BMC Plant Biology

Research article

Open Access

# A new genomic resource dedicated to wood formation in *Eucalyptus*

David Rengel[†1], Hélène San Clemente[†1], Florence Servant[1,2], Nathalie Ladouce[1], Etienne Paux[1,3], Patrick Wincker[4], Arnaud Couloux[4], Pierre Sivadon[1,5] and Jacqueline Grima-Pettenati*[1]

Address: [1]UMR CNRS/Université Toulouse III 5546, Pôle de Biotechnologies Végétales, 24 chemin de Borde Rouge, BP42617 Auzeville, 31326 Castanet Tolosan, France, [2]Current address : Syngenta Seeds SAS, BP27, 31790 Saint Sauveur, France, [3]Current address : INRA-UBP, UMR 1095, INRA Site de Crouël, 234 avenue du Brézet, 63100 Clermont-Ferrand, France, [4]Génoscope, CNRS, UMR 8030 and Université d'Evry, 91057 Evry, France and [5]Current address : Université de Pau et des Pays de l'Adour, UMR CNRS 5254 IPREM, IBEAS – BP1155, 64013 Pau Cedex, France

Email: David Rengel - rengel@scsv.ups-tlse.fr; Hélène San Clemente - sancle@scsv.ups-tlse.fr; Florence Servant - florence.servant@syngenta.com; Nathalie Ladouce - ladouce@scsv.ups-tlse.fr; Etienne Paux - etienne.paux@clermont.inra.fr; Patrick Wincker - pwincker@genoscope.cns.fr; Arnaud Couloux - acouloux@genoscope.cns.fr; Pierre Sivadon - pierre.sivadon@univ-pau.fr; Jacqueline Grima-Pettenati* - grima@scsv.ups-tlse.fr

* Corresponding author †Equal contributors

Published: 27 March 2009

*BMC Plant Biology* 2009, **9**:36 doi:10.1186/1471-2229-9-36

Received: 29 September 2008
Accepted: 27 March 2009

This article is available from: http://www.biomedcentral.com/1471-2229/9/36



## Abstract

**Background:** Renowned for their fast growth, valuable wood properties and wide adaptability, *Eucalyptus* species are amongst the most planted hardwoods in the world, yet they are still at the early stages of domestication because conventional breeding is slow and costly. Thus, there is huge potential for marker-assisted breeding programs to improve traits such as wood properties. To this end, the sequencing, analysis and annotation of a large collection of expressed sequences tags (ESTs) from genes involved in wood formation in *Eucalyptus* would provide a valuable resource.

**Results:** We report here the normalization and sequencing of a cDNA library from developing *Eucalyptus* secondary xylem, as well as the construction and sequencing of two subtractive libraries (juvenile *versus* mature wood and *vice versa*). A total of 9,222 high quality sequences were collected from about 10,000 cDNA clones. The EST assembly generated a set of 3,857 wood-related unigenes including 2,461 contigs (Cg) and 1,396 singletons (Sg) that we named 'EUCAWOOD'. About 65% of the EUCAWOOD sequences produced matches with poplar, grapevine, *Arabidopsis* and rice protein sequence databases. BlastX searches of the Uniref100 protein database allowed us to allocate gene ontology (GO) and protein family terms to the EUCAWOOD unigenes. This annotation of the EUCAWOOD set revealed key functional categories involved in xylogenesis. For instance, 422 sequences matched various gene families involved in biosynthesis and assembly of primary and secondary cell walls. Interestingly, 141 sequences were annotated as transcription factors, some of them being orthologs of regulators known to be involved in xylogenesis. The EUCAWOOD dataset was also mined for genomic simple sequence repeat markers, yielding a total of 639 putative microsatellites. Finally, a publicly accessible database was created, supporting multiple queries on the EUCAWOOD dataset.

**Conclusion:** In this work, we have identified a large set of wood-related *Eucalyptus* unigenes called EUCAWOOD, thus creating a valuable resource for functional genomics studies of wood formation and molecular breeding in this economically important genus. This set of publicly available annotated sequences will be instrumental for candidate gene approaches, custom array development and marker-assisted selection programs aimed at improving and modulating wood properties.

## Background

Wood is the major component of terrestrial plant biomass and is expected to play a significant role in future sustainable development as a renewable and environmentally acceptable source for fibers, solid wood and biofuel products [1,2]. Furthermore, wood is an important sink for atmospheric $CO_2$, an excess of which is a major cause of global warming.

The production of wood or secondary xylem by xylogenesis is a remarkable example of terminal differentiation, producing a complex three-dimensional tissue specialized in conduction and mechanical support. This differentiation process comprises four major steps: cell division, cell expansion, deposition of lignified secondary cell wall and programmed cell death. The vascular cambium is the meristem tissue responsible for this differentiation process and, thus, for the extensive radial secondary growth of trees, ensuring regular renewal of functional secondary xylem and phloem during the lifespan of these perennial species.

Trees are long-living organisms that grow in a variable environment and are subject to developmental cues. As a consequence, wood is highly variable at the tissue level (in the proportions of different cell types) as well as at the cellular level (in cell size, shape, cell wall structure and composition). Anatomical, chemical and physical differences in wood properties are not only widespread from tree to tree, but also within a single tree [2]. For instance, variations between juvenile and mature wood present within the same tree produce distinct wood properties such as density and pulp yield [3].

The genus *Eucalyptus* is one of the main sources of wood worldwide and is the most widely used tree species in industrial plantations. Many *Eucalyptus* species are renowned for their fast growth, straight form, valuable wood properties, wide adaptability to soils and climates, and ease of management through coppicing [[4] and references therein]. According to the United Nations Food and Agriculture Organization [5], *Eucalyptus* is the principal hardwood species used for pulp extraction, with 19 million hectares of industrial plantations worldwide.

Because of their comparatively long generation times, forest trees are still at the early stages of domestication compared to crop species, with most breeding programs only one or two generations away from the wild. Nevertheless, the genetics of *Eucalyptus* is becoming one of the most advanced in forestry [4]. Nowadays, wood traits, which rely mainly on lignified secondary cell wall properties, are the key focus to many breeding programs. *Eucalyptus* breeding programs will thus benefit from genomic technologies that could significantly speed up the process of genetic improvement [4].

The genomes of most *Eucalyptus* species are very similar to those of poplar species, with a relatively small size (370–700 Mbp) and diploid inheritance ($n$ = 11). In addition, the *Eucalyptus* trees are fast growing, most species are amenable to clonal propagation and some can be genetically transformed. These features make *Eucalyptus* particularly suitable for genomic technologies and a growing number of genetic tools (genetic, physical maps and quantitative trait loci) as well as EST collections are becoming available for some species. However, the huge commercial potential of eucalypts has fostered a situation in which access to genomic resources is restricted to a small number of private research consortia. These limitations may be overcome by the initiative of an International *Eucalyptus* Genome Consortium [6], which promoted the sequencing project of the *Eucalyptus grandis* genome undertaken by the US Department of Energy.

Because wood quality is a major trait that tree breeders would like to improve by using marker-assisted selection, it is important to increase publicly available *Eucalyptus* genomic resources, including putative candidate genes involved in the genetic control of wood properties. Indeed, recent advances in the molecular study of xylogenesis have revealed that wood formation is under strong genetic control, notably at the transcriptional level [7,8]. The production and analysis of ESTs from wood-forming tissues has increased our understanding of gene regulation involved in wood formation in tree species including loblolly pine [9-11], poplar [7,12], and white spruce [13]. Similarly, large scale sequencing of ESTs will be instrumental for the annotation of the *Eucalyptus* genome sequence. As a first step towards this goal, we have generated two secondary xylem subtractive libraries (xylem *versus* leaves and xylem *versus* phloem) rendering 487 unigenes preferentially or specifically expressed in differentiating secondary *Eucalyptus gunnii* secondary xylem [14,15], and providing a useful tool for gene profiling [16].

Here we present the sequencing of 9,216 normalized clones from a *E. gunnii* secondary xylem cDNA library generated in our laboratory [17]. In addition, we report the construction and sequencing of two suppression subtractive hybridization (SSH) libraries aimed at identifying genes differentially expressed in juvenile *vs* mature wood and *vice versa*. Sequencing of these EST libraries was performed in the framework of the French project FOREST [18] whose goal was to release ESTs sequences from woody species through public databases. *Eucalyptus* EST sequences produced in our lab have been assembled into a unigene dataset called EUCAWOOD and the unigenes have been functionally annotated and compared with other plant species. The functional annotation of the unigene set is discussed in the context of the wood formation process.

## Results and Discussion

### *Construction and sequencing of normalized libraries*

With the aim of sequencing a large number of ESTs representative of the set of mRNAs expressed in secondary xylem, we chose a cDNA library prepared from the differentiating secondary xylem of *E. gunnii* [$Xyl_{cDNA}$] containing $1.5 \times 10^6$ clones [17], which has already proven a good source of genes expressed during wood formation [17-23]. Because in cDNA libraries, each cDNA occurs at a frequency proportional to that of its corresponding mRNA in the tissue it was prepared from, prevalent and intermediate frequency classes of mRNAs are expected to be overwhelming in a random large scale sequencing program. In order to minimize this redundancy and increase the chance of identifying low-expressed genes, we decided to normalize the $Xyl_{cDNA}$ library according to the protocol of Bonaldo [24]. During the normalization procedure, human *desmin* cDNA was added at 1,000 copies to the non-normalized library whereas *EgCAD2*, of which 31 cDNA copies were present before normalization, served as an internal control. After normalization, six copies of *desmin* and five copies of *EgCAD2* were recovered, demonstrating that redundancy in the library was drastically reduced by the normalization procedure. Thus, the representation of the different genes expressed in secondary xylem was expected to be increased among the 9,216 clones of the normalized $Xyl_{cDNA}$ library as compared to the original library.

All 9,216 $Xyl_{cDNA}$ clones were sequenced from the 5' end. Following vector and low-quality sequence trimming, 8,043 high quality sequences with an average length of 566 nucleotides (nt) were retained. Sixty three percent (5,060) of the sequences were longer than 500 nt and only six percent (486) were shorter than 200 nt, indicating the quality of the library. These 8,043 sequences were deposited in the EMBL-EBI nucleotide database [EMBL: CT980028 to CT988078].

To complement this set of ESTs, we decided to seek for genes that are differentially expressed in juvenile and mature secondary xylem tissues. The transition from juvenile to mature xylem is known to be an important source of variation in wood quality [3]. We took advantage of SSH technology, known to equalize the level of representation of rare and abundant fragments [25], to reciprocally subtract cDNAs prepared from juvenile and mature secondary xylem tissues. Thus, we produced two SSH libraries: a juvenile *vs* mature (Jm) and a mature *vs* juvenile (Mj) secondary xylem library. Altogether, 818 clones were obtained and sequenced from both sides of the cloning site. A total of 1,179 good quality sequences with an average length of 412 nt were obtained, 604 from the Jm library and 575 from the Mj library. The sequences were deposited in the EMBL-EBI nucleotide database [EMBL: CT988079 to CT989251].

### *EST assembly*

The assembly of the 9,222 good quality sequences described above together with the ESTs and core nucleotide sequences publicly available in the GenBank and EMBL databases, generated 17,087 unigenes, comprising 7,921 contigs and 9,166 singletons. Among these, we discarded all sequences whose size was below 100 nt and selected for further analysis only the 3,857 unigenes (2,461 Cg and 1,396 Sg) which contained at least one sequence originating from one of our libraries including two SSH libraries previously obtained in the laboratory, *i.e.* a secondary xylem *vs* secondary phloem SSH library (Xp) [14] and a secondary xylem *vs* leaves SSH library (Xl) [15]. The rationale for this was to select a subset of secondary xylem-related sequences that we called 'EUCAWOOD' (see Additional file 1). The EUCAWOOD unigenes had an average length of 640 nt and a size distribution as shown in Figure 1. To mine this new *Eucalyptus* genome resource, we have developed a publicly accessible database that supports multiple queries on the EUCAWOOD unigenes and their functional annotation [26].

The Venn diagram in Figure 2A illustrates the number of unigenes shared between the cDNA library ($Xyl_{cDNA}$) and each of the four different SSH libraries. Interestingly, most of the contigs containing sequences originating from at least one of the SSH libraries (*i.e.* 269) were not present in the $Xyl_{cDNA}$ library. Only 107 contigs contained ESTs originating from the $Xyl_{cDNA}$ and one of the SSH libraries (Figure 2A). This little overlap confirms the utility of combining cDNA and SSH libraries to identify new genes expressed in *Eucalyptus* secondary xylem: the SSH libraries contain many clones not recovered from the total cDNA library.

Figure 2B illustrates the low number of overlapping sequences between the four different SSH libraries. For instance, the Jm and the Mj subtractive libraries we generated assembled into 279 unigenes of which only 17 contained ESTs from both libraries. This limited overlap (6%) between the two libraries illustrates the efficacy of the subtraction procedure in the SSH technique. Most interestingly, the low overlap between the four libraries demonstrates the advantage of using several subtractive libraries to recover new genes distinct to each tissue.

### *Sequence comparisons with other species*

Homology searches were conducted using the BlastX program [27] to compare the "EUCAWOOD" unigene set with predicted protein and gene model databases for arabidopsis, poplar and rice, four plant species whose genomes have been sequenced [28-31]. These homology searches allowed us to assess the overlap between the EUCAWOOD unigenes and the protein sequence databases of these three model plants (Figure 3). Approximately 55% of the unigenes (2,150) matched sequences

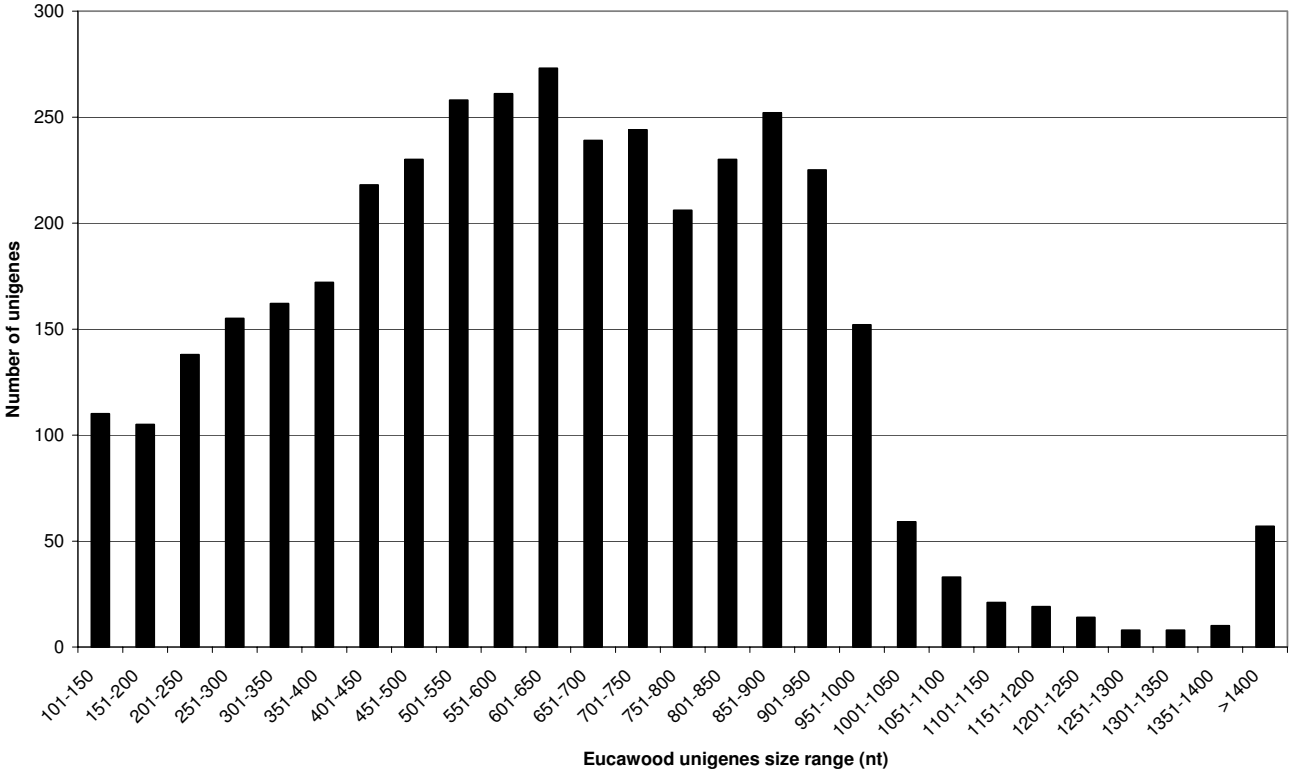


**Figure 1**
**Size distribution of the EUCAWOOD unigenes after assembly**.

occurring in all three species and 65% (2,567) matched sequences in at least one of these three species. The highest number of hits were obtained with the two woody angiosperms, *i.e.* poplar (2,474) and grapevine (2,451), followed by arabidopsis (2,350) and rice (2,243) predicted protein sequences. Interestingly, 171 unigenes matched only against poplar and/or grapevine sequences, the only woody species whose genomes have been sequenced so far (see Additional file 2). Most of these 171 unigenes corresponded to unknown proteins, 45 of them only matched predicted proteins of *Vitis vinifera*, 52 had no other hit than gene models from poplar at an E-value cut-off of $10^{-10}$, 74 were common to poplar and grapevine. Further investigation is needed to verify whether these latter sequences correspond to genes specifically expressed during wood formation in trees.

### *Functional annotation*

To further allocate protein annotations to the EUCAWOOD unigenes, BlastX searches were performed against the Uniref100 database [32]. GO terms [33] associated with the best Uniref100 hit were then automatically assigned to the corresponding EUCAWOOD unigenes. Functional annotation data are presented in Additional file 1 as well as in the public EUCAWOOD database [26]. Overall, 2,466 (64%) unigenes produced matches to proteins in Uniref100. A total of 2,850 GO terms were allocated to 1,316 unigenes, filed under 'Biological Process' (1,018 terms), 'Molecular Function' (1,138 terms) and 'Cellular Component' (694 terms) (Figure 4 and Additional file 3). The vast majority of the 1,018 GO terms allocated to Biological Process genes fell under the categories 'Metabolism' (819 terms) and 'Cellular process' (767 terms) (Figure 4). The large proportion of unigenes involved in metabolic and biosynthetic processes confirms that differentiating secondary xylem is a very active tissue with a high metabolic rate. A large number of the terms allocated to 'Molecular Function' were in genes in the subcategories 'Catalytic Activity' (668 terms) and 'Binding' (665 terms) (Figure 4). The most represented activities in Catalytic Activity were transferases (230 terms), hydrolases (197 terms) and oxidoreductases (156 terms). The most abundant Binding activites were nucleotide binding (219 terms), iron binding (208 terms), nucleic acid binding (161 terms) and protein binding (119 terms).

In a parallel annotation approach, we related the best Uniref100 hit of every unigene to the PFAM database [34,35] in order to identify protein families and domains in the EUCAWOOD unigene set. A total of 1,453 unigenes (37%) were assigned at least one PFAM identifier (ID)

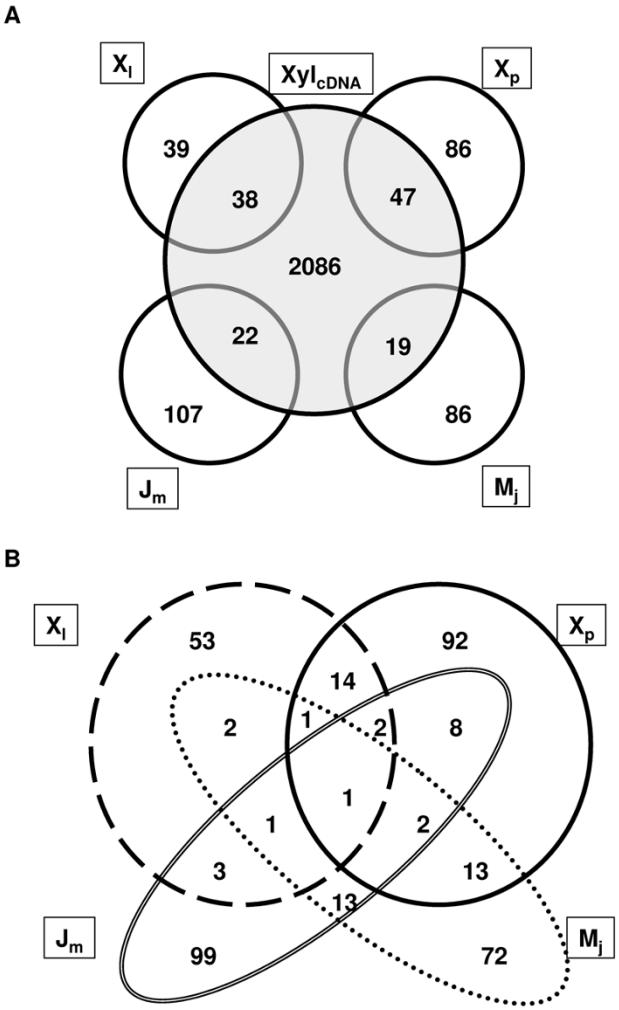


**Figure 2**
**Overlap between the EUCAWOOD unigenes**. (A) Venn diagram showing the overlap between unigenes originating from the cDNA library [$Xyl_{cDNA}$] and each of the SSH libraries [Jm: juvenile *vs* mature secondary xylem; Mj: mature *vs* juvenile secondary xylem; Xl: secondary xylem *vs* leaves; Xp: secondary xylem *vs* secondary phloem]. (B) Venn diagram showing the overlap of unigenes derived from the four different SSH libraries.


**Figure 3**
**Number of EUCAWOOD unigenes with similarities to predicted proteins from four plant species**. BLASTX searches (E value ≤ $e^{-10}$) were conducted to identify EUCAWOOD unigenes in the JGI Poplar Proteins v*1.1*, *Arabidopsis TAIR7 Peptides*, *TIGR Rice Genome Annotation* and NCBI *(Vitis vinifera)* databases.

and, overall, 825 PFAM protein families and domains were represented among EUCAWOOD unigenes. Remarkably, PFAM IDs related to signal transduction and cell wall metabolism formed the majority of the 20 most abundant protein families (Figure 5 and Additional file 4). These PFAM matches showed that the most abundant protein families in the EUCAWOOD unigene set were also among the most represented in comparable studies with other plant species [13,36]. Similar examination of the various protein families represented in the subtractive libraries (Jm and Mj) revealed a completely different pattern from that of the EUCAWOOD dataset, in which the large majority of the unigenes originate from the $Xyl_{cDNA}$ library (Additional file 4). EUCAWOOD unigenes containing ESTs from Jm or Mj libraries produced matches with 57 and 47 different protein families, respectively, including only five families common to both Jm and Mj libraries. Among these 99 protein families, only seven appeared among the 20 most abundant families in the EUCAWOOD dataset. The PFAM annotation of the Uniref100 matches confirmed the little overlap between both libraries at the protein family level with only five common PFAM IDs.

Finally, 1,261 (32,7%) of EUCAWOOD unigenes produced no match against Uniref100, arabidopsis, poplar, grapevine or rice proteinsand were therefore considered as 'No Hits' at E value ≤ $e^{-10}$ (Additional file 5). The average length of the "No Hits" was remarkably shorter than that of the unigenes showing at least one BLASTX hit (447 nt *vs* 738 nt). Consistent with this, the percentage of unigenes shorter than 400 nt was much higher among the 'No Hits' than among the 'Hits' (47.6% *vs* 10.1%). The opposite was also true for unigenes longer than 800 nt: the percentage of unigenes longer than 800 nt was much lower among the "No Hits" than among the "Hits" (12% *vs* 36%). The "No Hits" group is enriched in 3' sequences, which are usually less conserved than those upstream in the gene.

### *Cell wall-related genes*

One of the crucial stages in xylem differentiation is the formation of the secondary cell wall, which is largely composed of cellulose, lignin and hemicelluloses together with other less abundant polysaccharides and structural

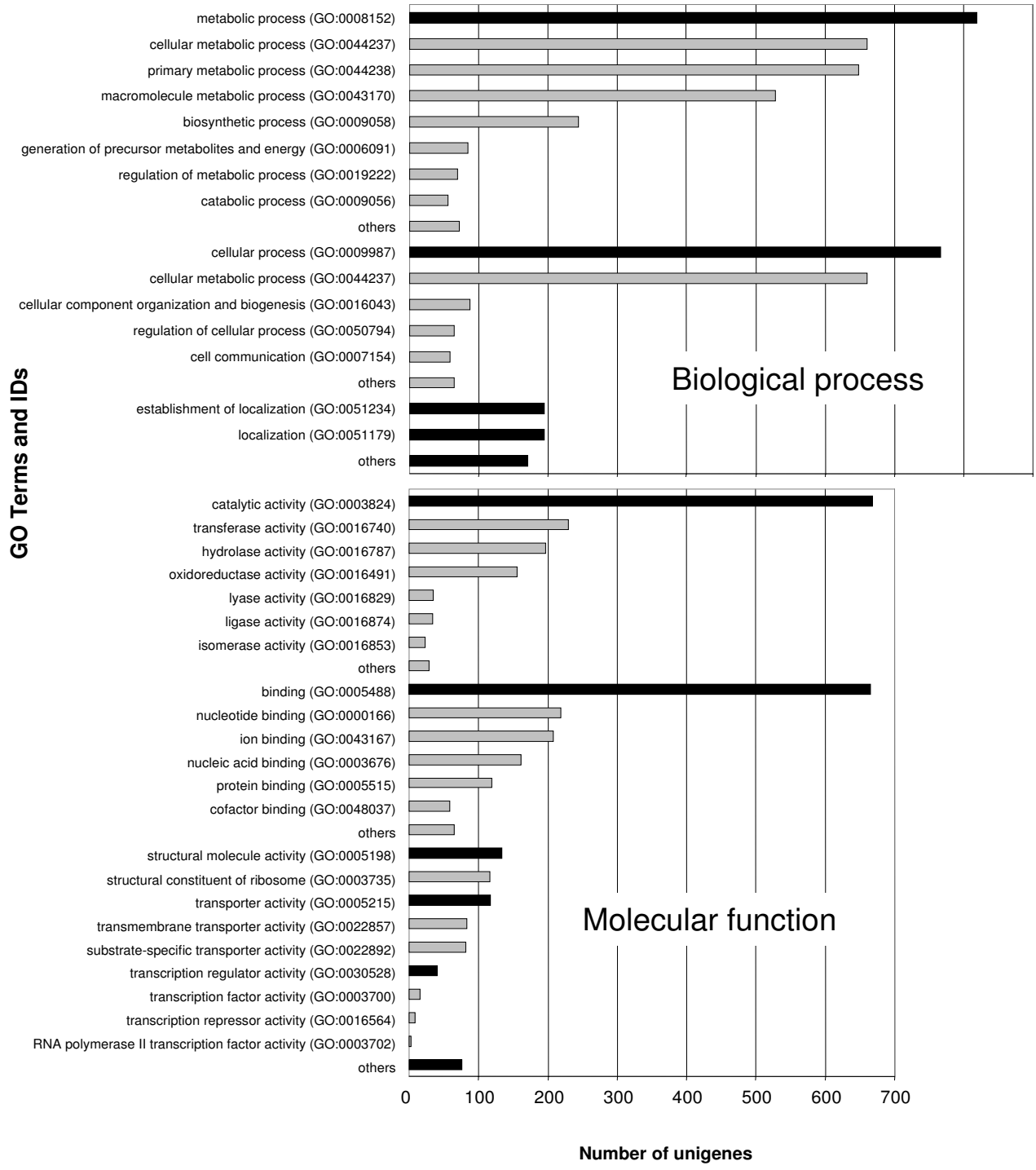


**Figure 4**

**Gene ontology assignments to EUCAWOOD unigenes**. GO terms were allocated to EUCAWOOD unigenes according to their best hit in searches of the Uniref100 database (E value $\leq e^{-10}$). Terms and IDs belonging to the 'Biological Process' and 'Molecular Function' categories are shown. Black bars indicate the main subcategories whereas the grey bars immediately below them illustrate subcategories therein. (Terms and IDs belonging to 'Cellular Component' category can be found in Additional file 1.)

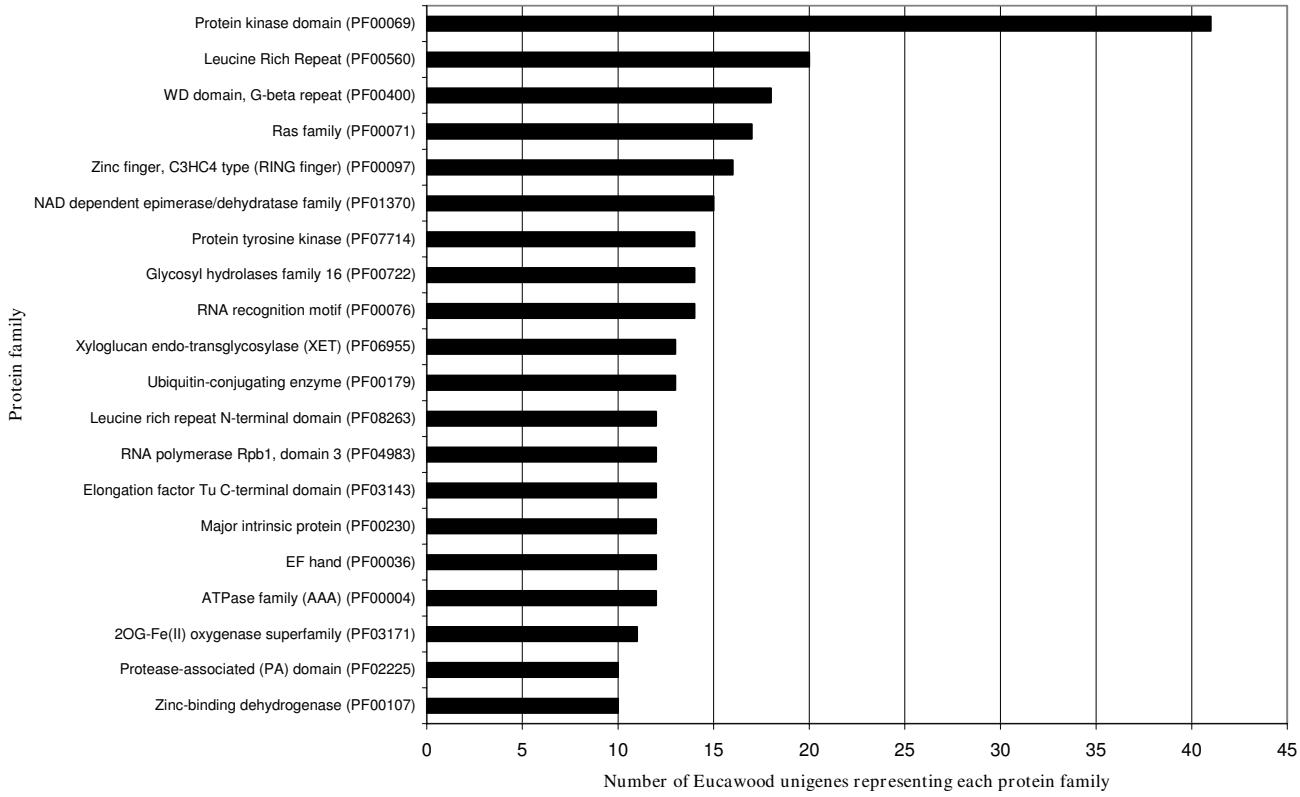


**Figure 5**
**Protein families among EUCAWOOD unigenes**. A total of 825 protein families from the PFAM protein family database were represented in the EUCAWOOD dataset. The black bars indicate the occurrence of the 20 most abundant protein families.

proteins [37]. We therefore mined the EUCAWOOD unigene set for genes involved in lignin biosynthesis, carbohydrate and cell wall metabolism. We performed BlastX searches using both the Cell Wall Navigator (CWN) [38,39], and the MAIZEWALL databases [40]. Altogether, 422 EUCAWOOD unigenes matched cell wall-related genes, with 142 and 380 hits with CWN and MAIZEWALL, respectively (Additional file 6). Among those, 101 were common to both databases, and 279 were found only in MAIZEWALL representing altogether the totality of the 18 categories described in this database. Most of the hits found only in MAIZEWALL were secondary cell wall-related genes including phenylpropanoid and lignin biosynthetic genes.

### *Lignin biosynthesis genes*

All the gene families involved in the monolignol biosynthesis pathway were represented in the EUCAWOOD dataset including 18 unigenes (Additional file 7) with similarities to the set of lignin biosynthetic genes identified in *Arabidopsis* by Raes *et al* [41]. The EUCAWOOD set contained three distinct genes encoding hydroxycinnamoyl-CoA:shikinimate/quinate hydroxycinnamoyltransferase (*HCT*) suggesting that HCT in *Eucalyptus* is encoded by a small gene family as in poplar [42] rather than a single *HCT* gene, as in *Arabidopsis* [41]. Interestingly, eight ATP-binding cassette (ABC) transporters were present among EUCAWOOD unigenes, which might be involved in the transport of the lignin monomers to the cell wall through direct membrane pumping [43]. The molecular mechanism by which monolignols are incorporated into the lignin polymer is thought to involve key oxidation steps catalyzed by laccases and peroxidases [44]. Six putative laccases were found among the EUCAWOOD unigenes, one of which was most similar to *TT10*/*AtLAC15*, which has recently been proven to play a role in lignin synthesis [45]. Three of these six unigenes were similar to *IRX12*/*LAC4*, a gene involved in cell wall biosynthesis [46]. The expression of *IRX12*/*LAC4* might be regulated by *At*MYB26/MALE STERILE34, a MYB transcription factor involved in secondary thickening of the anthers in *Arabidopsis* [47]. Eight EUCAWOOD unigenes were annotated as encoding peroxidases. Three of them are homologues of *At*PER12 and *At*PER64, two proteins whose precise biochemical functions remain elusive but which have been located in the cell wall [48].

*Carbohydrate active enzymes and cell wall metabolism genes*

The three-step process of cellulose biosynthesis was represented within the EUCAWOOD unigenes set [49,50]. Three sucrose synthases (SuSy) were found: one was similar to *At*SUS1 whereas the other two were similar to *At*SUS4 [51]. In addition, five unigenes homologous to members of the cellulose synthase (*CesA*) multigene family were also found that correspond to the *EgCesA1-EgCesA5* genes recently described in *E. grandis* [52]. *EgCesA1, EgCesA2 and EgCesA3* are specifically expressed during secondary cell wall biosynthesis, whereas expression of *EgCesA4* and *EgCesA5* is linked to the synthesis of primary cell wall. Two unigenes similar to *KORRIGAN (KOR)* proteins were alsoretrieved from EUCAWOOD. Several studies have proven the importance of KOR proteins in the formation of the plant cell wall in various species. For instance, *Arabidopsis irx2* and *kor1* mutations, which map to the same gene, both affect secondary growth [53].

The EUCAWOOD set also contained unigenes with homologies to *Arabidopsis* proteins dedicated to hemicellulose and pectin biosynthesis including three putative cellulose synthase-like genes, thought to be involved in the synthesis of the backbone structures of mannans, glucomannans and galactomannans [54]. We also found eight unigenes similar to UDP-xylose synthases, one to UDP-xylose epimerase, two to β-xylosidases, one to glucuronic acid epimerase, two to pectin esterases, four to pectate lyases and four to polygalacturonases.

Several unigenes similar to other gene families thought to be involved in cell wall formation were also found. Two unigenes were similar to *PttGH19A*, which encodes a chitinase-like protein highly expressed during poplar secondary cell wall biosynthesis [55]. Mutation of two genes similar to *PttGH19A* in *Arabidopsis (At1g05850* and *At3g16920*) caused deficient biosynthesis and incorporation of cellulose into the cell wall, as well as ectopic lignin deposition and aberrant cell shapes with incomplete cell walls [56].

Genes encoding proteins involved in loosening and rearrangement of the cell wall were also present among the EUCAWOOD unigenes, including, for instance, two expansin genes. Expansins are thought to directly promote cell expansion by hydrolysing noncovalent bonds between cellulose and hemicelluloses in the cell wall [57]. The action of expansins is facilitated by xyloglucan endotransglycosylases (XETs)/hydrolases (XEHs), also known as XTHs, which incorporate and modify xyloglucans into the cell wall [58]. XTH proteins are members of the glycosyl hydrolase (GH) family 16, which is the most abundant carbohydrate-metabolising enzyme group among the EUCAWOOD matches in the CWN database, represented by 19 unigenes. A total of 41 gene models belonging to the GH16 family have been recorded in the genome of poplar [54].

Whereas carbohydrates and lignin constitute the bulk of cell wall materials, structural proteins also form a network that contributes to the architecture and functionality of the cell wall. This is the case for fasciclin-like proteins (FLA), a subgroup of arabinogalactan proteins involved in processes such as growth and cell proliferation. Five FLAs were identified in the EUCAWOOD unigene set. All five are similar to *AtFLA11* and *AtFLA12*, whose expression is linked to secondary cell wall biosynthesis and maturation [59].

***Transcription factors***

Given the importance of transcriptional regulation during wood formation, we carried out BlastX searches comparing the EUCAWOOD unigene set with the Plant Transcription Factor Database (PTFD) [60] and the Database of *Arabidopsis* Transcription Factors (DATF) [61]. A total of 141 unigenes (110 Cg and 31 Sg) had at least one hit in either database. PTFD and DATF produced 136 and 103 hits respectively, with 98 unigenes having a hit in both databases (Additional file 8). Interestingly, 90 of the 136 PTFD hits corresponded to poplar sequences, whereas only 24 matched *Arabidopsis* and 10 matched rice proteins. The 141 hits identified 41 transcription factor families, some of which are known to play a role in secondary growth and wood formation [8,62]. The 'C2H2 zinc-finger' family was the most frequently represented among the EUCAWOOD unigenes, with 15 putative members, followed by the MYB and NAC families, each represented by 11 putative unigenes. A number of plant MYB proteins, including *Eucalyptus* and other woody species, have already been proven to regulate the biosynthesis of phenolic compounds, including lignin [22,23,62,63]. Putative orthologs of NAC factors known to play a role in xylem differentiation were found among the EUCAWOOD sequences. For instance, the NAC secondary wall thickening promoting factor genes *NST1* and *NST3* are implicated in the formation and thickening of secondary wall in *Arabidopsis* [64,65]; *ANAC012/SND1*, a member of the IIb group of the *NAC* family, has recently been described as a key regulator of xylary fiber development [66,67]. A putative ortholog of the negative regulator of both secondary cell wall synthesis and programmed cell death, *ANAC104/XND1* [68], was also present in EUCAWOOD. Three unigenes resemble LIM transcription factors, some of which have been shown to regulate the expression of lignin biosynthetic genes [69,70]. In fact, Cg2892 is similar to *EcLIM1* from *E. camaldulensis*, which shares 86% homology with *Nicotiana tabacum NtLIM1*. Suppression of *NtLIM1*expression caused the downregulation of lignin biosynthesis genes such as phenylalanine ammonia-lyase (*PAL*), 4-coumarate CoA ligase (*4CL*), cinnamate 4-hydroxylase (*C4H*), and cinnamyl alcohol dehydrogenase (*CAD*) [69,70].

The auxin-inducible factor (AUX/IAA) and auxin-response factor (ARF) families were represented by four EUCAWOOD unigenes. One is similar to *IAA13* and its closely related *BDL/IAA12*, whose mutation disrupts the normal cell and tissue organization along the apical-basal axis resulting in discontinuous and reduced vascular formation [71].

Six homedomain-leucine zipper proteins were present in the EUCAWOOD dataset. Among them, one contig (Cg3498) is similar to *ATHB15* and *ATHB8*, members of class III (HD-ZIPIII). These proteins are involved in vascular development and wood formation and share antagonistic functions with other HD-ZIPIII proteins such as REVOLUTA, PHABULOSA (PHB), and PHAVOLUTA [72]. A putative ortholog of PHB, known to positively regulate the size of the vascular bundles, was also found in the EUCAWOOD set [72].

#### *Core xylem genes*

Expression profiling has been used in several studies to report sets of genes differentially expressed during xylem development, notably in arabidopsis [46,73,74]. Comparison of the EUCAWOOD unigenes with sets of genes expressed during xylem differentiation in arabidopsis, revealed four candidate genes common to all the above-mentioned studies. They encode IRX9 (At2g37090; a GT family 43), COBL4/IRX6 (At5g15630; a COBRA-like protein), IRX8 (At5g54690; a GT family 8) as well as a protein of unknown function (At4g27435). These four genes belong to a group of 52 arabidopsis genes defined by Ko and collaborators as 'core xylem-specific genes' in their comparative transcriptome analysis [74].

#### In silico *identification of simple sequence repeat (SSR) markers*

Genomic SSR markers or microsatellites have already been developed in *Eucalyptus* species [75,76], however, to the best of our knowledge, only one very recent paper was dedicated to EST-SSRs [77]. To mine the EUCAWOOD dataset for EST-SSRs, we looked for di- and tri-nucleotide repeats stretching for at least 12 nt and also tetra- to hexanucleotides repeated at least three times. A total of 639 putative microsatellites were thus found in 512 EUCAWOOD unigenes (Additional file 9). That is, 13.3% of the EUCAWOOD unigenes contain at least one putative SSR. This agrees with the frequency of SSR-ESTs found in other dicotyledonous species, which ranges from 2.65–16.82% [78].

Tri-nucleotide repeats (TNRs) were the most abundant motifs (46.3% of the total 639 SSRs), followed by di-nucleotide repeats (DNRs, 29.4%). This is consistent with most similar studies of monocots as well as dicots [78,79]. Among the TNRs, the most abundant motifs were AAG/AGA/GAA/CTT/TTC/TCT (96 EST-SSRs) representing 32.3% of TNRs and 14.9% of all SSRs. The DNR the most represented was AG/GA/CT/TC (165 EST-SSRs), which accounted for 87.8% of all DNRs and 25.9% of all SSRs. These motifs have also been found to be the predominant DNRs and TNRs among the EST-SSRs in more than 20 plant species [78,79].

#### *The EUCAWOOD database*

EUCAWOOD [26] is a MySQL database allowing four types of queries through a web interface consisting of check boxes and pull-down menus. Query 1 is a library filter query allowing retrieval of all unigenes or a selection of them from the user-specified libraries. EST assembly, Blast hits against several databases (Uniref 100, CWN, MAIZEWALL ...), GO and PFAM annotations can also be retrieved. Query 2 retrieves unigenes by name (aliases), key words, PFAM or GO annotations, or hits in Blast (accession number or name). Query 3 allows Blast searches (blastn, tblastx, tblastn) for a user-specified sequence (or batch of sequences) in the EUCAWOOD database. Query 4 gives access to a tree view showing the number of unigenes by GO terms.

## Conclusion

We report the sequencing, assembly and annotation of approximately 10,000 ESTs derived from a normalised full-length secondary xylem cDNA library as well as subtractive libraries. Our data demonstrate the benefit for large-scale gene/EST discovery of using normalized libraries that minimize redundancies and increase the representation of the different genes expressed in a chosen tissue. They also illustrate the advantage of sequencing, in parallel, ESTs from subtracted libraries, which are enriched in clones not found in cDNA libraries and are a valuable source of new genes. The combination of a normalised secondary xylem library and subtractive libraries allowed us to assemble a large set of wood-related *Eucalyptus* unigenes, called EUCAWOOD, thus substantially increasing the representation of *Eucalyptus* ESTs available in public databases. The number of sequences available for this economically important genus has increased significantly during the past months [80-82] but is still low in comparison to other forest tree species such as poplar or pine. The major part of this new data set is composed of short sequences whose number is expected to increase dramatically in the future thanks to the development of the high-throughput '454' technology [81].

The EUCAWOOD dataset currently provides the most comprehensive list of unigenes dedicated to wood formation in the genus *Eucalyptus*. We have provided a public database supporting multiple queries that will be a particularly valuable resource for the correct annotation of genomic sequences and for the functional analysis of

genes and their products. The most immediate application of the EUCAWOOD unigene set reported in this study is the development of a wood reference microarray for *Eucalyptus*.

Finally, the EUCAWOOD dataset is also a valuable source of microsatellite markers as 639 EST-SSRs were identified from it. The usefulness of these EST-derived SSRs is superior to that of the genomic SSRs especially in looking for markers for important traits using the Gene Candidate approach. They are also usually more conserved and, therefore, may be easily transferred between species. The microsatellites reported for all these unigenes might be used to produce genetic maps, providing resources, for instance, for trait/gene association and candidate gene identification for wood quality traits.

## Methods

### *Normalization of a* Eucalyptus *secondary xylem cDNA library*

A library of directionally-cloned cDNAs prepared from the developing secondary xylem tissue of *Eucalyptus gunnii* was constructed in the λ ZapII vector (Stratagene, Amsterdam, The Netherlands) [17]. The library normalization process was based on the reassociation of an excess of cDNA inserts (driver DNA) to the cDNA library in the form of single-stranded circles (tracer DNA) as described by Bonaldo et al. [24]. A pBluescript SK vector carrying a *Homo sapiens* desmin cDNA (accession N° BC032116) was added at 1,000 copies to the initial library in order to assess the normalization efficiency. Single-stranded pBluescript phagemid DNA was generated *in vivo* from approximately $1.5 \times 10^6$ library clones and purified by hydroxyapatite (HAP) chromatography. Double-stranded driver DNA was generated by PCR from 1 ng of single-stranded library plasmid DNA with SK and T7 primers flanking the pBluescript vector multicloning site. PCR products were purified on a Qiagen Spin Column PCR Purification kit (Qiagen, Courtaboeuf, France) and eluted in TE buffer. Hybridization was performed by mixing 250 ng of single-stranded library phagemids with an excess of the PCR-amplified driver DNA and of each 3', 5' and oligo d(T20) blocking oligonucleotides. Hybridization was performed at 30°C for 24 h (C*o*t = 5). Single-stranded phagemids were purified by using HAP chromatography and converted to double strands by using SEQUENASE v2.0 DNA polymerase (USB, Staufen, Germany) and M13 primer. Double-stranded plasmids were electroporated into *Escherichia coli* DH10B cells (Invitrogen, Cergy Pontoise, France) and transformed cells were selected by growth on ampicillin.

### *Preparation of juvenile and mature secondary xylem RNAs*

Juvenile and mature secondary xylem samples were harvested from four-year-old and 10-year-old trees, respectively. Samples were collected from trees of a single *Eucalyptus globulus* genotype (clone vc9, RAIZ, Portugal). Tissue collection and RNA extraction were performed as described by Southerton et al. [83]. Remaining traces of DNA were removed with RQ1-RNAase-free DNAase (Promega, Madison, WI, USA) according to the manufacturer's procedure. RNA quality was checked by both agarose gel electrophoresis and spectrophotometry.

### *Construction and normalization of EST subtractive libraries*

The secondary xylem subtractive libraries were constructed by using the SSH technique [25]. SSH was performed with the PCR-Select cDNA Subtraction kit (Clontech Laboratories, Mountain View, CA, USA), according to the manufacturer's procedure. The subtracted PCR products generated by SSH were inserted into pGEM-T Easy Vector (Promega) and cloned into *E. coli* DH5α. Clones of recombinant bacteria were tested for complementation [84]. White colonies were picked with a BioPick robot (Genomic Solutions, Huntingdon, Cambridgeshire, UK) and arrayed in 384-well plates containing ampicillin (100 μg/ml)-supplemented LB freezing medium (25 g/l LB broth, 6.3 g/l $K_2HPO_4$, 1.8 g/l $KH_2PO_4$, 0.5 g/l sodium citrate, 1 g/l $MgSO_4$, 0.9 g/l ammonium sulfate, 4.4% glycerol). All recombinant clones were grown at 37°C overnight then stored at -80°C. High-density colony arrays (HDCA) were produced, hybridized and analyzed in order to eliminate false-positive clones. For this purpose all bacterial clones were spotted onto nylon membranes and hybridized with labeled SMART cDNAs from two independent juvenile and mature xylem probes as previously described [14,15]. ANOVA was performed on normalized data enabling us to keep 818 clones showing a significant relative expression level change (ratio of 1.2) between the two developmental stages.

### *Data processing and assembly*

Sequencing of the *Eucalyptus* secondary xylem cDNA library and SSH libraries was done at the Genoscope facilities (Centre National de Séquençage, Evry, France). Crossmatch software [85] was used to trim vector from the sequences. Subsequently, a home-made script was run to detect chimeras and remove low quality sequences. Sequences longer than 50 nucleotides and with a 'phred20 score' in at least 80% of the sequence were selected as good quality sequences suitable for assembly. The presence of poly A and poly T in the middle of the sequence was regarded as an indication of a chimeric sequence, which was then split in two and treated as two independent sequences. Good quality sequences were submitted to EMBL or GenBank according to the database curators' instructions.

Publicly available *Eucalyptus* ESTs and mRNA sequences were downloaded from the GenBank database at the NCBI server using the Entrez tool in March 2008. Wood-

related ESTs produced in our lab were eliminated to avoid redundancy. The FASTA files containing good quality wood-related sequences and the GenBank *Eucalyptus* sequences were combined and then assembled with CAP3 software [86] using default parameters.

#### *Sequence annotation*

The NCBI BLAST program version 2.2.6 was used to perform BLASTX similarity searches of various protein databases: UniRef100 was downloaded from the EBI ftp site [31] in March 2008; TAIR7 peptides were downloaded from The *Arabidopsis* Information Resource web site [28]; PoplarProteins1.1b JamboreeModels were downloaded from the Joint Genome Institute *Populus trichocarpa* v1.1 web site [29]; rice sequences were downloaded from the TIGR Rice Genome Annotation website [30]. The expectation (E)-value threshold used for BlastX searches was $10^{-10}$.

For functional characterization purposes, the Gene Ontology (GO) terms associated with the Uniref100 database were downloaded from the ftp site at the EBI [87]. The GO terms allocated to the best Uniref100 hit were assigned to the corresponding unigene. BLASTX searches (E value ≤ $10^{-10}$) were likewise conducted against the Cell Wall Navigator Database [38,39], the Maizewall database [40], the Plant Transcription Factor Database [60] and the Database of *Arabidopsis* Transcription Factors [61].

#### *Data storage*

A MySQL database was developed to store the raw ESTs, the good quality ESTs, the assembly results (contigs and singletons), the BlastX results as well as the GO and PFAM annotations [26]. Programs written in PHP and PERL languages were developed to load and export the data. The reports, figures and tables were built by querying the database using the SQL query language.

#### In silico *identification of simple sequence repeat (SSR) markers*

The MREPS software [88] was used to identify and localize the microsatellite motifs in the assembled EUCAWOOD unigene set. We looked for di- and tri-nucleotide repeats stretching for at least 12 nucleotides and also tetra- to hexa-nucleotides repeated at least three times.

## List of the abbreviations used

ABC: ATP-binding cassette; CesA: cellulose synthase; Cg: contig; CWN: Cell Wall Navigator database; DATF: Database of *Arabidopsis* Transcription Factors; DNR: di-nucleotide repeat; EST: expressed sequence tag; FLA: fasciclin-like arabinogalactan protein; GH: glycosyl hydrolase; GO: gene ontology; GT: glycosyl transferase; HCT: hydroxycinnamoyl-CoA:shikinimate/quinate hydroxycinnamoyltransferase; ID: identifier; Jm: *Eucalyptus globulus* juvenile vs mature secondary xylem SSH library; Mbp: million base pairs; Mj: *Eucalyptus globulus* mature vs juvenile secondary xylem SSH library; nt: nucleotide; PTFD: Plant Transcription Factor Database; Sg: singleton; SSH: suppression subtractive hybridisation; SSR: simple sequence repeat; SuSy: sucrose synthase; TNR: tri-nucleotide repeat; XET: xyloglucan endotransglycosylases; Xl: *Eucalyptus gunnii* secondary xylem *vs* mature leaf SSH library; Xp: *Eucalyptus gunnii* secondary xylem *vs* secondary phloem SSH library; $Xyl_{cDNA}$: *Eucalyptus gunnii* secondary xylem full-length cDNA library.

## Competing interests

The authors declare that they have no competing interests.

## Authors' contributions

DR designed bioinformatics approaches, analyzed the sequence and bioinformatics data, sorted the $M_j$ and $J_m$ SSH libraries and wrote the manuscript; HSC designed and created the EUCAWOOD database and its interface; HSC and FS designed bioinformatics approaches and provided bioinformatics tools, pipelines and scripts; NL normalized the $Xyl_{cDNA}$ library and participated in managing and sequencing all the libraries described in the manuscript; EP built the $M_j$ and $J_m$ SSH libraries and contributed to discussions on the manuscript; PW and AC sequenced the libraries; PS and JGP contributed to writing the manuscript as well to discussions on its content; JGP supervised and coordinated the project. All authors read and approved the manuscript.

## Additional material

### Additional file 1

***EUCAWOOD unigenes annotation.*** *Data concerning all the 3,857 EUCAWOOD unigenes are shown, including: the wood-related ESTs in every unigene; the best BlastX hit for each unigene (E value ≤ $e^{-10}$) in all the searched protein databases (*i.e. *Uniref100, JGI poplar proteins v1.1,* Arabidopsis *TAIR7 peptides, TIGR Rice genome annotation, NCBI grapevine), as well as the GO and PFAM annotation terms and IDs allocated to the first Uniref100 hit of each unigene.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-9-36-S1.xls]

### Additional file 2

***EUCAWOOD unigenes matching poplar and/or grapevine but not* Arabidopsis *or rice sequences.*** *Lists of EUCAWOOD unigenes matching poplar, grapevine or both, but not* Arabidopsis *or rice sequences.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-9-36-S2.xls]

### Additional file 3

***GO terms and ID assignments to EUCAWOOD unigenes.*** *GO terms and PFAM ID assignments to EUCAWOOD unigenes.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-9-36-S3.xls]

## Additional file 4

***PFAM assignments for EUCAWOOD unigenes. Comparison of Jm and Mj libraries.*** *PFAM assignments for EUCAWOOD unigenes. Comparison of Jm and Mj libraries.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-9-36-S4.xls]

## Additional file 5

***EUCAWOOD 'No Hits'.*** *Unigenes are shown that have no matches in the Uniref100, poplar, grapevine,* Arabidopsis *or rice protein databases.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-9-36-S5.xls]

## Additional file 6

***Identification of cell wall related genes in EUCAWOOD.*** *EUCAWOOD unigenes presenting BLASTX hits against the Cell Wall Navigator [38,39] and/or MAIZEWALL [40] databases are shown with the best hit in either database (E value ≤ $e^{-10}$).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-9-36-S6.xls]

## Additional file 7

***EUCAWOOD lignification toolbox.*** *EUCAWOOD unigenes were mined for homologous genes in the* Arabidopsis *lignification toolbox [41]. The table shows the best hit of each unigene against Uniref100 and TAIR7 Peptides databases. For every hit, the corresponding alignment score (E value ≤ $e^{-10}$) is shown.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-9-36-S7.xls]

## Additional file 8

***Transcription factors among EUCAWOOD unigenes.*** *EUCAWOOD unigenes with BLASTX hits in the* Arabidopsis *Transcription Factor database (TFD) or in the Plant Transcription Factors database (PDFD) are shown, along with the best hit in either database, and the corresponding alignment score (E value ≤ $e^{-10}$).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-9-36-S8.xls]

## Additional file 9

***EUCAWOOD putative SSRs or microsatellites.*** In silico *identification of simple sequence repeat (SSR) markers in the EUCAWOOD database.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-9-36-S9.xls]

## Acknowledgements

The authors are grateful to Cristina Marques and Victor Carocha (RAIZ) for the gift of the juvenile and mature wood samples, to Jean Marc Frigerio for useful advice and help with assembly procedures and microsatellites analysis, to Christian Brière for help with statistical tests and to Christophe Plomion (INRA, Pierroton) for coordinating the Forest project. The authors also wish to thank the Bioinformatic platform "GenoToul Midi-Pyrénées" http://www.bioinfo.genotoul.fr for providing calculation facilities for Blasts, GO and Pfam searches). RAIZ and ENCE provided financial support for the construction of the SSH libraries. DR was supported by an EU Marie Curie Intra-European Fellowship.

## References

1. Boudet AM, Kajita S, Grima-Pettenati J, Goffner D: **Lignins and lignocellulosics: a better control of synthesis for new and improved uses.** *Trends Plant Sci* 2003, **8:**576-581.
2. Plomion C, Leprovost G, Stokes A: **Wood formation in trees.** *Plant Physiol* 2001, **127:**1513-1523.
3. Zobel B, Sprague J: *Juvenile wood in forest trees* Berlin: Springer-Verlag; 1998.
4. Myburg AA, Potts BM, Marques CMP, Kirst M, Gion J-M, Grattapaglia D, Grima-Pettenati J: **Eucalyptus.** In *Genome Mapping & Molecular Breeding in Plants. Forest Trees Volume 7*. Edited by: Kole CR. Heidelberg: Springer; 2007. ISBN: 978-3-540-34540-4
5. **Food and Agriculture Organization of the United Nations (Global Forest Resources Assessment, 2005)** [http://www.fao.org/docrep/008/a0400e/a0400e0400.htm]
6. **EUCAGEN** [http://www.ieugc.up.ac.za]
7. Hertzberg M, Aspeborg H, Schrader J, Andersson A, Erlandsson R, Blomqvist K, Bhalerao R, Uhlen M, Teeri TT, Lundeberg J, Sundberg B, Nilsson P, Sandberg G: **A transcriptional roadmap to wood formation.** *Proc Natl Acad Sci USA* 2001, **98:**14732-14737.
8. Demura T, Fukuda H: **Transcriptional regulation in wood formation.** *Trends Plant Sci* 2007, **12:**64-70.
9. Allona I, Quinn M, Shoop E, Swope K, Cyr S, Carlis J, Riedl J, Retzel E, Campbell MM, Sederoff R, Whetten RW: **Analysis of xylem formation in pine by cDNA sequencing.** *Proc Natl Acad Sci USA* 1998, **95:**9693-9698.
10. Whetten R, Sun Y-H, Zhang Y, Sederoff R: **Functional genomics and cell wall biosynthesis in loblolly pine.** *Plant Mol Biol* 2001, **47:**275-291.
11. Lorenz WW, Dean JF: **SAGE profiling and demonstration of differential gene expression along the axial developmental.** *Tree Physiol* 2002, **22:**301-10.
12. Sterky F, Regan S, Karlsson J, Hertzberg M, Rohde A, Holmberg A, Amini B, Bhalerao R, Larsson M, Villarroel R, Van Montagu M, Sandberg G, Olssoni O, Teeri TT, Boerjan W, Gustafsson P, Uhlen M, Sundberg B, Lundeberg J: **Gene discovery in the wood-forming tissues of poplar: Analysis of 5,692 expressed sequence tags.** *Proc Natl Acad Sci USA* 1998, **95:**13330-13335.
13. Pavy N, Paule C, Parsons L, Crow JA, Morency MJ, Cooke J, Johnson JE, Noumen E, Guillet-Claude C, Butterfield Y, Barber S, Yang G, Liu J, Stott J, Kirkpatrick R, Siddiqui A, Holt R, Marra M, Seguin A, Retzel E, Bousquet J, MacKay J: **Generation, annotation, analysis and database integration of 16,500 white spruce EST clusters.** *BMC Genomics* 2005, **6:**144.
14. Foucart C, Paux E, Ladouce N, San-Clemente H, Grima-Pettenati J, Sivadon P: **Transcript profiling of a xylem *vs* phloem cDNA subtractive library identifies new genes expressed during xylogenesis in *Eucalyptus*.** *New Phytol* 2006, **170:**739-752.
15. Paux E, Tamasloukht M, Ladouce N, Sivadon P, Grima-Pettenati J: **Identification of genes preferentially expressed during wood formation in Eucalyptus.** *Plant Mol Biol* 2004, **55:**263-280.
16. Paux E, Carocha V, Marques C, Mendes de Sousa A, Borralho N, Sivadon P, Grima-Pettenati J: **Transcript profiling of *Eucalyptus* xylem genes during tension wood formation.** *New Phytol* 2005, **167:**89-100.
17. Poeydomenge O, Boudet AM, Grima-Pettenati J: **A cDNA encoding S-adenosyl-L-methionine:caffeic acid 3-O-methyltransferase from *Eucalyptus*.** *Plant Physiol* 1994, **105:**749-750.
18. **GENOSCOPE Eucalyptus EST sequencing project** [http://wwwgenoscopecnsfr/externe/Francais/Projets/Projet_LG/organisme_LGhtml]
19. Lacombe E, Hawkins S, Van Doorsselaere J, Piquemal J, Goffner D, Poeydomenge O, Boudet AM, Grima-Pettenati J: **Cinnamoyl CoA reductase, the first committed enzyme of the lignin branch biosynthetic pathway: Cloning, expression and phylogenetic relationships.** *Plant J* 1997, **11:**429-441.
20. Feuillet C, Boudet AM, Grima-Pettenati J: **Nucleotide sequence of a cDNA encoding cinnamyl alcohol dehydrogenase from *Eucalyptus*.** *Plant Physiol* 1993, **103:**1447.
21. Gion J-M, Rech P, Grima-Pettenati J, Verhaegen D, Plomion C: **Mapping candidate genes in Eucalyptus with emphasis on lignification genes.** *Mol Breed* 2000, **6:**441-449.

22. Goicoechea M, Lacombe E, Legay S, Mihaljevic S, Rech P, Jauneau A, Lapierre C, Pollet B, Verhaegen D, Chaubet-Gigot N, *et al.*: ***Eg* MYB2, a new transcriptional activator from *Eucalyptus* xylem, regulates secondary cell wall formation and lignin biosynthesis.** *Plant J* 2005, **43:**553-567.
23. Legay S, Lacombe E, Goicoechea M, Briere C, Seguin A, Mackay J, Grima-Pettenati J: **Molecular characterization of *Eg* MYB1, a putative transcriptional repressor of the lignin biosynthetic pathway.** *Plant Sci* 2007, **173:**542-549.
24. Bonaldo M, Lennon G, Soares M: **Normalization and subtraction: two approaches to facilitate gene discovery.** *Genome Res* 1996, **6:**791-806.
25. Diatchenko L, Lau Y-FC, Campbell AP, Chenchik A, Moqadam F, Huang B, Lukyanov S, Lukyanov K, Gurskaya N, Sverdlov ED, Siebert PD: **Suppression subtractive hybridization: A method for generating differentially regulated or tissue-specific cDNA probes and libraries.** *Proc Natl Acad Sci USA* 1996, **93:**6025-6030.
26. **EUCAWOOD** [http://www.polebio.scsv.ups-tlse.fr/eucalyptus/eucawood]
27. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D: **Gapped BLAST and PSI-BLAST: A new generation of protein database search programs.** *Nucl Acids Res* 1997, **25:**3389-3402.
28. **The *Arabidopsis* Information Resource peptides (TAIR7 peptides)** [ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR7_genome_release/TAIR7_blastsets/TAIR7_pep_20070425]
29. **JGI *Populus tricocharpa* proteins annotation v1.1** [http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.download.ftp.html]
30. **TIGR Rice Genome Annotation** [http://www.tigr.org/tdb/e2k1/osa1/data_download.shtml]
31. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyère C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pè ME, Valle G, Morgante M, Caboche M, Adam-Blondon AF, Weissenbach J, Quétier F, Wincker P: **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.** *Nature* 2007, **449:**463-467.
32. **Uniref100 ftp** [ftp://ftp.ebi.ac.uk/pub/databases/uniprot/uniref/uniref100/]
33. **Gene Ontology** [http://www.geneontology.org]
34. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer ELL, Bateman A: **The Pfam protein families database.** *Nucl Acids Res* 2008, **36:**D281-D288.
35. **Pfam database** [http://Pfam.sanger.ac.uk/]
36. Vettore AL, da Silva FR, Kemper EL, Souza GM, da Silva AM, Ferro MI, Henrique-Silva F, Giglioti EA, Lemos MV, Coutinho LL, Nobrega MP, Carrer H, França SC, Bacci Júnior M, Goldman MH, Gomes SL, Nunes LR, Camargo LE, Siqueira WJ, Van Sluys MA, Thiemann OH, Kuramae EE, Santelli RV, Marino CL, Targon ML, Ferro JA, Silveira HC, Marini DC, Lemos EG, Monteiro-Vitorello CB, Tambor JH, Carraro DM, Roberto PG, Martins VG, Goldman GH, de Oliveira RC, Truffi D, Colombo CA, Rossi M, de Araujo PG, Sculaccio SA, Angella A, Lima MM, de Rosa Júnior VE, Siviero F, Coscrato VE, Machado MA, Grivet L, Di Mauro SM, Nobrega FG, Menck CF, Braga MD, Telles GP, Cara FA, Pedrosa G, Meidanis J, Arruda P: **Analysis and functional annotation of an Expressed Sequence Tag collection for tropical crop sugarcane.** *Genome Res* 2003, **13:**2725-2735.
37. Mellerowicz EJ, Baucher M, Sundberg B, Boerjan W: **Unravelling cell wall formation in the woody dicot stem.** *Plant Molecular Biology* 2001, **47:**239-274.
38. Girke T, Lauricha J, Tran H, Keegstra K, Raikhel N: **The Cell Wall Navigator database. A systems-based approach to organism-unrestricted mining of protein families involved in cell wall metabolism.** *Plant Physiol* 2004, **136:**3003-3008.
39. **Cell Wall Navigator database** [http://bioweb.ucr.edu/Cellwall/index.pl]
40. Guillaumie S, San-Clemente H, Deswarte C, Martinez Y, Lapierre C, Murigneux A, Barriere Y, Pichon M, Goffner D: **MAIZEWALL. Database and developmental gene expression profiling of cell wall biosynthesis and assembly in maize.** *Plant Physiol* 2007, **143:**339-363 [http://www.polebio.scsv.ups-tlse.fr/MAIZEWALL/index.html].
41. Raes J, Rohde A, Christensen JH, Peer Y Van de, Boerjan W: **Genome-Wide Characterization of the Lignification Toolbox in *Arabidopsis*.** *Plant Physiol* 2003, **133:**1051-1071.
42. Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Dejardin A, Depamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjarvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leple JC, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouze P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Peer Y Van de, Rokhsar D: **The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray).** *Science* 2006, **313:**1596-1604.
43. Samuels L, Rensing K, Douglas C, Mansfield S, Dharmawardhana P, Ellis B: **Cellular machinery of wood production: differentiation of secondary xylem in *Pinus contorta* var. latifolia.** *Planta* 2002, **216:**72-82.
44. Boerjan W, Ralph J, Baucher M: **Lignin biosynthesis.** *Ann Rev Plant Biol* 2003, **54:**519-546.
45. Liang M, Davis E, Gardner D, Cai X, Wu Y: **Involvement of *At* LAC15 in lignin synthesis in seeds and in root elongation of *Arabidopsis*.** *Planta* 2006, **224:**1185-1196.
46. Brown DM, Zeef LAH, Ellis J, Goodacre R, Turner SR: **Identification of novel genes in *Arabidopsis* involved in secondary cell wall formation using expression profiling and reverse genetics.** *Plant Cell* 2005, **17:**2281-2295.
47. Yang C, Xu Z, Song J, Conner K, Vizcay Barrena G, Wilson ZA: ***Arabidopsis* MYB26/MALE STERILE35 regulates secondary thickening in the endothecium and is essential for anther dehiscence.** *Plant Cell* 2007, **19:**534-548.
48. Bayer E, Bottrill A, Walshaw J, Vigouroux M, Naldrett M, Thomas C, Maule A: **Cell wall proteome defined using multidimensional protein identification technology.** *Proteomics* 2006, **6:**301-311.
49. Joshi CP, Bhandari S, Ranjan P, Kalluri UC, Liang X, Fujino T, Samuga A: **Genomics of cellulose biosynthesis in poplars.** *New Phytol* 2004, **164:**53-61.
50. Gardiner JC, Taylor NG, Turner SR: **Control of cellulose synthase complex localization in developing xylem.** *Plant Cell* 2003, **15:**1740-1748.
51. Bieniawska Z, Paul Barratt DH, Garlick AP, Thole V, Kruger NJ, Martin C, Zrenner R, Smith AM: **Analysis of the sucrose synthase gene family in *Arabidopsis*.** *Plant J* 2007, **49:**810-828.
52. Ranik M, Myburg A: **Six new cellulose synthase genes from *Eucalyptus* are associated with primary and secondary cell wall biosynthesis.** *Tree Physiol* 2006, **26:**545-556.
53. Szyjanowicz PMJ, McKinnon I, Taylor NG, Gardiner J, Jarvis MC, Turner SR: **The irregular xylem 2 mutant is an allele of KORRIGAN that affects the secondary cell wall of *Arabidopsis thaliana*.** *Plant J* 2004, **37:**730-740.
54. Geisler-Lee J, Geisler M, Coutinho PM, Segerman B, Nishikubo N, Takahashi J, Aspeborg H, Djerbi S, Master E, Andersson-Gunneråas S, Sundberg B, Karpinski S, Teeri TT, Kleczkowski LA, Henrissat B, Mellerowicz EJ: **Poplar carbohydrate-active enzymes. Gene identification and expression analyses.** *Plant Physiol* 2006, **140:**946-962.
55. Aspeborg H, Schrader J, Coutinho PM, Stam M, Kallas A, Djerbi S, Nilsson P, Denman S, Amini B, Sterky F, Master F, Sandberg G, Mellerowicz E, Sundberg B, Henrissat B, Teeri TT: **Carbohydrate-active enzymes involved in the secondary cell wall biogenesis in hybrid aspen.** *Plant Physiol* 2005, **137:**983-997.
56. Zhong R, Kays SJ, Schroeder BP, Ye Z-H: **Mutation of a chitinase-like gene causes ectopic deposition of lignin, aberrant cell shapes, and overproduction of ethylene.** *Plant Cell* 2002, **14:**165-179.

57. Cosgrove DJ: **Enzymes and other agents that enhance cell wall extensibility.** *Ann Rev Plant Physiol and Plant Mol Biol* 1999, **50:**391-417.
58. Rose JKC, Braam J, Fry SC, Nishitani K: **The XTH family of enzymes involved in xyloglucan endotransglucosylation and endohydrolysis: Current perspectives and a new unifying nomenclature.** *Plant Cell Physiol* 2002, **43:**1421-1435.
59. Lafarguette F, Leple J-C, Dejardin A, Laurans F, Costa G, Lesage-Descauses M-C, Pilate G: **Poplar genes encoding fasciclin-like arabinogalactan proteins are highly expressed in tension wood.** *New Phytol* 2004, **164:**107-121.
60. **Plant Transcription Factor Database** [http://plntfdb.bio.uni-potsdam.de/v1.0/]
61. **Database of *Arabidopsis* Transcription Factors** [http://datf.cbi.pku.edu.cn/]
62. Rogers LA, Campbell MM: **The genetic control of lignin deposition during plant growth and development.** *New Phytol* 2004, **164:**17-30.
63. Bomal JC, Bedon F, Caron S, Mansfield SD, Levasseur C, Cooke J, Blais S, Tremblay L, Morency MJ, Pavy N, Grima-Pettenati J, Séguin A, MacKay J: **Involvement of *Pinus taeda MYB1* and *MYB8* in phenylpropanoid metabolism and secondary cell wall biogenesis: a comparative *in planta* analysis.** *J Exp Bot* 2008, **59:**3925-3939.
64. Mitsuda N, Seki M, Shinozaki K, Ohme-Takagi M: **The NAC transcription factors NST1 and NST2 of *Arabidopsis* regulate secondary wall thickenings and are required for anther dehiscence.** *Plant Cell* 2005, **17:**2993-3006.
65. Mitsuda N, Iwase A, Yamamoto H, Yoshida M, Seki M, Shinozaki K, Ohme-Takagi M: **NAC Transcription factors, NST1 and NST3, are key regulators of the formation of secondary walls in woody tissues of *Arabidopsis*.** *Plant Cell* 2007, **19:**270-280.
66. Ko J-H, Yang SH, Park AH, Lerouxel O, Han K-H: **ANAC012, a member of the plant-specific NAC transcription factor family, negatively regulates xylary fiber development in *Arabidopsis thaliana*.** *Plant J* 2007, **50:**1035-1048.
67. Zhong R, Demura T, Ye Z-H: **SND1, a NAC domain transcription factor, is a key regulator of secondary wall synthesis in fibers of *Arabidopsis*.** *Plant Cell* 2006, **18:**3158-3170.
68. Zhao C, Avci U, Grant EH, Haigler CH, Beers EP: **XND1, a member of the NAC domain family in Arabidopsis thaliana, negatively regulates lignocellulose synthesis and programmed cell death in xylem.** *Plant J* 2008, **53:**425-436.
69. Kawaoka A, Kaothien P, Yoshida K, Endo S, Yamada K, Ebinuma H: **Functional analysis of tobacco LIM protein *Nt* LIM1 involved in lignin biosynthesis.** *Plant J* 2000, **22:**289-301.
70. Kawaoka A, Nanto K, Ishii K, Ebinuma H: **Reduction of lignin content by suppression of expression of the LIM domain transcription factor *Eucalyptus camadulensis*.** *Silvae Genetica* 2006, **55:**269-277.
71. Scarpella E, Meijer AH: **Pattern formation in the vascular system of monocot and dicot plant species.** *New Phytol* 2004, **164:**209-242.
72. Prigge MJ, Otsuga D, Alonso JM, Ecker JR, Drews GN, Clark SE: **Class III homeodomain-leucine zipper gene family members have overlapping, antagonistic, and distinct roles in *Arabidopsis* development.** *Plant Cell* 2005, **17:**61-76.
73. Persson S, Wei H, Milne J, Page GP, Somerville CR: **Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets.** *Proc Natl Acad Sci USA* 2005, **102:**8633-8638.
74. Ko J-H, Han K-H, Park S, Yang J: **Plant body weight-induced secondary growth in *Arabidopsis* and its transcription phenotype revealed by whole-transcriptome profiling.** *Plant Physiol* 2004, **135:**1069-1083.
75. Brondani R, Williams E, Brondani C, Grattapaglia D: **A microsatellite-based consensus linkage map for species of Eucalyptus and a novel set of 230 microsatellite markers for the genus.** *BMC Plant Biol* 2006, **6:**20.
76. Brondani RPV, Brondani C, Tarchini R, Grattapaglia D: **Development, characterization and mapping of microsatellite markers in *Eucalyptus grandis* and *E. urophylla*.** *Theor Appl Genet* 1998, **97:**816-827.
77. Yasodha R, Sumathi R, Chezian P, Kavitha S, Ghosh M: **Eucalyptus microsatellites mined *in silico* : survey and evaluation.** *J Genet* 2008, **87:**21-25.
78. Kumpatla S, Mukhopadhyay S: **Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species.** *Genome* 2005, **48:**985.
79. Morgante M, Hanafey M, Powell W: **Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes.** *Nat Genet* 2002, **30:**194-200.
80. Costa da Cruz M, Gomes Caldas D, Tozelli Carneiro R, Moon DH, Salvatierra G, Franceschini LM, de Andrade A, Fiorani Celedon PA, Oda S, Labate CL: **SAGE transcript profiling of the juvenile cambial region of *Eucalyptus grandis*.** *Tree Physiol* 2008, **28:**905-919.
81. Novaes E, Drost DR, Farmerie WG, Pappas GJ, Grattapaglia D, Sederoff RR, Kirst M: **High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome.** *BMC Genomics* 2008, **9:**312.
82. Qiu D, Wilson LW, Gan S, Washusen R, Moran GF, Southerton SG: **Gene expression in *Eucalyptus* branch wood with marked variations in cellulose microfibril orientation and lacking G-layers.** *New Phytol* 2008, **179:**94-103.
83. Southerton SG, Marshall H, Mouradov A, Teasdale RD: **Eucalypt MADS-box genes expressed in developing flowers.** *Plant Physiol* 1998, **118:**365-372.
84. Sambrook J, Fritsch E, Maniatis T: **Molecular cloning: A laboratory manual.** 2nd edition. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1989.
85. **PHRAP** [http://www.phrap.org/phredphrapconsed.html]
86. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Res* 1999, **9:**868-877.
87. **Gene Ontology terms for Uniprot database** [ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/]
88. **MREPS** [http://bioinfo.lifl.fr/mreps/index.php]