the leaves of white teak. Xylem expression of HCT, although lower, could be enough to maintain the lignification process.

Biosynthetic genes involved in non-lignin components of secondary cell wall

Development of the xylem cells requires coordinated synthesis of the different elements constituting the secondary cell wall and programmed cell death. Some of the genes involved in these processes showed highly specific expression patterns. This is the case of *IRX* (Irregular xylem) genes, whose mutations affect the phenotypic development at the level of xylem cells [62] as well as *PGSIP* (plant glycogenin-like starch initiation proteins) genes, also known as *Gux*, that constitute a group of genes involved in xylan synthesis and whose

function has been specifically related with secondary wall formation [63]. The *IRXs* genes are involved in synthesis of celluloses and hemicelluloses: *IRX1*, *IRX3* and *IRX5*, for example, are celluloses synthases (CesA) specifically expressed in secondary cell wall [64]. Interestingly, key protease encoding genes such as *XCP1*, *XCP2* and *VPE*, known to be involved in programmed cell death during xylem development, have been identified amongst xylem upregulated genes [65].. Furthermore, concomitant upregulation of transcription factors like VNI2 and XND1, reported as specifically involved in the tight regulation of this process, has also been observed in our transcriptomic profile. [11].

On the other hand,, specific activation of laccase genes such as *LAC4*, *LAC11*, *LAC17*, *LAC10* and *LAC2*, known to be involved in monolignol polymerization [66-68], may reflect the importance of these enzymes for xylem formation,. Finally, upregulation of the *FLA11* gene in xylem is in accordance with previous reports of its induction during the biosynthesis of the secondary cell wall in Arabidopsis and Eucalyptus, where a key role for these fasciclin- like arabinogalactan proteins in cell wall development biomechanics and development has been proposed[64].

Other key genes showed a different pattern of expression like those coding for cellulose synthases (*CESA*), and cellulose synthase-like proteins (*CSL*), for which a significant downregulation in stem was observed. This could be related to fluctuations in the expression of these genes according to the type of cell wall, and the developmental stage. In *Arabidopsis* for example, expression of *CesA1*, *CesA2*, *CesA3*, *CesA5*, *CesA6* and *CesA9* genes was shown to be related to formation of primary cell wall, rather than secondary cell wall [69]. In rice and *Eucalyptus camaldulensis*, differences in the patterns of expression of some *CesA* were found in different types of tissue,cell wall or development stages [70, 71]. In the case

of the *CSL* genes, in white teak most of them presented a predominant expression in leaves. About this, Lerouxel et al.[72], and Muthamilarasan et al. [73], indicate that these proteins have a relevant role in the synthesis of polysaccharides that are not necessarily part of the secondary cell wall hemicellulose matrix, and that environmental factors may affect their expression patterns.

Thus, xylem differentially expressed genes bring molecular knowledge on key functional and anatomical processes seemingly important for white teak's secondary xylem development, like the activation of programmed cell death, the activation of biosynthetic pathways related to lignin formation and other components of the secondary cell wall, or other associated regulatory processes.

CONCLUSIONS

Transcriptomic profiling of leaves and wood of young white teak (*Gmelina arborea* Roxb.) trees was carried out, which constitutes an important genomic resource for this tropical timber. Differential expression analysis allowed to identify for the first time in this species, major genes related with lignin biosynthesis and other components of the secondary cell wall, as well the main transcription factors implicated in its regulation. Also, a catalog of intragenic microsatellite markers was obtained that may be useful in the future establishment of strategies for marker assisted selection of traits related with lignin formation, wood and/or secondary cell wall development in this economically important tree species. The transcriptome obtained could contribute significantly to increase the knowledge on wood and lignin formation that is still scarce in white teak, and will be highly useful for other non-model tropical wood tree species.

METHODS

Plant material and RNA isolation

Plant material was obtained from approximately one-year-old trees, located in the commercial plantation “El Neme”, located at Coello (Tolima, Colombia).of municipality of Coello (Tolima, Colombia). Leaves and stem cuttings from six different individual plants were collected and stored in liquid nitrogen. For RNA isolation from stem (with secondary xylem), external tissues that constitute the bark (phloem and periderm), and pith were removed from stems. Wood was chopped into small pieces using a sterile scalpel and grounded in liquid nitrogen. Total RNA was obtained using the protocol developed for RNA extraction from the pine wood by Chang et al. [74]. The leaf RNA was isolated using the Isolate I RNA isolation kit (Bioline, BIO-52040). RNA samples were quantified using a Nanodrop spectrophotometer (2000, Thermo Scientific, USA) and its integrity was verified using 1% agarose gel electrophoresis in denaturing conditions.

Library preparation and RNAseq

RNA samples with best integrity and concentration values were further validated using a bioanalyzer (2100 Agilent, USA) and samples with a RIN value > 7 were selected for sequencing. Nine RNA samples of each, xylem and leaves, were used to make 3 pools of 3 different individuals for each tissue type. From each pool of RNA, sequencing libraries were generated using the TruSeq library prep kit (Illumina, catalog no. RS-122-210, USA), obtaining six indexed libraries with three replicates for each tissue. All the libraries were sequenced using the NextSeq500 platform (Illumina, USA) to generate paired-end reads of 2 x 150 bases length.

Bioinformatic analysis and *de novo* assembly of reference transcriptomes

Raw reads were evaluated for quality, and sequences with a Q score < 20 were eliminated. Adapters were eliminated by trimming the 10 bp from the 5' ends of the reads using Trimmomatic (version 0.36) [75]. Additionally, the reads corresponding to rRNAs were aligned and eliminated using the program bowtie2 (version 2.3.5) [76] and the SILVA database [77]. Finally, overrepresented sequences identified as contaminants or low complexity sequences were also eliminated from the further processing.

A *de novo* transcriptome assembly strategy was chosen discarding the alternative of reference genome-guided assembly, because the most closely related genome sequence available belongs to a relatively distant member of the Lamiaceae family, and a different genus (*Tectona grandis*). Thus, *de novo* assembly of the reference transcriptome was performed using Trinity (version 2.1.1) [78], setting as parameters a minimum length of 200 bases and a k-mer value of 25. To obtain the reference transcriptome of secondary xylem, only the reads from stem were used in the assembly process. Additionally, a *de novo* assembly using the pooled filtered reads from xylem (stem) and leaves was also performed for the differential expression analysis. All the necessary programs for the computational analysis were run using the High-Performance Computational Center (HPCC) at the Texas Tech University and the ZINE Cluster of Xavierian University.

Annotation of reference transcriptome

Reference transcriptome was annotated using the public databases (TAIR10, NR, and UNIPROT/SWISSPROT) using the BLASTX similarity search program [79] with an e-value of 1e-5. Categories of gene ontology (GO) were assigned using the GO annotation tool in

TAIR [80]. For visualization of GO categories, the system of classification of Wego was used [81]. TAIR annotation was also used for the identification of transcription factors using the AGRIS transcription factors database [82]. KO identifiers necessary for the annotation in KEGG pathways were obtained using the Uniprot tools [83]. PFAM domains were identified using HMMER tools [84]. Additionally, the TRAPID tool was used to perform a quick annotation based in RAPSearch and identify ORFs in the transcripts [85]. For the validation of the identity of some genes with full length ORFs, a multiple alignment–based phylogenetic analysis of their derived protein sequences was performed with selected homologous sequences of plant model and tree species obtained from gene bank, Uniprot, TAIR, PlantTFDB and iTAK plant transcription factor database, using the MEGA 7 software [86] .

SSRs identification

The reference transcriptome was further analyzed for the presence of microsatellite markers using the MISA tool [87], considering a minimum of 5 motif repetition for the dinucleotides (DNRs), trinucleotides (TNRs), tretranucleotides (TtNRs), pentanucleotides (PNRs) and hexanucleotides (HNRs).

Differential expression analysis

For differential expression analysis, the transcriptome assembly generated from the pooled reads of xylem (stem) and leaf tissues were used as reference transcriptome. Reads from each replicate and tissue were mapped against this transcriptome assembly using bowtie2 and samtools [88] and the counts of the mapped sequences were obtained using bedtools [89]. Counts were normalized to RPKMs (Reads Per Kilobase per Million mapped reads) and the differential expression analysis was performed using DESeq package [90] with the leaves

transcripts used as control tissue. A principal component analysis (PCA) of expression levels and using transcripts counts was performed to assess the variance in transcript profiling simultaneously amongst samples (replicates) and treatments (i.e. tissues: xylem and leaves). PCA plot was obtained using ggplot2 R package [91].

Selection of differentially expressed genes between xylem and leaf tissue was done using a binomial test with an adjusted p-value ($p < 0.05$) and values of logarithmic change in fold expression (Log_2 Fold change) ≥ 2 indicating up- or down-regulation of the xylem genes in comparison with leaves. The functional annotation of differentially expressed transcripts was performed using the Mercator [92] and TRAPID tools. Visualization of the key metabolic pathways with differentially expressed genes was performed using the MapMan program [93].

Differential expression validation of genes using RT-qPCR

To validate the differential expression of a selection of genes upregulated in xylem, RT-qPCR was performed. cDNA of xylem (stem) and leaves were prepared using the Transcriptor first strand cDNA synthesis kit (Roche, USA): 1µg of total RNA per 40 µl final reaction volume was used following manufacturer operating procedure. Primers were designed for the selected 13 candidate genes related to the monolignol biosynthetic pathway, cellulose and hemicellulose synthesis, and transcription factors involved in the regulation of secondary cell wall biosynthesis. *UBIQUITIN5* (*UBQ5*), *β-TUBULIN* (*β-TUB*) and *HISTONE3* (*HIS3*) genes were evaluated as reference genes based on the transcriptome data and finally *UBIQUITIN5* (*UBQ5*) was used for the normalization of RT-qPCR data. Primer3 tool was used for the primer designing taking into account the criteria for qPCR primers [94] (Supplementary Table 3).

RT-qPCR was run in a Lightcycler 96 real time PCR (Roche, USA) using the Fast start ™ SYBR green (Roche, USA in a 96-well plate with 3 biological replicates and 3 technical replicates for each gene. Reactions were performed by manufacturer operating procedure in a final volume of 20 µl with 10 µl of SYBR mix, 5 µl of five-fold diluted cDNA (equivalent to 25 ng of reverse transcribed total RNA) and primers at a final concentration of 0.5 pmol/µl. Three negative template controls per primer pair were included in each plate. Running conditions were: a pre incubation phase at 95°C for 10 minutes, 45 cycles of amplification with 3 steps: 95°C for 10 seconds, 58°C for 10 seconds and 72°C for 10 seconds, a melting phase with 3 steps: 95°C for 10 seconds, 65°C for 60 seconds and 97°C for 1 second, finally a cooling phase at 37°C for 30 seconds. Melting curves were analyzed to verify the presence of only one product and the absence of primer dimers. The ΔΔCt comparative method [95] was used for the estimation of the change of gene expression between the two tissues.

Abbreviations

CAD: cinnamyl alcohol dehydrogenase

CRISPR-CAS: Clustered Regularly Interspaced Short Palindromic Repeats- CRISPR Associated.

GO: gene ontology

PAL: Phenylalanine ammonia-lyase

C4H: Trans-cinnamate 4-hydroxylase

4CL:4-coumarate-CoA ligase

HCT: p-hydro-xycinnamoyl-CoA

CCoAOMT: caffeoyl-CoA O-methyltransferase

CCR: hydroxycinnamoyl-CoA reductase

F5H: ferulate 5-hydroxylase

COMT: caffeic acid O-methyltransferase

SSR: Single Sequence Repeats

KEGG: kyoto encyclopedia of genes and genomes

ORF: Open read frame

RPKMs: reads Per Kilobase per Million mapped reads

PCA: Principal component analysis

Log2FC: log 2 fold change

DEGs: Differentially expressed genes

RT-qPCR: quantitative reverse transcription PCR

IRX: Irregular xylem

CesA: celluloses synthases

PGSIP: plant glycogenin-like starch initiation proteins

CSL: cellulose synthase-like proteins

Declarations

Ethics approval and consent to participate:

Not applicable.

Consent for publication:

Not applicable.

Availability of data and materials:

Sequences for comparative phylogenetic analysis were downloaded from Uniprot (https://www.uniprot.org/), NCBI protein (https://www.ncbi.nlm.nih.gov/protein/), TAIR (https://www.arabidopsis.org), PlantTFDB (http://planttfdb.gao-lab.org/) and iTAK plant transcription factor databases (http://itak.feilab.net/cgi-bin/itak/index.cgi). Accession numbers are listed in Figure 3 and Figure 9.

All assembled sequences of this transcriptomic resource have been deposited in the European Nucleotide Archives (ENA) public database under the following accession numbers: PRJEB36634 (ERP119847).

The data supporting the conclusions of this article are included within the article and its supplementary information files.

Competing interests:

The authors declare that they have no competing interests.

Funding:

This work was supported by the Pontifical Xavierian University research grant 00565; MLYL was a recipient of a graduate studies fellowship from Colciencias.

Authors' contributions:

MLYL and WT conceived and designed RNAseq experiments. MLYL carried out all experimental and bioinformatic work and drafted first versions of the manuscript. KMR and VKB assisted on bioinformatic analysis and RT-qPCR validation respectively, and both

assisted on manuscript preparation. WT supervised RNaseq experimental work. LD an VM supervised bioinformatic, RT-qPCR and statistical analysis and were major contributors in polishing the manuscript. WT revised final version of manuscript. All authors read and approved the final manuscript.

Acknowledgements: We thank Suzen, “el mono” and Felipe Ballesteros, owners of “El Neme” natural reserve at Coello (Tolima), for allowing field sampling of white teak trees, and Juliana Vásquez Ardila for her kind assistance during field sampling.

References

1. Hinchee M, Rottmann W, Mullinax L, Zhang C, Chang S, Cunningham M, et al. Short-rotation woody crops for bioenergy and biofuels applications. In Vitro Cell Dev Biol Plant. 2009;45(6):619-29. Epub 2009/11/26. https://doi.org/10.1007/s11627-009-9235-5. PubMed PMID: 19936031; PubMed Central PMCID: PMCPMC2778772.2. Wang JP, Matthews ML, Williams CM, Shi R, Yang C, Tunlaya-anukit S, et al. Improving wood properties for wood utilization through multi-omics integration in lignin biosynthesis. Nat Commun. 2018;9(1579). https://doi.org/10.1038/s41467-018-03863-z.

3. Chanoca A, De Vries L, Boerjan W. Lignin Engineering in Forest Trees. Front Plant Sci. 2019;10. https://doi.org/10.3389/fpls.2019.00912.

4. Zhang G, Wang L, Li X, Bai S, Li Z, Yanting ST, et al. Distinctively altered lignin biosynthesis by site-modification of OsCAD2 for enhanced biomass saccharification in rice. GCB Bioenergy. 2020;13:305–19. https://doi.org/10.1111/gcbb.12772.

679 5. Gui J, Lam PY, Tobimatsu Y, Umezawa T, Li L. Fibre-specific regulation of lignin
680 biosynthesis improves biomass quality in Populus. New Phytol. 2020;226:1074–87.
681 https://doi.org/10.1111/nph.16411
682 6. Dvorak WS. World view of Gmelina arborea : opportunities and challenges. New For.
683 2004;28:111–26.7. Roshetko J.M, Mulawarman, Purnomosidhi P. Gmelina arborea– a
684 viable species for smallholder tree farming in Indonesia?. New For. 2004;28 207–15.
685 https://doi.org/10.1023/B:NEFO.0000040948.53797.c5
686 8. Onyekwelu JC. Managing Short Rotation Tropical Plantations as Sustainable Source of
687 Bioenergy. Silviculture in the Tropics. Tropical Forestry. 2011;8:109-17.
688 https://doi.org/10.1007/978-3-642-19986-8_9.
689 9.Basumaraty S, Deka D, Deka DC. Composition of biodiesel from Gmelina arborea seed
690 oil. Advances in applied science research. 2012;3(5):2745-53.10. Moya R, Tenorio C.
691 Características de combustibilidad de diez especies de plantaciones de rápido crecimiento
692 en Costa Rica. Rev For Mes Kurú. 2013;10(24):26-33.
693 https://doi.org/10.18845/rfmk.v10i24.1321.
694 11. Bollhoner B, Prestele J, Tuominen H. Xylem cell death: emerging understanding of
695 regulation and function. J Exp Bot. 2012;63(3):1081-94. Epub 2012/01/04.
696 https://doi.org/10.1093/jxb/err438. PubMed PMID: 22213814.
697 12. Mendu V, Harman-Ware AE, Crocker M, Jae J, Stork J, Morton S, et al. Identification
698 and thermochemical analysis of high-lignin feedstocks for biofuel and biochemical
699 production. Biotechnol Biofuels. 2011;4(1):43. https://doi.org/10.1186/1754-6834-4-43.
700 13. Somerville C. Cellulose Synthesis in Higher Plants. Ann Rev Cell Dev Biol.
701 2006;22(1):53-78. https://doi.org/10.1146/annurev.cellbio.22.022206.160206. PubMed
702 PMID: 16824006.

14. Ko JH, Kim WC, Han KH. Ectopic expression of MYB46 identifies transcriptional regulatory genes involved in secondary wall biosynthesis in Arabidopsis. Plant J. 2009;60(4):649-65. Epub 2009/08/14. https://doi.org/ 10.1111/j.1365-313X.2009.03989.x. PubMed PMID: 19674407.

15. Zhong R, Ye ZH. Secondary cell walls: biosynthesis, patterned deposition and transcriptional regulation. Plant Cell Physiol. 2015;56(2):195-214. Epub 2014/10/09. https://doi.org/ 10.1093/pcp/pcu140. PubMed PMID: 25294860.

16. Silva-Moura JC, Bonine CA, de Oliveira Fernandes Viana J, Dornelas MC, Mazzafera P. Abiotic and biotic stresses and changes in the lignin content and composition in plants. J Integr Plant Biol. 2010;52(4):360-76. Epub 2010/04/10. https://doi.org/ 10.1111/j.1744-7909.2010.00892.x. PubMed PMID: 20377698.

17. Vogt T. Phenylpropanoid biosynthesis. Mol Plant. 2010;3(1):2-20. Epub 2009/12/26. https://doi.org/ 10.1093/mp/ssp106. PubMed PMID: 20035037.

18. Fraser CM, Chapple C. The phenylpropanoid pathway in Arabidopsis. Arabidopsis Book. 2011;9:e0152. Epub 2012/02/04. https://doi.org/ 10.1199/tab.0152. PubMed PMID: 22303276; PubMed Central PMCID: PMCPMC3268504.

19. Vanholme R, Demedts B, Morreel K, Ralph J, Boerjan W. Lignin Biosynthesis and Structure. Plant Physiol. 2010;153(3):895-905. https://doi.org/ 10.1104/pp.110.155119.

20. Guillaumie S, Mzid R, Mechin V, Leon C, Hichri I, Destrac-Irvine A, et al. The grapevine transcription factor WRKY2 influences the lignin pathway and xylem development in tobacco. Plant Mol Biol. 2010;72(1-2):215-34. Epub 2009/11/11. https://doi.org/ 10.1007/s11103-009-9563-1. PubMed PMID: 19902151.

21. Weng JK, Chapple C. The origin and evolution of lignin biosynthesis. New Phytol. 2010;187(2):273-85. Epub 2010/07/21. https://doi.org/ 10.1111/j.1469-8137.2010.03327.x. PubMed PMID: 20642725.

22. Xie M, Zhang J, Tschaplinski TJ, Tuskan GA, Chen J-G, Muchero W. Regulation of Lignin Biosynthesis and Its Role in Growth-Defense Tradeoffs. Front Plant Sci. 2018;9(1427). https://doi.org/ 10.3389/fpls.2018.01427.

23. Zhong R, Ye Z-H. Transcriptional regulation of lignin biosynthesis. Plant Signal Behav. 2009;4(11):1028-34. https://doi.org/ 10.4161/psb.4.11.9875. PubMed PMID: 19838072.

24. Zhao Q, Dixon RA. Transcriptional networks for lignin biosynthesis: more complex than we thought? Trends Plant Sci. 2011;16(4):227-33. Epub 2011/01/14. https://doi.org/ 10.1016/j.tplants.2010.12.005. PubMed PMID: 21227733.

25. Liu J, Osbourn A, Ma P. MYB Transcription Factors as Regulators of Phenylpropanoid Metabolism in Plants. Mol Plant. 2015;8(5):689-708. Epub 2015/04/04. https://doi.org/ 10.1016/j.molp.2015.03.012. PubMed PMID: 25840349.

26. Van Acker R, Vanholme R, Storme V, Mortimer JC, Dupree P. Lignin biosynthesis perturbations affect secondary cell wall composition and saccharification yield in Arabidopsis thaliana. Biotechnol Biofuels. 2013;6:1–17. https://doi.org/10.1186/1754-6834-6-46.

27. Zhang J, Tuskan GA, Tschaplinski TJ, Muchero W, Chen J-G. Transcriptional and Post-transcriptional Regulation of Lignin Biosynthesis Pathway Genes in Populus. Front Plant Sci. 2020;11 May:1–11. https://doi.org/ 10.3389/fpls.2020.00652.

747 28. Li X, Weng JK, Chapple C. Improvement of biomass through lignin modification.
748 Plant J. 2008;54(4):569-81. Epub 2008/05/15. https://doi.org/ 10.1111/j.1365-
749 313X.2008.03457.x. PubMed PMID: 18476864.
750 29. Wang H, Xue Y, Chen Y, Li R, Wei J. Lignin modification improves the biofuel
751 production potential in transgenic Populus tomentosa. Ind Crop Prod. 2012;37(1):170-7.
752 https://doi.org/ 10.1016/j.indcrop.2011.12.014.
753 30. Ziebell A, Gjersing E, Hinchee M, Katahira R, Sykes RW, Johnson DK, et al.
754 Downregulation of p-Coumaroyl Quinate/Shikimate 3′-Hydroxylase (C3′H) or Cinnamate-
755 4-hydrolylase (C4H) in Eucalyptus urophylla × Eucalyptus grandis Leads to Increased
756 Extractability. BioEnergy Res. 2016;9(2):691-9. https://doi.org/ 10.1007/s12155-016-9713-
757 7.
758 31. Rojas A, Moreno L, Melgarejo LM, Rodríguez M. Physiological response of
759 gmelina (*Gmelina arborea* Roxb) to hydric conditions of the colombian Caribbean.
760 Agronomía colombiana. 2012;30(1):52-8.
761 32. Malavasi UC, Davis AS, Malavasi MDM. Lignin in Woody Plants under Water Stress :
762 A Review. Floresta e Ambient. 2016;23:589–97. https://doi.org/10.1590/2179-
763 8087.143715.
764 33. Pappas M, Pappas GJ, Grattapaglia D. Genome-wide discovery and validation of
765 Eucalyptus small RNAs reveals variable patterns of conservation and diversity across
766 species of Myrtaceae. BMC Genomics. 2015;16:1113. Epub 2015/12/31. https://doi.org/
767 10.1186/s12864-015-2322-6. PubMed PMID: 26714854; PubMed Central PMCID:
768 PMCPMC4696225.
769 34. Hefer CA, Mizrachi E, Myburg AA, Douglas CJ, Mansfield SD. Comparative
770 interrogation of the developing xylem transcriptomes of two wood-forming species:

771 Populus trichocarpa and Eucalyptus grandis. New Phytol. 2015;206(4):1391-405. Epub
772 2015/02/11. https://doi.org/ 10.1111/nph.13277. PubMed PMID: 25659405.
773 35. Li X, Wu HX, Southerton SG. Seasonal reorganization of the xylem transcriptome
774 at different tree ages reveals novel insights into wood formation in Pinus radiata. New
775 Phytol. 2010;187(3):764-76. Epub 2010/06/22. https://doi.org/ 10.1111/j.1469-
776 8137.2010.03333.x. PubMed PMID: 20561208.
777 36. Wong MM, Cannon CH, Wickneswari R. Identification of lignin genes and
778 regulatory sequences involved in secondary cell wall formation in *Acacia auriculiformis*
779 and *Acacia mangium* via de novo transcriptome sequencing. BMC Genomics. 2011;12:342.
780 Epub 2011/07/07. https://doi.org/ 10.1186/1471-2164-12-342. PubMed PMID: 21729267;
781 PubMed Central PMCID: PMCPMC3161972.
782 37. Galeano E, Vasconcelos TS, Vidal M, Mejia-Guerra MK, Carrer H. Large-scale
783 transcriptional profiling of lignified tissues in *Tectona grandis*. BMC Plant Biol.
784 2015;15:221. Epub 2015/09/16. https://doi.org/ 10.1186/s12870-015-0599-x. PubMed
785 PMID: 26369560; PubMed Central PMCID: PMCPMC4570228.
786 38. Yasodha R, Vasudeva R, Balakrishnan S, Sakthi AR, Abel N, Binai N, et al. Draft
787 genome of a high value tropical timber tree , Teak (Tectona grandis L . f): insights into
788 SSR diversity , phylogeny and conservation. DNA Res. 2018;25:409–19. https://doi.org/
789 10.1093/dnares/dsy013.
790 39. Zhao D, Hamilton JP, Bhat WW, Johnson R, Godden GT, Kinser TJ, et al. A
791 chromosomal-scale genome assembly of Tectona grandis reveals the importance of tandem
792 gene duplication and enables discovery of genes in natural product biosynthetic pathways.
793 Gigascience. 2019;8:1–10. https://doi.org/ 10.1093/gigascience/giz005.
794

795 40. Yoon J, Choi H, An G. Roles of lignin biosynthesis and regulatory genes in plant
796 development. J Integr Plant Biol. 2015;57(11):902-12. Epub 2015/08/25. https://doi.org/
797 10.1111/jipb.12422. PubMed PMID: 26297385; PubMed Central PMCID:
798 PMCPMC5111759.
799 41. Ranade S, Lin Y, Zuccolo A, Van de Peer Y, García-Gil M. Comparative in silico
800 analysis of EST-SSR in angiosperm and gymnosperm tree genera. BMC Plant Biology.
801 2014;14(220):1-10.
802 42. Sonah H, Deshmukh RK, Sharma A, Singh VP, Gupta DK, Gacche RN, et al.
803 Genome-wide distribution and organization of microsatellites in plants: an insight into
804 marker development in Brachypodium. PLoS One. 2011;6(6):e21298. Epub 2011/06/30.
805 https://doi.org/ 10.1371/journal.pone.0021298. PubMed PMID: 21713003; PubMed Central
806 PMCID: PMCPMC3119692.
807 43. Zhang L, Zuo K, Zhang F, Cao Y, Wang J, Zhang Y, et al. Conservation of
808 noncoding microsatellites in plants: implication for gene regulation. BMC Genomics.
809 2006;7:323. Epub 2006/12/26. https://doi.org/ 10.1186/1471-2164-7-323. PubMed PMID:
810 17187690; PubMed Central PMCID: PMCPMC1781443.
811 44. Wu G, Lin W, Huang T, Poethig S, Springer P, Kerstetter RA. KANADI1 regulates
812 adaxial-abaxial polarity in Arabidopsis by directly repressing the transcription of
813 assymetric leaves2. PNAS. 2008;105(42):16392-7.
814 https://doi.org/10.1073/pnas.0803997105.
815 45. Zhang X, Ju HW, Chung MS, Huang P, Ahn SJ, Kim CS. The R-R-type MYB-like
816 transcription factor, AtMYBL, is involved in promoting leaf senescence and modulates an
817 abiotic stress response in Arabidopsis. Plant Cell Physiol. 2011;52(1):138-48. Epub
818 2010/11/26. https://doi.org/ 10.1093/pcp/pcq180. PubMed PMID: 21097474.

819 46. Song Y, Yang C, Gao S, Zhang W, Li L, Kuai B. Age-triggered and dark-induced
820 leaf senescence require the bHLH transcription factors PIF3, 4, and 5. Mol Plant.
821 2014;7(12):1776-87. Epub 2014/10/10. https://doi.org/ 10.1093/mp/ssu109.
822 47. Muthamilarasan M, Bonthala VS, Mishra AK, Khandelwal R, Khan Y, Roy R, Prasad
823 M. C2H2 type of zinc finger transcription factors in foxtail millet define response to abiotic
824 stresses. Funct Integr Genomics. 2014;14(3):531-43. https://doi.org/ 10.1007/s10142-014-
825 0383-2. Epub 2014 Jun 11. PMID: 24915771.
826 48. Jiang AL, Xu ZS, Zhao GY. et al. Genome-Wide Analysis of the C3H Zinc Finger
827 Transcription Factor Family and Drought Responses of Members in Aegilops tauschii .
828 Plant Mol Biol Rep. 2014;32:1241–1256. https://doi.org/10.1007/s11105-014-0719-z
829 49. Ambawat S, Sharma P, Yadav NR, Yadav RC. MYB transcription factor genes as
830 regulators for plant responses: an overview. Physiol Mol Biol Plants. 2013;19(3):307-21.
831 Epub 2014/01/17. https://doi.org/ 10.1007/s12298-013-0179-1. PubMed PMID: 24431500;
832 PubMed Central PMCID: PMCPMC3715649.
833 50. Nuruzzaman M, Sharoni AM, Kikuchi S. Roles of NAC transcription factors in the
834 regulation of biotic and abiotic stress responses in plants. Front Microbiol. 2013;4:248.
835 Epub 2013/09/24. https://doi.org/ 10.3389/fmicb.2013.00248. PubMed PMID: 24058359;
836 PubMed Central PMCID: PMCPMC3759801.
837 51. Nakano Y, Yamaguchi M, Endo H, Rejab NA, Ohtani M. NAC-MYB-based
838 transcriptional regulation of secondary cell wall biosynthesis in land plants. Front Plant Sci.
839 2015;6:288. Epub 2015/05/23. https://doi.org/ 10.3389/fpls.2015.00288. PubMed PMID:
840 25999964; PubMed Central PMCID: PMCPMC4419676.
841 52. Hussey SG, Mizrachi E, Creux NM, Myburg AA. Navigating the transcriptional
842 roadmap regulating plant secondary cell wall deposition. Front Plant Sci. 2013;4:325. Epub

2013/09/07. https://doi.org/ 10.3389/fpls.2013.00325. PubMed PMID: 24009617; PubMed Central PMCID: PMCPMC3756741.

53. Yang JH, Wang H. Molecular Mechanisms for Vascular Development and Secondary Cell Wall Formation. Front Plant Sci. 2016;7:356. Epub 2016/04/06. https://doi.org/ 10.3389/fpls.2016.00356. PubMed PMID: 27047525; PubMed Central PMCID: PMCPMC4801872.

54. Mitsuda N, Iwase A, Yamamoto H, Yoshida M, Seki M, Shinozaki K, et al. NAC transcription factors, NST1 and NST3, are key regulators of the formation of secondary walls in woody tissues of Arabidopsis. Plant Cell. 2007;19(1):270-80. Epub 2007/01/24. https://doi.org/ 10.1105/tpc.106.047043. PubMed PMID: 17237351; PubMed Central PMCID: PMCPMC1820955.

55. Saito K, Yonekura-Sakakibara K, Nakabayashi R, Higashi Y, Yamazaki M, Tohge T, et al. The flavonoid biosynthetic pathway in Arabidopsis: structural and genetic diversity. Plant Physiol Biochem. 2013;72:21-34. Epub 2013/03/12. https://doi.org/ 10.1016/j.plaphy.2013.02.001. PubMed PMID: 23473981.

56. Hirano K, Aya K, Kondo M, Okuno A, Morinaka Y, Matsuoka M. OsCAD2 is the major CAD gene responsible for monolignol biosynthesis in rice culm. Plant Cell Reports. 2011;31(1):91-101. https://doi.org/ 10.1007/s00299-011-1142-7.

57. Shen H, Mazarei M, Hisano H, Escamilla-Trevino L, Fu C, Pu Y, et al. A genomics approach to deciphering lignin biosynthesis in switchgrass. Plant Cell. 2013;25(11):4342-61. Epub 2013/11/29. https://doi.org/ 10.1105/tpc.113.118828. PubMed PMID: 24285795; PubMed Central PMCID: PMCPMC3875722.

58. Raes J, Rohde A, Christensen JH, Van de Peer Y, Boerjan W. Genome-wide characterization of the lignification toolbox in Arabidopsis. Plant Physiol.

2003;133(3):1051-71. Epub 2003/11/13. https://doi.org/ 10.1104/pp.103.026484. PubMed PMID: 14612585; PubMed Central PMCID: PMCPMC523881.

59. Bethke G, Pecher P, Eschen-Lippold K, Katagiri F. Activation of the Arabidopsis thaliana mitogen-actived protein kinase MPK11 by the felagellin-derived elicitor peptide, flg22. MPMI. 2012;25(4):471-80.

60. Cheng X, Li M, Li D, Zhang J, Jin Q, Sheng L, et al. Characterization and analysis of CCR and CAD gene families at the whole-genome level for lignin synthesis of stone cells in pear (*Pyrus bretschneideri*) fruit. Biol Open. 2017;6(11):1602-13. Epub 2017/11/17. https://doi.org/ 10.1242/bio.026997. PubMed PMID: 29141952; PubMed Central PMCID: PMCPMC5703608.

61. Besseau S, Hoffmann L, Geoffroy P, Lapierre C, Pollet B, Legrand M. Flavonoid accumulation in Arabidopsis repressed in lignin synthesis affects auxin transport and plant growth. Plant Cell. 2007;19(1):148-62. Epub 2007/01/24. https://doi.org/ 10.1105/tpc.106.044495. PubMed PMID: 17237352; PubMed Central PMCID: PMCPMC1820963.

62. Carpita NC, McCann MC. Characterizing visible and invisible cell wall mutant phenotypes. J Exp Bot. 2015;66(14):4145-63. Epub 2015/04/16. https://doi.org/ 10.1093/jxb/erv090. PubMed PMID: 25873661.

63. Rennie EA, Hansen SF, Baidoo EE, Hadi MZ, Keasling JD, Scheller HV. Three members of the Arabidopsis glycosyltransferase family 8 are xylan glucuronosyltransferases. Plant Physiol. 2012;159(4):1408-17. Epub 2012/06/19. https://doi.org/ 10.1104/pp.112.200964. PubMed PMID: 22706449; PubMed Central PMCID: PMCPMC3428776.

64. MacMillan CP, Mansfield SD, Stachurski ZH, Evans R, Southerton SG. Fasciclin-like arabinogalactan proteins: specialization for stem biomechanics and cell wall architecture in Arabidopsis and Eucalyptus. Plant J. 2010; 62(4):689–703. https://doi.org/10.1111/j.1365-313X.2010.04181.x

65. Zamyatnin AA. Plant Proteases Involved in Regulated Cell Death. Usp Biol Khim. 2015;80:1701–15. https://doi.org/10.1134/S0006297915130064.

66. Wang J, Feng J, Jia W, Chang S, Li S, Li Y. Lignin engineering through laccase modification : a promising field for energy plant improvement. Biotechnol Biofuels. 2015;8:1–11. https://doi.org/10.1186/s13068-015-0331-y.

67. Berthet S, Demont-Caulet N, Pollet B, Bidzinski P, Cezard L, Le Bris P, et al. Disruption of LACCASE4 and 17 results in tissue-specific alterations to lignification of Arabidopsis thaliana stems. Plant Cell. 2011;23(3):1124-37. Epub 2011/03/31. https://doi.org/ 10.1105/tpc.110.082792. PubMed PMID: 21447792; PubMed Central PMCID: PMCPMC3082258.

68. Zhao Q, Nakashima J, Chen F, Yin Y, Fu C, Yun J, et al. Laccase is necessary and nonredundant with peroxidase for lignin polymerization during vascular development in Arabidopsis. Plant Cell. 2013;25(10):3976-87. Epub 2013/10/22. https://doi.org/ 10.1105/tpc.113.117770. PubMed PMID: 24143805; PubMed Central PMCID: PMCPMC3877815.

4143805; PubMed Central PMCID: PMCPMC3877815.

69. Endler A, Persson S. Cellulose synthases and synthesis in Arabidopsis. Mol Plant. 2011;4(2):199-211. Epub 2011/02/11. https://doi.org/ 10.1093/mp/ssq079. PubMed PMID: 21307367.

913 70. Wang L, Guo K, Li Y, Tu Y, Hu H, Wang B, et al. Expression profiling and
914 integrative analysis of the CESA/CSL superfamily in rice. BMC Plant Biol.
915 2010;10(282):1-16.
916 71. Lin Y, Kao Y-Y, Chen Z-Z, Chu F-H, Chung J-D. cDNA cloning and molecular
917 characterization of five cellulose synthase A genes from Eucalyptus camaldulensis. J Plant
918 Biochem Biot. 2013;23(2):199-210. https://doi.org/ 10.1007/s13562-013-0202-1.
919 72. Lerouxel O, Cavalier DM, Liepman AH, Keegstra K. Biosynthesis of plant cell wall
920 polysaccharides - a complex process. Curr Opin Plant Biol. 2006;9(6):621-30. Epub
921 2006/10/03. https://doi.org/ 10.1016/j.pbi.2006.09.009. PubMed PMID: 17011813.
922 73. Muthamilarasan M, Khan Y, Jaishankar J, Shweta S, Lata C, Prasad M. Integrative
923 analysis and expression profiling of secondary cell wall genes in C4 biofuel model *Setaria*
924 *italica* reveals targets for lignocellulose bioengineering. Front Plant Sci. 2015;6:965. Epub
925 2015/11/20. https://doi.org/ 10.3389/fpls.2015.00965. PubMed PMID: 26583030; PubMed
926 Central PMCID: PMCPMC4631826.
927 74. Chang S, Puryear J, Cairney J. A simple and efficient method for isolating RNA
928 from pine trees. Plant Mol Biol Rep. 1993;11(2):113-6. https://doi.org/
929 10.1007/bf02670468.
930 75. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina
931 sequence data. Bioinformatics. 2014;30(15):2114-20. Epub 2014/04/04. https://doi.org/
932 10.1093/bioinformatics/btu170. PubMed PMID: 24695404; PubMed Central PMCID:
933 PMCPMC4103590.
934 76. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat
935 Methods. 2012;9(4):357-9. Epub 2012/03/06. https://doi.org/ 10.1038/nmeth.1923. PubMed
936 PMID: 22388286; PubMed Central PMCID: PMCPMC3322381.

77. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41(Database issue):D590-6. Epub 2012/11/30. https://doi.org/ 10.1093/nar/gks1219. PubMed PMID: 23193283; PubMed Central PMCID: PMCPMC3531112.

78. Grabherr M, Haas B, Yassour M, Levin J, Thompson D, Amit I, et al. Trinity: Reconstructing a full length transcriptome without a genome from RNA-Seq data. Nat Biotechnol. 2013;29(7):644-52. https://doi.org/ 10.1038/nbt.1883.Trinity.

79. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-10. Epub 1990/10/05. https://doi.org/ 10.1016/s0022-2836(05)80360-2. PubMed PMID: 2231712.

80. Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, Miller N, Mueller L, et al. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community , Nucleic Acids Research. 2003;31(1):224–28. https://doi.org/ 10.1093/nar/gkg076.

81. Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, et al. WEGO: a web tool for plotting GO annotations. Nucl Acids Res. 2006;34(Web Server):W293-W7. https://doi.org/ 10.1093/nar/gkl031.

82. Palaniswamy SK, James S, Sun H, Lamb RS, Davuluri RV, Grotewold E. AGRIS and AtRegNet. a platform to link cis-regulatory elements and transcription factors into regulatory networks. Plant Physiol. 2006;140(3):818-29. Epub 2006/03/10. https://doi.org/ 10.1104/pp.105.072280. PubMed PMID: 16524982; PubMed Central PMCID: PMCPMC1400579.

961 83. The UniProt Consortium, UniProt: the universal protein knowledgebase. Nucl Acids
962 Res. 2017;45(D1):D158–69. https://doi.org/ 10.1093/nar/gkw1099.
963 84. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence
964 similarity searching. Nucl Acids Res. 2011;39(Web Server issue):W29–W37.
965 https://doi.org/10.1093/nar/gkr367.
966 85. Van Bel M, Proost S, Neste CV, Deforce D, Van De Peer Y, Vandepoele K.
967 TRAPID : an efficient online tool for the functional and comparative analysis of de novo
968 RNA-Seq transcriptomes. Genome Biol. 2013;14:1-10.
969 86. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics
970 Analysis Version 7.0 for Bigger Datasets, Mol Biol Evol. 2016;33(7):1870–74.
971 https://doi.org/10.1093/molbev/msw054.
972 87. Thiel T, Michalek W, Varshney K, Graner A. Exploiting EST databases for the
973 development and characterization of gene-derived SSR-markers in barley (*Hordeum*
974 *vulgare* L .). Theor Appl Genet. 2003;106:411-22. https://doi.org/ 10.1007/s00122-002-
975 1031-0.
976 88. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
977 Alignment / Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
978 https://doi.org/ 10.1093/bioinformatics/btp352.
979 89. Quinlan AR, Hall IM. BEDTools : a flexible suite of utilities for comparing
980 genomic features. Bioinformatics. 2010;26(6):841-2. https://doi.org/
981 10.1093/bioinformatics/btq033.
982 90. Anders S, Huber W. Differential expression analysis for sequence count data.
983 Genome Biol. 2010;11(R106):1-12.

91. Wickham H. Ggplot2: Elegant Graphics for Data Analysis. 2nd Edition, Springer, New York. 2009. https://doi.org/10.1007/978-0-387-98141-3.

92. Lohse M, Nagel A, Herter T, May P, Schroda M, Zrenner R, et al. Mercator : a fast and simple web server for genome scale functional annotation of plant sequence data. Plant Cell Environ. 2014;37:1250-8. https://doi.org/10.1111/pce.12231.

93. Thimm O, Blasing O, Gibon Y, Nagel A, Meyer S, Kru P, et al. MAPMAN : a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J. 2004;37:914-39. https://doi.org/ 10.1111/j.1365-313X.2004.02016.x.

94. Thornton B, Basu C. Real-time PCR (qPCR) primer design using free online software. Biochem Mol Biol Educ. 2011;39(2):145-54. Epub 2011/03/30. https://doi.org/ 10.1002/bmb.20461. PubMed PMID: 21445907.

95. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods. 2001;25(4):402-8. https://doi.org/10.1006/meth.2001.1262. PMID: 11846609.

Figures.

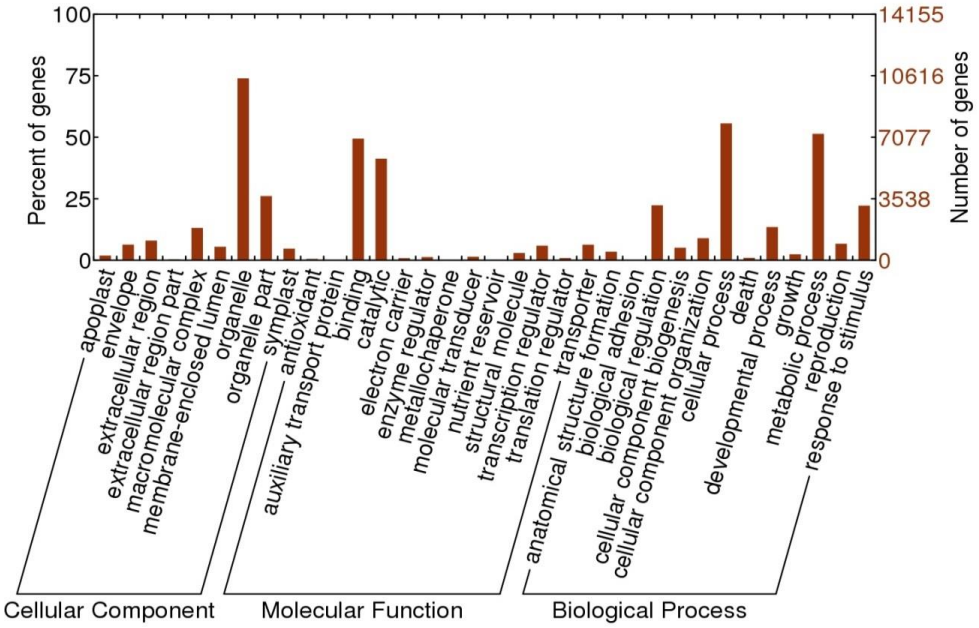
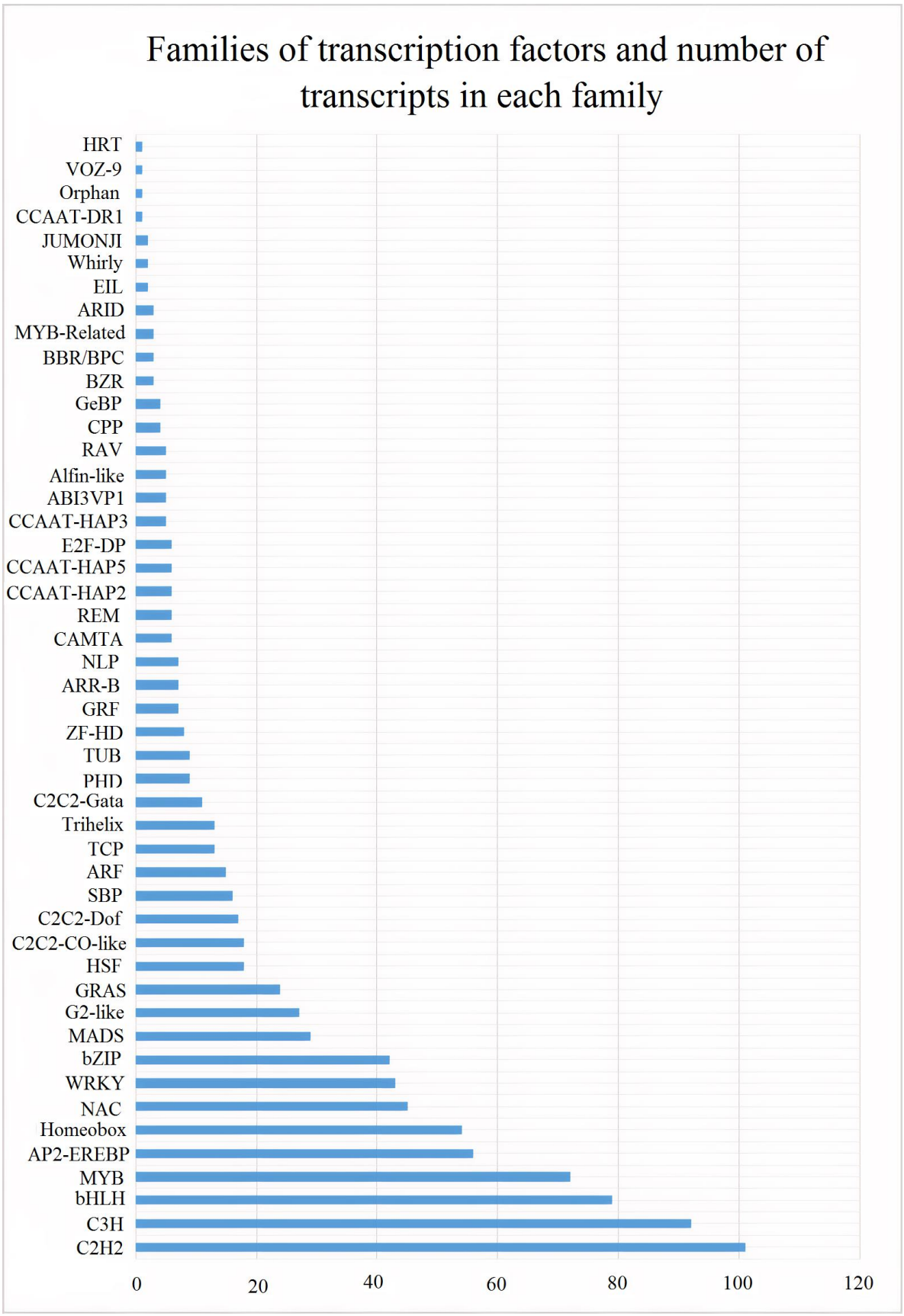


Fig 1. Main GO categories assigned to xylem reference transcriptome of *Gmelina arborea*.


Families of transcription factors and number of transcripts in each family
HRT
VOZ-9
Orphan
CCAAT-DR1
JUMONJI
Whirly
EIL
ARID
MYB-Related
BBR/BPC
BZR
GeBP
CPP
RAV
Alfin-like
ABI3VP1
CCAAT-HAP3
E2F-DP
CCAAT-HAP5
CCAAT-HAP2
REM
CAMTA
NLP
ARR-B
GRF
ZF-HD
TUB
PHD
C2C2-Gata
Trihelix
TCP
ARF
SBP
C2C2-Dof
C2C2-CO-like
HSF
GRAS
G2-like
MADS
bZIP
WRKY
NAC
Homeobox
AP2-EREBP
MYB
bHLH
C3H
C2H2
0
20
40
60
80
100
120

1003

Fig 2. Main families of transcription factors identified in the xylem transcriptome**.** Blue bars indicate the number of transcripts belonging to each family.

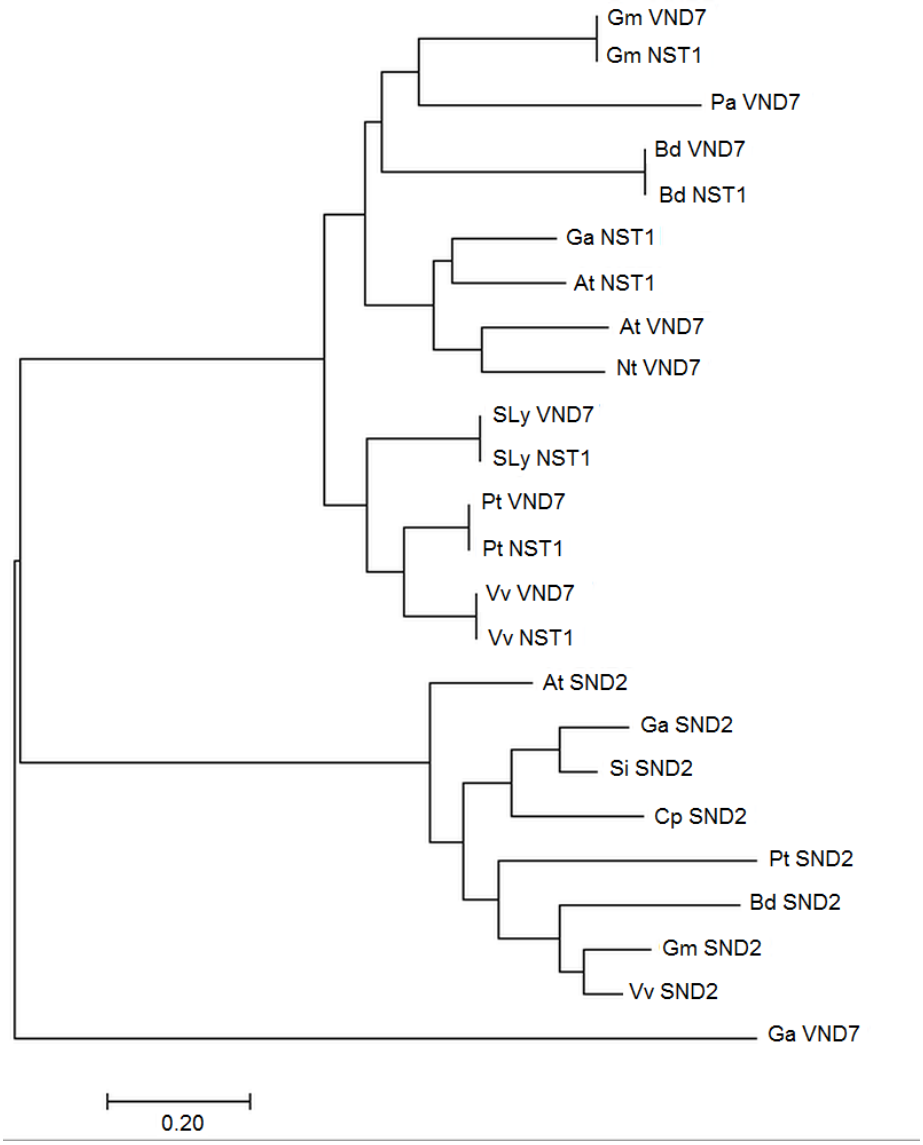


Fig 3. Phylogenetic analysis of *G. arborea* NAC transcription factors: VND7, NST1 and SND2 protein sequences identified from the reference transcriptome of *Gmelina arborea* (Ga) were compared to homologs from other species: At: *Arabidopsis thaliana* (*Q9C8W9*,

Q84WP6, *O49459*), Bd: *Brachypodium distachyum* (*Bradi1g04150.1.p*, *Bradi1g06970.1.p*, *Bradi1g37898.1.p*), Cp: *Carica papaya* (*XP_021889039*), Gm: *Glycine max* (*XP_006589457.1*, *Glyma.01G046800.1.p*, *Glyma.01G005500.1.p*), Nt: *Nicotiana tabacum* (*XP_016440678.1*), Pa: *Picea abis* (*MA_101849g0010*), Pt: *Populus trichocarpa* (*XP_024447115.1*, *Potri.001G061200.1*, *Potri.001G343800.1*), Si: *Sesamum indicum* (*XP_011096365*), Sly: *Solanum lycopersicum* (*Solyc01g009860.2.1*, *Solyc01g102740.2.1*), Vv: *Vitis vinifera* (*GSVIVT01000940001*, *XP_002267383*, *GSVIVT01015274001*). The clustering method used for dendrogram construction was neighbor-joining. Line length indicates the evolutionary distance. Uniprot, NCBI protein, TAIR and PlantTFDB accession IDs are shown in parenthesis. In the case of *Picea abis*, accession was obtained from iTAK plant transcription factor database (http://itak.feilab.net/cgi-bin/itak/db_gene_seq.cgi?trans_ID=MA_101849g0010).

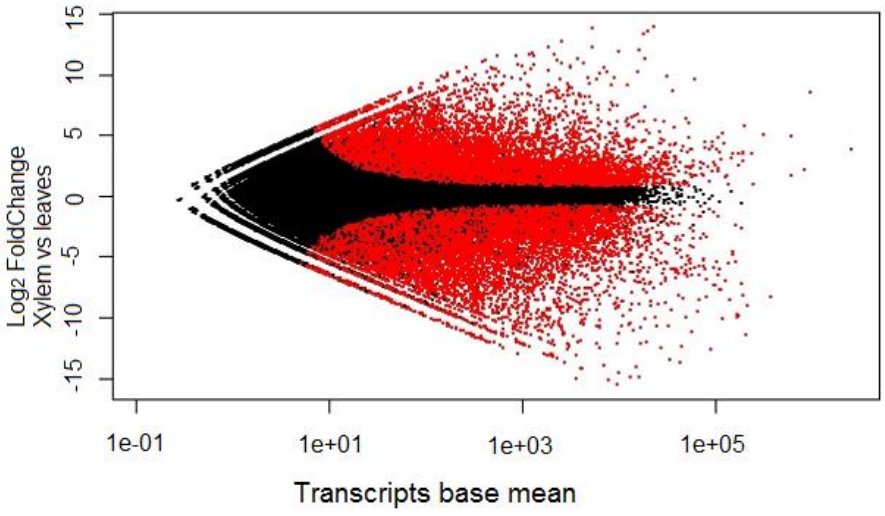


Fig 4. Distribution of differentially expressed transcripts (DEG) with a p-value < 0.05. DEG are shown in red and the non-DEG are shown in black.

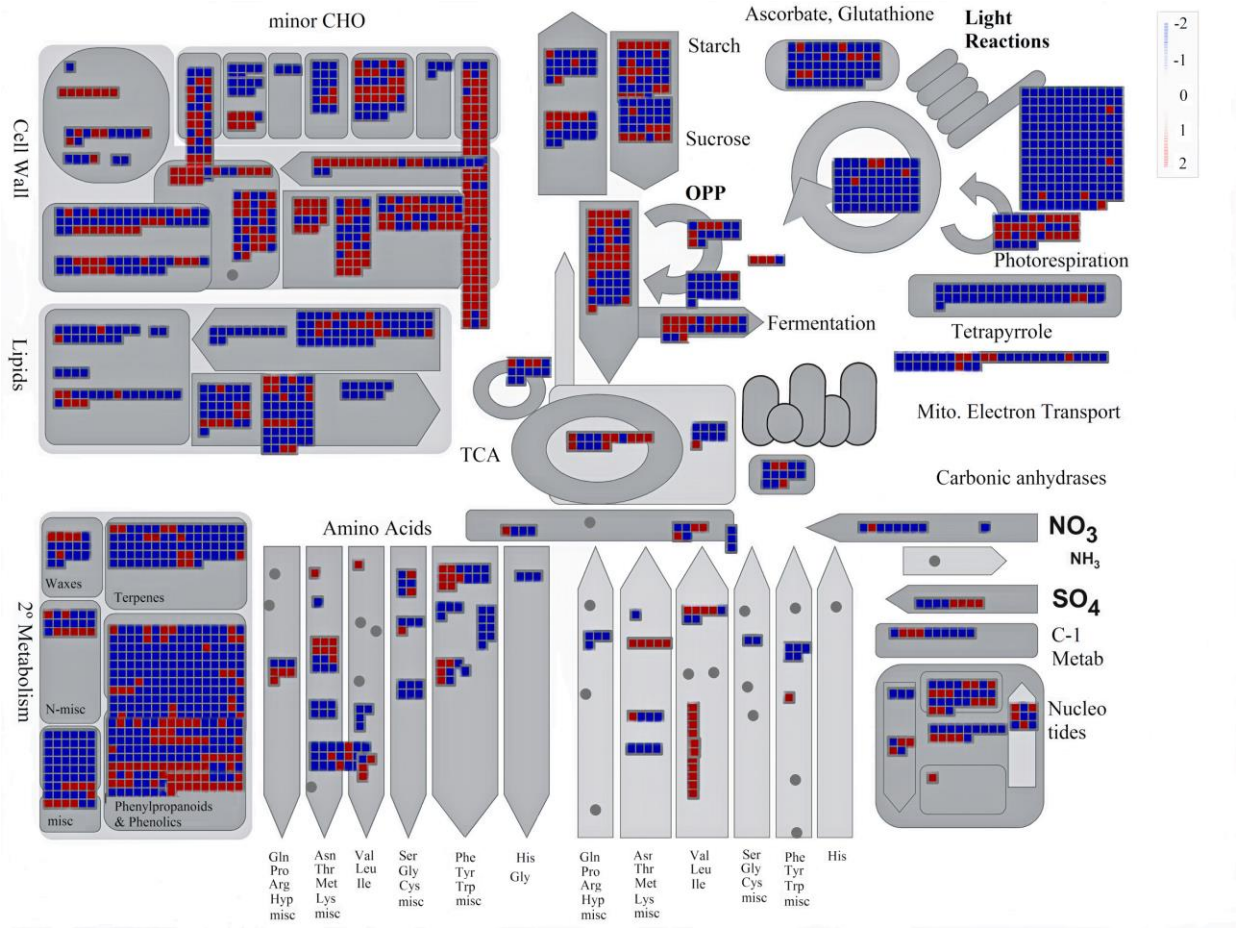


Fig 5. Differential expression between leaf and stem according to the main metabolic processes in which they are involved. The logarithm of changes of expression for each transcript is represented in red color (induction in stem, $Log_2FC \geq 2$) and blue (repression in xylem, induction in leaf, $Log_2FC \leq 2$). Analysis was performed using MapMan software.

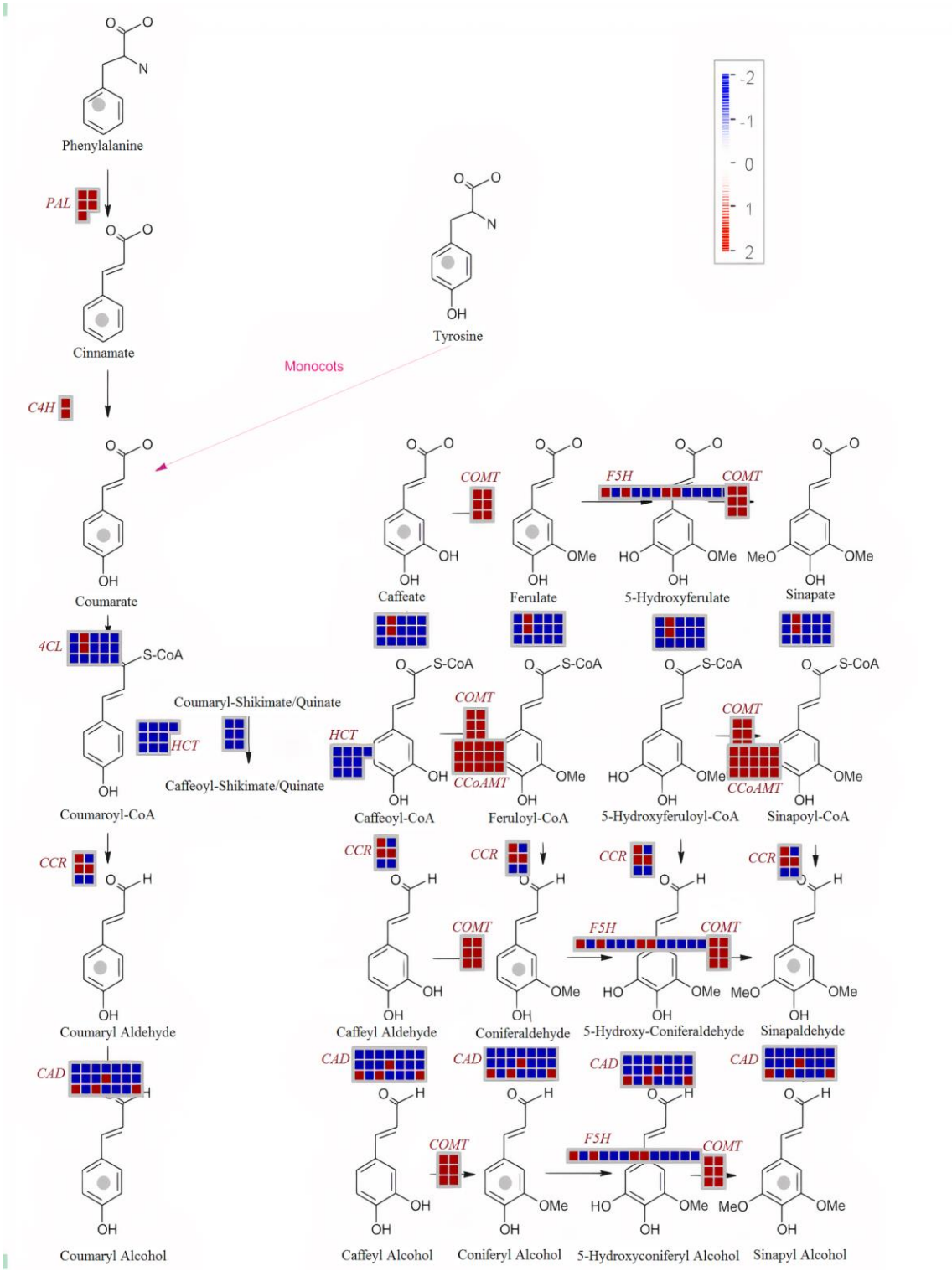


Fig 6. Differential expression of genes of the monolignol pathway, according to the logarithm of fold change (Log_2FC). Transcripts corresponding to each gene are represented in squares. In red are represented the Log2FC values ≥ 2 (induction in xylem) and in blue

the Log_2FC values ≤ 2 (repressed in xylem). Pathway analysis was performed using MapMan software.

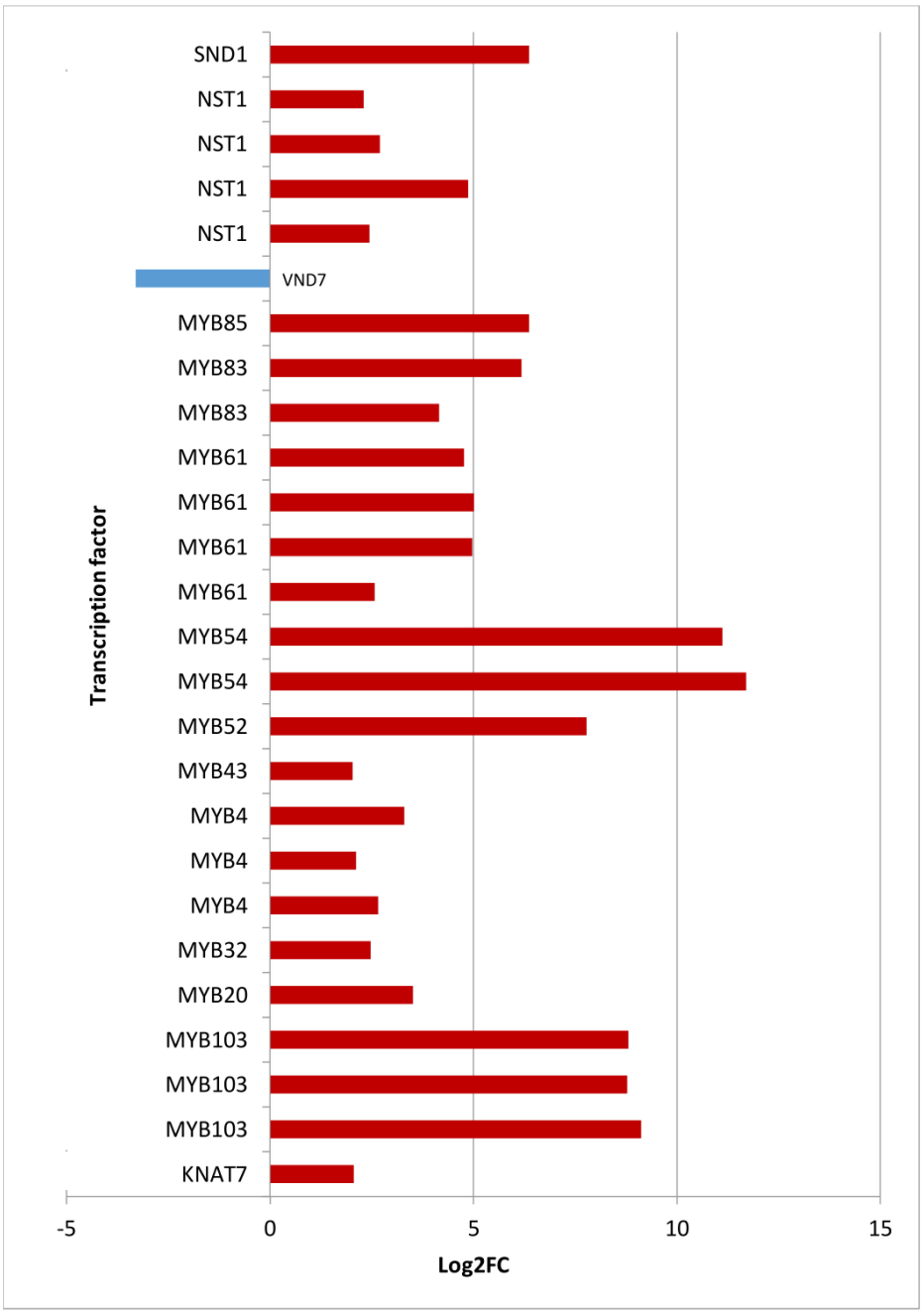


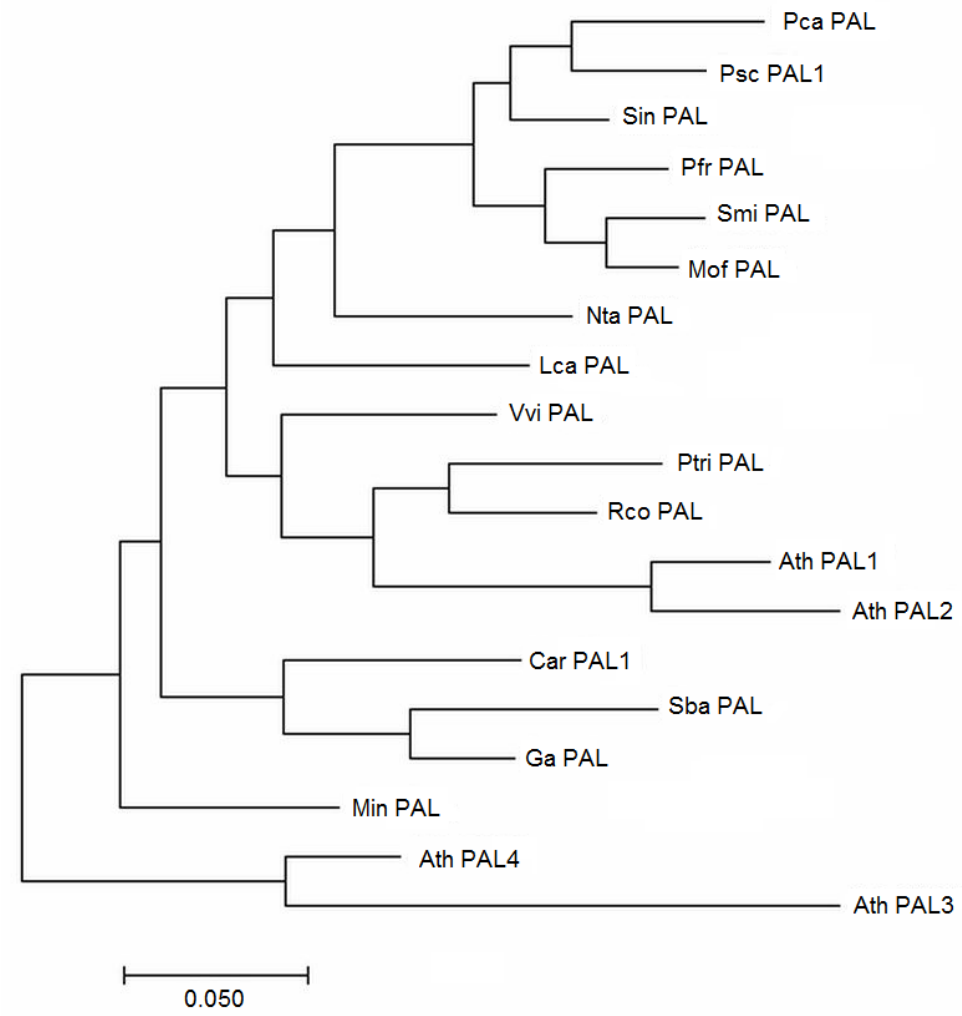
Fig 7. Differential expression of different transcripts identified as transcription factors involved in the regulation of the monolignol pathway. In red are shown the Log_2FC values

≥ 2 (induction in xylem) and in blue the Log_2FC values ≤ 2 (repressed in xylem, induction in leaf).

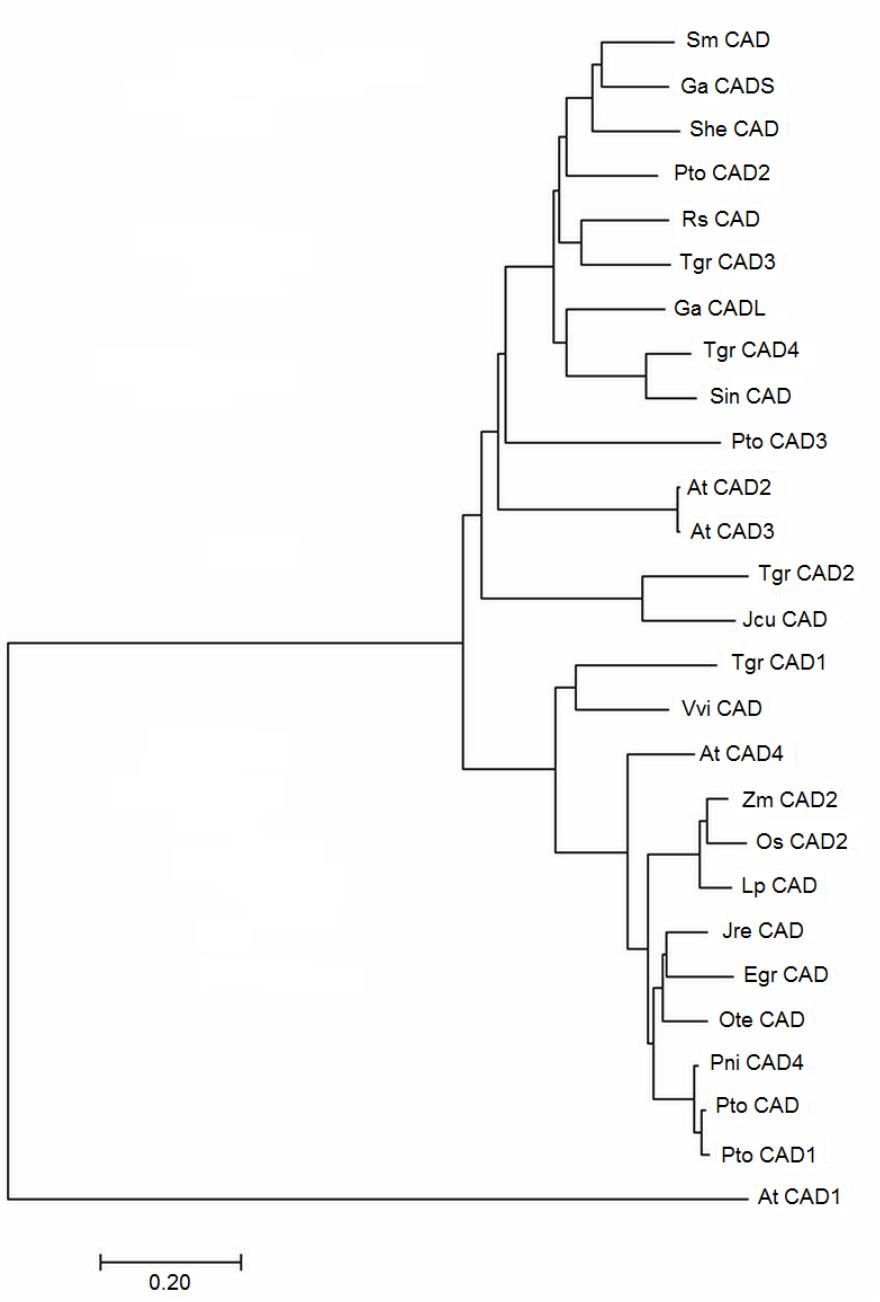
Group	Gene	Log2FC				
Celluloses synthesis	CESA2	-2	2	3,33		
	CESA4	-8	4,67	4,86		
	CESA5	3,7				
	CESA9	-3	3,79			
	IRX1	8,4	9,56			
	IRX3	9,1				
hemicelluloses synthesis	CSLA02	-3	-2,1			
	CSLC5	-4				
	CSLC6	-4				
	CSLD5	2,3	5,11			
	CSLE1	-9	-8	-5,3	-3	2,31
	CSLE2	-7	-3,2			
	CSLE3	-7				
	CSLG3	-8	-2,7			
	FRA8	3,6	4,07			
	GAUT12	2,6				
	IRX14	4,6				
	IRX6	-2	4,45	7,95	13	
	IRX9	3,1	3,57			
	PARVUS/GLZ1	3,9				
	PGSP1	3,1				
	PGSPI3	4,8	5,55	5,73	6,5	11,5
	PGSPI4	-3	-2,3	-2,1		
Laccases	LAC10	5,5	8,16			
	LAC11	6,3				
	LAC12	2,5	3,18	4,07	6	
	LAC13	4,6				
	LAC14	-3				
	LAC17	4,3	3,53	6,81	5,4	9,7
	LAC2	7,2				
	LAC6	2,9				
Programed cell death	XND1	2,6				
	VEP1	-5	-11	2,3	2	
	XCP1	3,6				
	XCP2	3,5	3,6			
Others	FLA11	4,7	5,5	6,2	13	

Fig 8. Genes related to the synthesis of other elements of the secondary cell wall with differential gene expression between stem and leaf. Red color represents Log_2FC values ≥ 2 (induction in xylem), blue colors Log_2FC values ≤ -2 (repressed in xylem).

A



B



1066

1067

1068 **Fig 9.** Phylogenetic analysis of *G. arborea* PAL (A) and CAD (B) proteins. Protein sequences

1069 of PAL and CAD enzymes obtained from *G. arborea* full length cognate transcripts were

1070 compared to homologous sequences belonging to other plant species. Dendrograms were

1071 constructed using the neighbor-joining clustering method. Line length indicates the

1072 evolutionary distance.

In addition to *G. arborea* (Ga) putative PAL1 sequence, other protein sequences used in PAL phylogenetic analysis were: *Ath: Arabidopsis thaliana*, with four paralogs of PAL included in the analysis, *AthPAL1* (*P35510*), *AthPAL2* (*OAP06573*), *AthPAL*3 (*OAO94639*) and *AthPAL4* (*OAP02490.1*). *Car: Coffea arabica (AEL21616), Lca: Lonicera caerulea (ALU09327), Nta: Nicotiana tabacum (NP_001312352.1), Min*: *Mangifera indica (AIY24975.1), Mof: Melissa officinalis (CBJ23826.1), Pfr: Perilla frutescens (AEZ67457.1), Psc: Plectranthus scutellarioides (AFZ94859.1), Pca: Pogostemon cablin (AJO53272.1), Ptri: Populus trichocarpa (P45730), Rco: Ricinus communis (AGY49231.1), Smi: Salvia miltiorrhiza (ABD73282), Sba: Scutellaria baicalensis (ADN32766.1), Sin: Sesamum indicum (XP_011094662), Vvi: Vitis vinifera (ABM67591),*

Protein sequences used in CAD phylogenetic analysis, included two possible variants of *Gmelina arborea (Ga),* the first one induced in stem (CADS, putative CAD3) and the second one induced in leaves (CADL). Other CAD protein sequences used were: *Ath: Arabidopsis thaliana* CAD1 *(OAP16446.1)* and CAD2 *(NP_179765), Egr: Eucalyptus grandis (XP_010024064.1), Jcu: Jatropha curcas (XP_012086572.1), Jre: Juglans regia (XP_018827699.1), Lp: Lolium perenne (AAB70908), Ote: Ocimum tenuiflorum (ADO16245.1), Os: Oryza sativa (Q6ZHS4), Pni: Populus nigra (AFR37935.1), Pto: Populus tomentosa (AAR83343.1), Rs: Rauvolfia serpentine (ALW82980.1), Sm: Salvia miltiorrhiza (ADN78309.1), Sin: Sesamum indicum (XP_011097452.1), She*: *Sinopodophyllum hexandrum (AEA36767.1), Tgr: Tectona grandis (ANG60951.1, ANG60952.1, ANG60953.1, ANG60954.1), Vvi: Vitis vinifera (RVW57228.1), Zm: Zea mays (NP_001105654).* Different CAD members were included for some species. Accession IDs from protein NCBI database are shown in parenthesis.

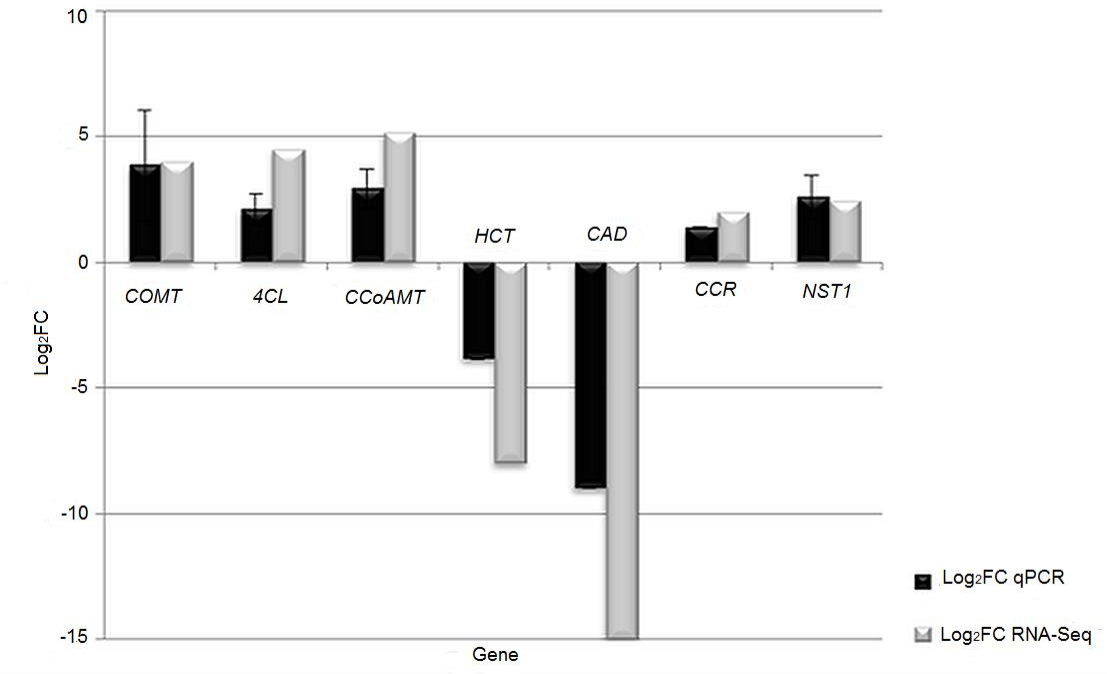
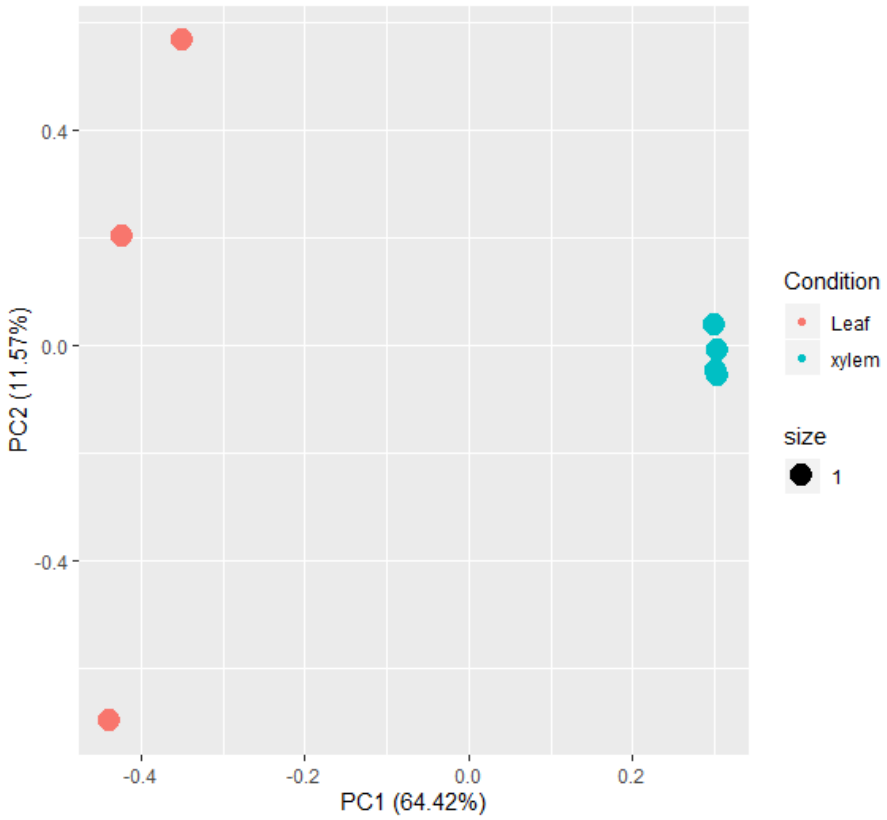
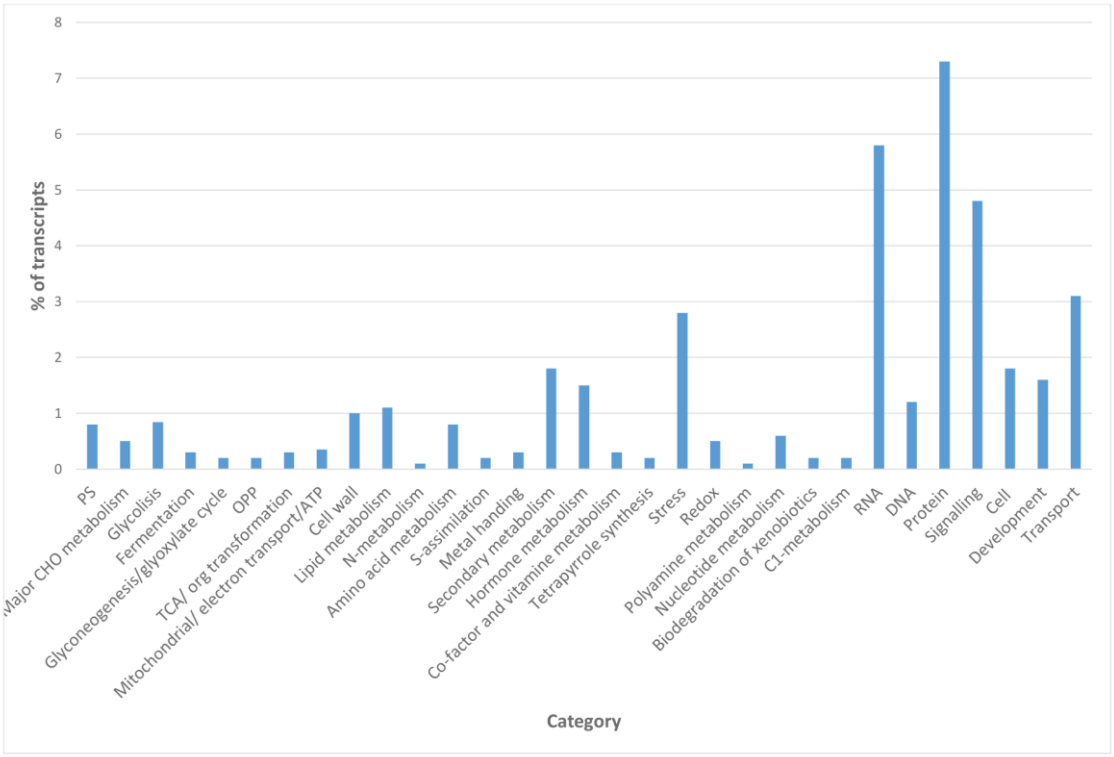


Fig 10. RT-qPCR differential expression validation of a selection of seven *G. arborea* genes. Bars indicate log2FC of xylem expression compared with leaf expression: black bars, mean log2FC values obtained from RT-qPCR assays; gray bars, mean log2FC values obtained from RNA-seq data.

Supplementary figures and tables



Supplementary Fig 1. Principal component analysis (PCA) of *G. arborea* expressed transcripts. Transcript read counts obtained in each sample were used. Difference between plant tissues (condition) is highlighted.



1133

1134 Supplementary Fig 2**. Main functional categories represented by DEG.**

1135

1136 **Supplementary table 1**. Summary of *G. arborea* de novo transcriptome assembly metrics

1137 combining leaves and xylem RNaseq data.

Assembly	
Total number of sequences obtained	147,130,884
Number of sequences used for the assembly	311,868,206
Number of transcripts obtained post assembly	151,229
N50 value (in bp)	1332
Average contig length (in bp)	782.84
Number of bases assembled	118,457,690
Annotation	

Full length ORFs	20,156
Quasi full length ORFs	16,706
Transcripts with hits in TAIR10 (blastx)	53,537
Transcripts with Interpro domains	47,884
Transcripts classified in gene families	57,075
Transcripts with GO terms	41,674
Number of GO terms	3106

Supplementary table 2. Frequency in the number of repetitions found for SSRs microsatellite markers.

Unity Size	Number of SSRs
2	20,634
3	4463
4	319
5	17
6	27

Supplementary table 3. Primers used for RT-qPCR validation of differentially expressed genes between xylem and leaf tissues.

Gene /Primer	Sequence 5´-3´	Amplicon Size
PAL-F	AAGGCATTGCATGGAGGGAA	201
PAL-R	CTCAGCACCCTTGAACCCAT	
4CL-R	TTGACGGTGATGACGAGCTC	209
4CL-F	CTCAGTGACAGCGGAACCAT	
CADx-F	GACTCAACAAACCTGGTATGCACA	392
CADx-R	CGTCTCTTTCATCCCTCCAATGC	
CADl-R	GATAGGCACAATGGATGGTATCG	171
CADl-F	GCGCTTCCAACTATCGCCC	
COMT-F	GACAGGGTCTTGATGGAGGC	232
COMT-R	CACCACCAACATCGACCAGA	
CCoAOMT-F	GCATCAGGAGGTTGGGCATA	238
CCoAOMT-R	AGAGTAGCCGGTGTAAACGC	
CCR-F	CTGGAACAGTGATGGGTCCT	232

CCR-R	GCCACCTTAGCAGCAAAATC	
HCT-F	CTTTGTGTGGCGAACTCGTA	229
HCT-R	TTACCCCATCCAAAATCTGC	
NST1-F	ATGGCCAGAAATCAGACTGG	
NST1-R	ATGAAGTGAGGGGGCTTTCT	177
MYB85-F	AGCTGCCTATTCAGGGATGA	
MYB85-R	TGTCGCTGAAACAATCGAAG	
FRA8-F	TGGCTGGCTGACTTTTTCTT	213
FRA8-R	ATCCTCCATGGTGTGAAAGC	
PGSIP3-R	CAGCTCACCGACTACGACAA	230
PGSIP3-F	TCGTTCAGAAAACCCTGGTC	
Ces-R	TTCCGAAGGCAAGCTCTTTA	169
Ces-F	AGGCATCTCTGTGCTTCGAT	
UBQ5-F	GATAGAGGTGGTGCTGAACGA	179
UBQ5-R	AGTCCTTGAGGGTGATGTGG	
HIST3-F	GTTGCCTTGAGGGAGATCAG	176
HIST3-R	TCTTAGCGTGAATCGCACAC	
βTUB-F	TGGTGATCTCAACCACCTCA	211
βTUB-R	GATACTGCTGGGAGCCTCTG	

1147