A peer-reviewed version of this preprint was published in PeerJ on 7 June 2016.

View the peer-reviewed version (peerj.com/articles/2097), which is the preferred citable publication unless you specifically need to cite this preprint.

Transcriptome analysis of immature xylem in Chinese fir at different developmental phases

Yunxing Zhang, Xiaojiao Han, Jian Sang, Xuelian He, Mingying Liu, Guirong Qiao, Renying Zhuo, Guiping He, Jianjun Hu

**Background:** Chinese fir [*Cunninghamia lanceolata* (Lamb.) Hook.], is one of the most important native tree species for timber production in southern China. An understanding of overall fast growing stage, stem growth stage and senescence stage cambium transcriptome variation is lacking. We used transcriptome sequencing to identify the repertoire of genes expressed during development of xylem tissue in Chinese fir, aiming to delineate the molecular mechanisms of wood formation. **Results:** We carried out transcriptome sequencing at three different cultivation ages (7Y, 15Y and 21Y) generating 68.71 million reads (13.88 Gbp). A total of 140,486 unigenes with a mean size of 568.64 base pairs (bp) were obtained via *de novo* assembly. Of these, 27,427 unigenes (19.52%) were further annotated by comparison to public protein databases. A total of 5,331 (3.79%) unigenes were mapped into 118 pathways by searching against the Kyoto Encyclopedia of Genes and Genomes Pathway database (KEGG). Differentially expressed genes (DEG) analysis identified 3, 16 and 5,899 DEGs from the comparison of 7Y vs. 15Y, 7Y vs. 21Y and 15Y vs. 21Y, respectively, in the immature xylem tissues, including 2,638 significantly up-regulated and 3,280 significantly down-regulated genes. Besides, five NAC transcription factors, 190 MYB transcription factors, and 34 WRKY transcription factors were identified respectively from Chinese fir transcriptome. **Conclusion:** Our results revealed the active transcriptional pathways and identified the DEGs at different cultivation phases of Chinese fir wood formation. This transcriptome dataset will aid in understanding and carrying out future studies on the molecular basis of Chinese fir wood formation and contribute to future artificial production and applications.
Transcriptome analysis of immature xylem in Chinese fir at different developmental phases

Yunxing Zhang\textsuperscript{1,2,3}, Xiaojiao Han\textsuperscript{1,2}, Jian Sang\textsuperscript{1}, Xuelian He\textsuperscript{4}, Mingying Liu\textsuperscript{1,2}, Guirong Qiao\textsuperscript{1,2}, Renying Zhuo\textsuperscript{1,2,*}, Guiping He\textsuperscript{2,*}, Jianjun Hu\textsuperscript{3,*}

\textsuperscript{1} State Key Laboratory of Tree Genetics and Breeding, Chinese Academy of Forestry, Beijing, China
\textsuperscript{2} Key Laboratory of Tree Breeding of Zhejiang Province, The Research Institute of Subtropical of Forestry, Chinese Academy of Forestry, Hangzhou Zhejiang, China
\textsuperscript{3} Institute of Architectural and Artistic Design, Henan Polytechnic University, Jiaozuo Henan, China
\textsuperscript{4} Biotechnology Research Center of China Three Gorges University, Yichang Hubei, China
\textsuperscript{5} Key Laboratory of Tree Breeding and Cultivation of State Forestry Administration, Research Institute of Forestry, Chinese Academy of Forestry, Beijing, China

* Corresponding author
E-mail: zhuory@gmail.com (RYZ)

These authors contributed equally to the work
Abstract:

**Background:** Chinese fir (*Cunninghamia lanceolata* (Lamb.) Hook.), is one of the most important native tree species for timber production in southern China. An understanding of overall fast growing stage, stem growth stage and senescence stage cambium transcriptome variation is lacking. We used transcriptome sequencing to identify the repertoire of genes expressed during development of xylem tissue in Chinese fir, aiming to delineate the molecular mechanisms of wood formation.

**Results:** We carried out transcriptome sequencing at three different cultivation ages (7Y, 15Y and 21Y) generating 68.71 million reads (13.88 Gbp). A total of 140,486 unigenes with a mean size of 568.64 base pairs (bp) were obtained via *de novo* assembly. Of these, 27,427 unigenes (19.52%) were further annotated by comparison to public protein databases. A total of 5,331 (3.79%) unigenes were mapped into 118 pathways by searching against the Kyoto Encyclopedia of Genes and Genomes Pathway database (KEGG). Differentially expressed genes (DEG) analysis identified 3, 16 and 5,899 DEGs from the comparison of 7Y vs. 15Y, 7Y vs. 21Y and 15Y vs. 21Y, respectively, in the immature xylem tissues, including 2,638 significantly up-regulated and 3,280 significantly down-regulated genes. Besides, five NAC transcription factors, 190 MYB transcription factors, and 34 WRKY transcription factors were identified respectively from Chinese fir transcriptome.

**Conclusion:** Our results revealed the active transcriptional pathways and identified the DEGs at different cultivation phases of Chinese fir wood formation. This transcriptome dataset will aid in
understanding and carrying out future studies on the molecular basis of Chinese fir wood formation and contribute to future artificial production and applications.

**Key words**: Transcriptome, RNA-Seq, Chinese fir, xylem, wood formation

**Introduction**

Chinese fir \([Cunninghamia lanceolata](\text{Lamb.})\) \(\text{Hook.}\), a fast growing evergreen coniferous tree \((2n = 2x = 22)\), is one of the most important native tree species for timber production in southern China and is also distributed in Vietnam. It is the third most commonly planted tree species in plantations worldwide (Del Lungo, Ball et al. 2006). Due to its high value in terms of adaptability, growth rate, timber quality, versatility and commercial value, the planting area of Chinese fir in China is around 9.215 million ha, accounting 28.54\% of all forested land (Lei 2005, Shi, Zhen et al. 2010, Huang, Xu et al. 2012) and for 20–30\% of the total commercial timber production in China (Anonymous 1973, Orwa, Mutua et al. 2013). Chinese fir growth and development can be divided into three phases including fast growing stage, stem growth stage and senescence stage (Duan, Zhang et al. 2004).

Wood formation involves various division and differentiation activities of cambium cells, including vascular cambium activation, secondary xylem differentiation, cell expansion, secondary wall deposition, programmed cell death, and heartwood formation (Ye and Zhong 2015). Significant progress has been made in the past decade in uncovering the molecular players involved in the developmental phases of wood formation in tree species. \(\text{Populus trichocarpa}\) is the first sequencing tree (Tyler 2006), creates opportunities for investigation of secondary growth, and secondary xylem (wood) development in woody plants (Brunner, Busov...

In transcriptional network, secondary wall NAC, MYB and WRKY transcription factors act as the top-level and second-level master switches, respectively (Zhong and Ye 2014). These findings represent an important step toward elucidating the molecular mechanisms controlling wood formation. To date, genome sequences have been released for four tree species (Ye and Zhong 2015), including the angiosperms *P. trichocarpa* and *Eucalyptus grandis* (Myburg, Grattapaglia et al. 2014), and the gymnosperms *Picea abies* (Nystedt, Street et al. 2013) and *P. glauca* (Birol, Raymond et al. 2013). Furthermore, transcriptome sequences have been obtained, for *Acacia auriculiformis* (Wong, Cannon et al. 2011), *E. camaldulensis* (Thumma, Sharma et al. 2012), *Fraxinus spp.* (Bai, Rivera-Vega et al. 2011), *C. lanceolata* (Huang, Xu et al. 2012), *Larix leptolepis* (Zhang, Zhang et al. 2012), *Populus simonii* × *Populus nigra* via (Chen, Jiang et al. 2012), *P. spp.* (Raherison, Rigault et al. 2012), *Pinus monticola* (Liu, Sturrock et al. 2013), *P. glauca* (Raherison, Giguère et al. 2015). The availability of these genome and transcriptome sequences together with an improvement of the methodologies used for generation of transgenic trees will enable researchers to directly employ tree species as models for studying wood formation.

Studies of Chinese fir have mainly focused on anatomical, biochemical, cytological, physiology, and ecological aspects, with only few reports relating to molecular mechanism of wood formation. Due to the limited genomic sources of Chinese fir, it is important to explore

Transcriptome analyses (Huang, Xu et al. 2012, Qiu, Wan et al. 2013, Wang, Chen et al. 2013) have identified a number of candidate genes and transcription factors correlated with changes in wood formation in Chinese fir. In the previous studies, however, the samples were collected from the same year or during the same cultivation phase.

In the present work, we used RNA sequencing (RNA-seq) technology to characterize the transcriptome at different stages of Chinese fir growth and development. The RNA samples from three different growth and development phases were sequenced with the high-throughput Illumina deep sequencing technique. Based on the bioinformatics analysis of assembled transcriptome data, we characterized immature xylem transcriptional pathways during the different cultivation phases of Chinese fir. Furthermore, we identified the DEGs subject to regulation during xylem development. The transcriptome sequencing of Chinese fir immature xylem may help to discover new genes and pathways. The data will promote future genetic and genomics studies on the molecular mechanisms of wood formation, and contribute to future applications, including artificial wood production.

Materials and Methods

Plant materials
The samples of Chinese fir \textit{Cunninghamia lanceolata} (Lamb.) Hook.] were collected from three different sites in Kaihua Country Forest Farm (29°08' 33.56 N, 118°23' 56.59 E), Zhejiang Province. No specific permits were required from the Forest Farm to select samples. The Forest Farm is not privately-owned and the field studies did not involve protected species. The immature xylem tissues were collected from three trees at every three different cultivation phases [7 years, 15 years and 21 years of cultivation (7Y, 15Y and 21Y)]. The each phase, samples (outer glutinous 1-1.2 mm layer comprising early developing xylem tissue) were harvested from approximately breast height (1.0-1.20 m) on the main stem after removal of the bark using razor blades as described by Huang and Eshchar (Mizrachi, Hefer et al. 2010, Huang, Xu et al. 2012). All of the tissue samples were immediately frozen in liquid nitrogen and stored at -80°C for future use.

RNA extraction, library construction and RNA-seq

Experimental procedures including sample preparation and sequencing were performed following the standard protocols (Illumina, Inc.). Total RNA was extracted separately from each sample using the R6827-01Plant RNA Kit (Guduo, Shanghai, China). Three biological replicates were performed at three different cultivation phases. The concentration of RNA was analyzed using a spectrophotometer (UV-Vis Spectrophotometer, Quawell Q5000, San Jose, CA, USA), and the integrity of RNA was evaluated with an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Equal quantities of high-quality RNA from each sample were combined into a single large pool for cDNA synthesis. The mRNA-seq library was constructed using Illumina’s TruSeq RNA Sample Preparation Kit
(Illumina Inc, San Diego, CA, USA). The mRNA isolation, fragment interruption, cDNA synthesis, adapter ligation, PCR amplification and RNASeq were performed at Beijing BioMarker Technologies (Beijing, China). The poly-A mRNA was enriched using oligo (dT) magnetic beads, and the mRNA was broken into fragments by fragmentation buffer. The cleaved RNA fragments were transcribed into first-strand cDNA using random hexamer primers, followed by second strand cDNA synthesis using DNA polymerase I and RNase H. The short fragments were purified with the QiaQuick PCR Purification Kit (Qiagen) and eluted in EB buffer for end-repaired by addition of poly(A) to 3’. Then the suitable fragments were separated by an agarose gel electrophoresis and selected for PCR amplification as sequencing templates. The constructed mRNA-seq library was sequenced on the Illumina HiSeq™ 2500 sequencing platform.

**Sequence data analysis and assembly**

To obtain high-quality clean data for *de novo* assembly, the raw reads were filtered by removing the adapter sequences, low quality sequences (reads with ambiguous bases ‘N’), and reads in which more than 20% of bases had a Q-value <30. Reads were assembled using the reference transcriptome sequence of Chinese fir using the Bowtie and RSEM packages (Grabherr, Haas et al. 2011). The clean reads were assembled into contigs using Trinity (http://trinityrnaseq.sourceforge.net/) (Grabherr, Haas et al. 2011). After Trinity *de novo* assembly and correction, the contigs without any gaps were linked into transcripts according to the paired-end information of the sequences. Related contigs were clustered into transcripts based on nucleotide sequence identity. The longest transcripts were regarded as unigenes.
redundancies were removed. Finally, the unigenes were combined to produce the final assembly used for annotation. The unigenes expression abundance was represented in reads per kilobase of exon model per million mapped reads (RPKM). The RPKM measure of read density reflects the molar concentration of a transcript for RNA length and for the total read number in the measurement.

**Functional annotation**

To determine the functional annotation of the unigenes, the assembled sequences were compared against the NCBI Nr database (Deng, Li et al. 2006), SwissProt (Apweiler, Bairoch et al. 2004), GO (Ashburner, Ball et al. 2000), COG (Tatusov, Galperin et al. 2000), and KEGG (Kanehisa, Goto et al. 2004) with an E-value ≤10^{-5}. Gene names were assigned based on the best BLAST hit (Altschul, Madden et al. 1997). Open reading frames (ORFs) were predicted using the “GetORF” program (http://emboss.sourceforge.net/apps/cvs/emboss/apps/getorf.html). The longest ORF extracted from each unigene was defined as coding sequence (CDS), and the CDSs were translated into amino sequences using the standard codon table. The Blast2GO program was applied to obtain GO annotation of unigenes with an E-value ≤10^{-5} including molecular function, biological process, and cellular component categories. The unigenes sequences were aligned to COG database to classify and predict possible functions. Annotations of Chinese fir unigenes were used to predict biochemical pathways using the pathways tools. The KEGG database was used to analyze gene products related to metabolism and gene function in cellular processes.

**Detection of candidate SSR markers**
The assembled sequences longer than 1kb were used for the detection of SSR markers. Potential SSR markers were detected among the 17,902 unigenes using MISA software (http://pgrc.ipk-gatersleben.de/misa/). The parameters were set for the identification of perfect dinucleotide motifs with a minimum of six repeats, and tri-, tetra-, penta-, and hexa-nucleotide motifs with a minimum of five repeats (Zeng, Xiao et al. 2010, Wei, Qi et al. 2011).

Identification of differentially expressed genes

DESeq was performed to detect the genes which were differentially expressed, based on a threshold false discovery rate (FDR) < 0.01 and an absolute value log2ratio ≥2. If the FDR (Q = V/R) is required to remain below a cutoff (e.g. 0.01), then the FDR can be calculated according to the Benjamini and Hochberg algorithm as: FDR = E (Q) = E \{V/ (V + S)\} = E (V/R) (Benjamini and Hochberg 1995). All of the DEGs were used for the Nr, Swissport, GO, KEGG and COG Functional annotation analyses.

Sequence retrieval of transcription factors related to NAC, MYB and WRKY

Hidden Markov Model (HMM) was employed factors. The profiles of the NAC, MYB and WRKY DNA-binding domain PF01849, PF00249 and PF03106 used for the HMM search (HMMER 3.1, http://hmmer.janelia.org/) were downloaded from the Pfam database (http://pfam.sanger.ac.uk/), respectively. There were 5 NAC transcription factors, 190 MYB transcription factors and 34 WRKY transcription factors were obtained with an E-value threshold of 0.1. Ultimately, the expression levels of these transcription factors were identified according to the results of DEGs.

Sequence alignments and phylogenetic constructions of transcription factors
Alignment of the amino acid sequences of the NAC, MYB, WRKY transcription factor domains were aligned with Clustal X using the default parameters. For the phylogenetic analysis, the neighbor-joining trees was constructed by MEGA6.0. Bootstrap values obtained after 1000 replications are indicated on the branches.

Results

RNA-Seq and de novo transcriptome assembly

To obtain a global overview of the Chinese fir transcriptome at different developmental phases, nine RNA samples from immature xylem at three different cultivation stages (fast growing stage, stem growth stage and senescence stage) were sequenced with Illumina HiSeq™ 2500. After stringent quality assessment and data filtering, a total of 68.71 million reads and 13.88 gigabase pairs (Gbp) were generated (Table 1). The reads with base quality greater than 30 (Q≥30) and no ambiguous “N” were defined as high-quality reads.

Reads were mapped against the reference transcriptome sequences of Chinese fir. After the removal of adaptor sequences and exclusion of contaminated or short reads, 33,966,473 high-quality reads were assembled into 6,590,556 contigs (Table S1) using SOAPdenovo (Li, Zhu et al. 2010). Using the Trinity de novo assembly program, next-generation short-read sequences were assembled into 232,138 transcripts with mean length of 819.51 base pairs (bp) and N50 length of 1,635 bp. The transcripts were subjected to cluster and assembly analyses. Finally 140,486 unigenes with a mean size of 568.64 bp were obtained, these included 17,902 unigenes (12.74%) with length greater than 1kb. An overview of the contigs, transcripts and unigenes is shown in Table 2.
The length distributions of contigs, transcripts and unigenes were shown in Table 2. As expected for a randomly fragmented transcriptome, there was a positive relationship between the length of a given unigene and the number of reads assembled into it (Fig. 1). Open Reading Frame (ORF) prediction analysis performed with GetORF (http://emboss.sourceforge.net/apps/cvs/emboss/apps/getorf.html) identified 71,044 unigenes (50.57%) as having ORFs starting with an ‘ATG’ codon. The raw reads of Chinese fir produced in this study have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database (accession number: SRP059404).

Mapped read depth [reads per kilo base per million reads (RPKM)] was used as a metric for the expression of each unigenes. The expression of the unigenes varied similarly with sequencing depth (Table S2). The expression of unigenes ranged from 0 to 7,443.78 RPKM with an average of 9.32 RPKM. Unigenes with low RPKM values were removed, because they may not have been reliable due to low abundance or statistical errors. Of 71,102 unigenes remaining, 60,631 (85.27%) had a very low expression level of less than 10 RPKM. Unigenes with high RPKM values included those related to metabolism, cell wall biogenesis and remodeling, signal transduction and stress, such as laccase, poly-ubiquitin, ARF-L1 protein, thaumatin-like protein.

**Functional annotation and classification**

The reads of Chinese fir in different cultivation phases (7Y, 15Y and 21Y) were assembled, and several complementary approaches were utilized to annotate the assembled sequences (Table 3). Only 19.52% of the unigenes (27,427) were able to be annotated based on aligning with sequences deposited in diverse protein databases, including the National Center for
Biotechnology Information (NCBI) nonredundant protein (Nr) database, Cluster of Orthologous Groups of proteins (COG), Kyoto Encyclopedia of Genes and Genomes (KEGG), and UniProt/Swiss-Prot. According to the BLASTX results, 26,305 (18.72%) unigenes had homologous proteins in the Nr protein database. We found that for 29% of the unigenes the most similar proteins sequence was from *P. sitchensis*, whereas 13% were most similar to sequences from *Vitis vinifera*, and 5% to *Theobroma cacao* (Fig. 2).

Gene Ontology (GO) analysis was used for functional classification of the assembled transcripts and gene products in terms of their likely associated biological processes, cellular components, and molecular functions. There were 26,305 unigenes annotated in the Nr database, among which 15,085 unigenes were assigned one or more GO terms, with 37.2% in cellular components, 18.9% in molecular functions, and 43.9% in biological processes (Fig. 3). To better review GO cellular components, the GO terms were further clustered to their parent terms. For biological processes, genes involved in metabolic processes, cellular processes, and response to stimulus were highly represented. For molecular functions, catalytic activity, binding, and transporter activity were the most highly represented. For cellular components, the three top classifications were cell part, cell and organelle.

In addition, all unigenes were aligned to the COG database for further functional prediction and classification. Overall, 9,942 of the 140,486 sequences were assigned to 24 COG categories, including RNA processing and modification, chromatin structure and dynamics, energy production and conversion, cell cycle control, cell division, and chromosome partitioning (Fig. 4).

The category of general function prediction only represented the largest group (2,306; 17.32%),
followed by replication, recombination and repair (1,795; 13.48%), transcription (1,105; 8.30%).

Only a few unigenes were assigned to chromatin structure and dynamics, cell motility and nuclear structure (80, 38 and 2 unigenes, respectively). Furthermore 379 unigenes were assigned to cell wall/membrane/envelope biogenesis and 158 unigenes were assigned to cytoskeleton. No unigene was assigned to extracellular structures.

KEGG is a public database for networks of molecular interactions in cells and their variants specific to particular organisms. To further examine the usefulness of the Chinese fir unigenes generated in the present study, the unigenes were compared with the KEGG database using BLASTX and the corresponding pathways were established. Only 5,331 (3.79%) unigenes were assigned 118 pathways (Table S3). The pathways with highest unigene representation were Ribosome (ko03010, 213 unigenes, 3.69%), followed by RNA transport (ko03013, 195 unigenes, 3.38%) and Spliceosome (ko03040, 173 unigenes, 3.00%).

**SSR marker discovery**

SSRs can be used as powerful molecular markers for genetics, evolution and breeding studies. To explore SSR profiles in the unigenes of Chinese fir, the 17,902 unigene sequences were searched for SSRs. In total, 2,784 sequences containing 3,267 SSRs were obtained, with 394 unigene sequences containing more than one SSR. Tri-nucleotide repeat motifs (65.02%) were the most abundant, followed by Di-nucleotide repeats (31.53%) (Table 4). The most abundant repeat type was AAG/CTT (232, 18.24%), followed by AG/CT (166, 13.05%), and AT/AT (161, 12.66%).

**Identification of differentially expressed genes**
A total of 140,486 unigenes were detected from the clean reads of all three samples as described above. To detect DEGs between the samples harvested from trees at different stages, DESeq was used with the criteria FDR≤0.01 and log2Ratio≥2. Whereas only a few DEGs were identified for the comparisons of the samples from 7-year-old trees, by far the most DEGs were identified from the 15Y vs. 21Y comparisons (Table 5). The representative genes of up- and down-regulated DEGs in xylem of Chinese fir at different phases were shown in Fig. 5 and Table S4. For the 7Y vs. 21Y comparison, most DEGs were up-regulated. Whereas the 15Y vs. 21Y comparison produced a roughly similar number of up-regulated DEGs and down-regulated DEGs.

The up- and down-regulated DEGs were further analyzed based on GO component, GO function and GO process (Tables 6-8). The main GO component categories were cell part, cell, and organelle. In GO function ontology, the major classifications for the DEGs were catalytic activity, binding, and transporter activity. Most of the DEGs were classified into GO process categories of metabolic process, cellular process, and single-organism process. These results indicate that most of the DEGs were related to metabolism, cell wall biogenesis and remodeling, signal transduction and stress.

Transcription factors of interest

Recent studies have demonstrated that many transcription factors, such as NAC, MYB, and WRKY gene families, regulate the formation of secondary wall (Rogers and Campbell 2004, Kubo, Udagawa et al. 2005, Zhong and Ye 2009, Wang, Avci et al. 2010). Five NAC transcription factors, 190 MYB transcription factors, and 34 WRKY transcription factors were
identified respectively from Chinese fir transcriptome. Phylogenetic analyses of the three clusters of transcription factors were shown in the Figure S1-3. The transcription factors of NAC, MYB, WRKY were clustered into two, three, three main classes separately. The expression profiles of differentially expressed transcription factors were analyzed in different developmental phases (Fig. 6). The transcript levels of 32 MYB and 3 WRKY transcription factors increased from 7Y to 21Y, while 28 MYB and 4 WRKY exhibited a reduced expression profiles. We chose all of the NAC, WRKY transcription factors and 60 MYB transcription factors (full sequence score > 100) matching to the DEGs results, and discovered five WRKY transcription factors including three up-regulated DEGs (c100548.graph_c0, c49459.graph_c0, c97419.graph_c0) and two down-regulated DEGs (c48904.graph_c0, c70863.graph_c0). Furthermore, 17 MYB transcription factors were dug out including five up-regulated DEGs (c89668.graph_c0, c84067.graph_c0, c90780.graph_c0, etc.) and 12 down-regulated DEGs (c104123.graph_c0, c106541.graph_c0, c98343.graph_c0, etc.) in 15Y vs. 21Y (Table S5).

Discussion

RNA-seq has emerged to be a valuable tool to discover molecular markers and identify novel genes. In recent years, the growing number of species for which significant genetic resources are available is sparking a new era of plant genetic study (Morozova and Marra 2008, Coppe, Pujolar et al. 2010). In this study, RNA-seq technology was applied to the Chinese fir transcriptome using Illumina HiSeq™ 2500 platform, and the transcriptome at three different cultivation phases was systematically investigated.

A total of 68.71 million reads and 13.88 gigabase pairs (Gbp) were generated, and 140,486
unigenes with a mean size of 568.64 bp were obtained. About 19.52% of the unigenes were successfully annotated, and the information regarding putative nucleotide variations offers intriguing leads for the analysis of the transcriptome and biological functions in Chinese fir. Among the unigenes 29%, 13%, and 5% appeared to be most closely related to genes from *P. sitchensis*, *Vitis vinifera*, and *Theobroma cacao* respectively. Chinese fir and *P. sitchensis* gymnosperms assigned to the Coniferopsida Coniferae. Thus, there is a close relationship between *P. sitchensis* and Chinese fir based on both systematic botany and molecular analysis. This research moves us toward identifying candidate genes for wood formation and clarifying the functions of the relevant pathways in Chinese fir. Compared with those from other conifer trees, our results using samples from different cultivation phases identified a much larger number of unigenes. In addition, the mean length of unigenes (568 bp) that we obtained is much longer than those in previous studies using the same technology, which reported 449 bp (Huang, Xu et al. 2012), 505 bp (Wang, Chen et al. 2013), and 497 bp (Qiu, Wan et al. 2013). To the best of our knowledge, this study represents the first attempt at de novo sequencing and assembly of the Chinese fir transcriptome using RNA-seq, to focus on different cultivation phases including fast growing stage, stem growth stage and senescence stage respectively. The results obtained in this research demonstrated that our final assembly quality was satisfactory and it therefore provides sequence resources and facilitates further gene cloning and functional analyses.

Several genes encoding the biosynthesis of wood components (cellulose, xylan, glucomannan, and lignin), such as Cel/TDIF/CLE/PXY-WOX4/MYB (Plomion, Leprovost et al. 2001, Ito, Nakanomyo et al. 2006, Hirakawa, Shinohara et al. 2008, Etchells and Turner 2010, Hirakawa,
Kondo et al. 2010, Ji, Strable et al. 2010, Suer, Agusti et al. 2011, Wang, Soler et al. 2011), have
been identified in angiosperms (Jansson and Douglas 2007). Unfortunately, only few of the
related genes have been identified and functionally characterized in gymnosperms. In this study,
the lack of a reference genome for Chinese fir, hampered our efforts to determine gene and their
functions. BLAST hits Perhaps these unigenes might play an important roles in Chinese fir and
quite different from the other species.

Differences in gene expression profiles can yield insight into mechanisms underlying
physiological changes, and DEGs were found among different developmental stages, tissues,
treatments, and species (Zeng, Xiao et al. 2010, Logacheva, Kasianov et al. 2011, Wang, Xu et al.
2013, Wu, Austin et al. 2013, Wu, Li et al. 2014). The 7Y, 15Y and 21Y samples that we
collected represented the different cultivation development phases (fast growing stage, stem
growth stage and senescence stage), and there were vary a few DEGs in 7Y compared to either
15Y or 21Y. However, many up-regulated (2,623) and down-regulated (3,276) DEGs were
detected in 15Y compared to 21Y. Furthermore, most DEG unigenes in 15Y vs. 21Y were
annotated to specific pathways using the KEGG database, including metabolic pathways, and
biosynthesis of secondary metabolites among others. These results indicate considerable changes
of gene expression in immature xylem during the transition from stem growth stage and
senescence stage. A similar phenomenon was reported for the transition from the active stage to
the dormant stage in vascular cambium, in which expression of the core cell cycle genes in
vascular cambium correlated well with the cessation of cambial cell division (Brown, Zeef et al.
We can conclude that the key cycle protein transcripts expressed preferentially at stem growth stage and senescence stage of Chinese fir may be essential for wood formation, and these DEGs are involved in a broad range of physiological functions between stem growth stage and senescence stage.

References:


Supporting Information:

Table S1 Assembly contigs

Table S2 Assessment of assembly quality

Table S3 Pathway assignment based on KEGG

Table S4 the representative genes of up- and down-regulated DEGs in xylem of Chinese fir at different developmental phases

Table S5 MYB, WRKY transcriptome factors in DEGs

Fig. S1 Phylogenetic tree of NAC transcription factors

Fig. S2 Phylogenetic tree of MYB transcription factors

Fig. S3 Phylogenetic tree of WRKY transcription factors
Fig. legends:

Fig. 1. Dependence of unigene lengths on the number of reads assembled into that unigenes.

Fig. 2. Species distribution of the top BLAST hits in Nr database. Top BLASTX results for 26305 unigenes were calculated. Species with proportions of more than 1% are shown.

Fig. 3. Functional annotation of assembled sequences based on gene ontology (GO) categorization. GO analysis was summarized into three categories: cellular component, molecular function and biological process.

Fig. 4. Clusters of orthologous group (COG) classification

Fig. 5. Heatmap of the relative expression levels of differentially expressed genes

Fig. 6. The differential expression patterns of representative TFs in different developmental phases
Tables legends:

Table 1. Summary of Illumina transcriptome sequencing for Chinese fir

Table 2. Length distribution of assembled contigs, transcripts and unigenes.

Table 3 Functional annotation of Chinese fir unigenes

Table 4 Frequency of candidate SSRs in Chinese fir

Table 5 Number of up- and down-regulated DEGs in xylem of Chinese fir at different ages

Table 6 Up- and down-regulated Chinese fir DEGs by GO component ontology

Table 7 Up- and down-regulated Chinese fir DEGs by GO function ontology

Table 8 Up- and down-regulated Chinese fir DEGs by GO process ontology
Figure 1 Dependence of unigene lengths on the number of reads assembled into that unigenes
Figure 2 Species distribution of the top BLAST hits in Nr database
Figure 3 Functional annotation of assembled sequences based on gene ontology (GO) categorization
Figure 4 Clusters of orthologous group (COG) classification

COG Function Classification of Consensus Sequence

- A: RNA processing and modification
- B: Chromatin structure and dynamics
- C: Energy production and conversion
- D: Cell cycle control, cell division, chromosome partitioning
- E: Amino acid transport and metabolism
- F: Nucleotide transport and metabolism
- G: Carbohydrate transport and metabolism
- H: Coenzyme transport and metabolism
- I: Lipid transport and metabolism
- J: Translation, ribosomal structure and biogenesis
- K: Transcription
- L: Replication, recombination and repair
- M: Cell wall/membrane/envelope biogenesis
- N: Cell motility
- O: Posttranslational modification, protein turnover, chaperones
- P: Inorganic ion transport and metabolism
- Q: Secondary metabolites biosynthesis, transport and catabolism
- R: General function prediction only
- S: Function unknown
- T: Signal transduction mechanisms
- U: Intracellular trafficking, secretion, and vesicular transport
- V: Defense mechanisms
- W: Extracellular structures
- Y: Nuclear structure
- Z: Cytoskeleton
Figure 5 Heatmap of the relative expression levels of differentially expressed genes
Figure 6 The differential expression patterns of representative TFs in different developmental phases
Table 1. Summary of Illumina transcriptome sequencing for Chinese fir
Table 1. Summary of Illumina transcriptome sequencing for Chinese Fir

<table>
<thead>
<tr>
<th>Sample</th>
<th>Total Reads</th>
<th>Total Bases</th>
<th>GC (%)</th>
<th>Q30%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese fir</td>
<td>68 719 634</td>
<td>13 879 230 822</td>
<td>43.24</td>
<td>96.42</td>
</tr>
</tbody>
</table>
Table 2. Length distribution of assembled contigs, transcripts and unigenes.
Table 2. Length distribution of assembled contigs, transcripts and unigenes.

<table>
<thead>
<tr>
<th>Nucleotide length(bp)</th>
<th>Contigs</th>
<th>Transcripts</th>
<th>Unigenes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-300</td>
<td>6,508,442</td>
<td>87,745</td>
<td>69,385</td>
</tr>
<tr>
<td>300-500</td>
<td>42,221</td>
<td>51,541</td>
<td>35,452</td>
</tr>
<tr>
<td>500-1000</td>
<td>21,099</td>
<td>36,109</td>
<td>17,747</td>
</tr>
<tr>
<td>1000-2000</td>
<td>11,681</td>
<td>31,744</td>
<td>10,593</td>
</tr>
<tr>
<td>2000+</td>
<td>7,113</td>
<td>24,999</td>
<td>7,309</td>
</tr>
<tr>
<td>Total number</td>
<td>6,590,556</td>
<td>232,138</td>
<td>140,486</td>
</tr>
<tr>
<td>Total length</td>
<td>404,940,125</td>
<td>190,238,773</td>
<td>79,886,333</td>
</tr>
<tr>
<td>N50 length</td>
<td>64</td>
<td>1,635</td>
<td>882</td>
</tr>
<tr>
<td>Mean length</td>
<td>61.44</td>
<td>819.51</td>
<td>568.64</td>
</tr>
</tbody>
</table>
Table 3 (on next page)

Table 3 Functional annotation of Chinese fir unigenes
<table>
<thead>
<tr>
<th>Annotated databases</th>
<th>Unigenes</th>
<th>≥300 bp</th>
<th>≥1000 bp</th>
<th>Percentage of unigenes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nr-annotation</td>
<td>26,305</td>
<td>12,136</td>
<td>14,169</td>
<td>18.72%</td>
</tr>
<tr>
<td>COG-Annotation</td>
<td>9,942</td>
<td>3,797</td>
<td>6,145</td>
<td>7.08%</td>
</tr>
<tr>
<td>GO-Annotation</td>
<td>15,085</td>
<td>6,206</td>
<td>8,879</td>
<td>10.74%</td>
</tr>
<tr>
<td>KEGG-Annotation</td>
<td>5,331</td>
<td>1,854</td>
<td>3,477</td>
<td>3.79%</td>
</tr>
<tr>
<td>Swissprot-Annotation</td>
<td>17,407</td>
<td>7,276</td>
<td>10,131</td>
<td>12.39%</td>
</tr>
<tr>
<td>Total</td>
<td>27,427</td>
<td>12,899</td>
<td>14,528</td>
<td>19.52%</td>
</tr>
</tbody>
</table>
Table 4 Frequency of candidate SSRs in Chinese fir

<table>
<thead>
<tr>
<th>Candidate SSR</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSR1</td>
<td>0.05</td>
</tr>
<tr>
<td>SSR2</td>
<td>0.01</td>
</tr>
<tr>
<td>SSR3</td>
<td>0.03</td>
</tr>
<tr>
<td>SSR4</td>
<td>0.02</td>
</tr>
<tr>
<td>SSR5</td>
<td>0.04</td>
</tr>
</tbody>
</table>
# Table 4 Frequency of candidate SSRs in Chinese fir

<table>
<thead>
<tr>
<th>Motif</th>
<th>Repeat number</th>
<th>Total</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Di</td>
<td>-</td>
<td>210</td>
<td>73</td>
</tr>
<tr>
<td>Tri</td>
<td>518</td>
<td>196</td>
<td>94</td>
</tr>
<tr>
<td>Tetra</td>
<td>-</td>
<td>23</td>
<td>3</td>
</tr>
<tr>
<td>Penta</td>
<td>-</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Hexa</td>
<td>-</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>518</td>
<td>437</td>
<td>175</td>
</tr>
<tr>
<td>%</td>
<td>40.72</td>
<td>34.36</td>
<td>13.76</td>
</tr>
</tbody>
</table>
**Table 5** (on next page)

Table 5 Number of up- and down-regulated DEGs in xylem of Chinese fir at different ages
Table 5 Number of up- and down-regulated DEGs in xylem of Chinese fir at different ages

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Number of DEGs</th>
<th>Up</th>
<th>Down</th>
</tr>
</thead>
<tbody>
<tr>
<td>7Y vs. 15Y</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>7Y vs. 21Y</td>
<td>16</td>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>15Y vs. 21Y</td>
<td>5,899</td>
<td>2,623</td>
<td>3,276</td>
</tr>
</tbody>
</table>
Table 6 (on next page)

Table 6 Up- and down-regulated Chinese fir DEGs by GO component ontology
Table 6 Up- and down-regulated Chinese fir DEGs by GO component ontology

<table>
<thead>
<tr>
<th>GO component ontology</th>
<th>7Y vs. 15Y</th>
<th>7Y vs. 21Y</th>
<th>15Y vs. 21Y</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Up</td>
<td>Down</td>
<td>P-value</td>
</tr>
<tr>
<td>cell part</td>
<td>0</td>
<td>1</td>
<td>8.72E-01</td>
</tr>
<tr>
<td>cell</td>
<td>0</td>
<td>1</td>
<td>8.55E-01</td>
</tr>
<tr>
<td>organelle</td>
<td>0</td>
<td>1</td>
<td>7.36E-01</td>
</tr>
<tr>
<td>membrane</td>
<td>0</td>
<td>1</td>
<td>3.82E-01</td>
</tr>
<tr>
<td>organelle part</td>
<td>0</td>
<td>0</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>macromolecular complex</td>
<td>0</td>
<td>0</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>membrane part</td>
<td>0</td>
<td>0</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>cell junction</td>
<td>0</td>
<td>1</td>
<td>4.51E-02</td>
</tr>
<tr>
<td>extracellular region</td>
<td>0</td>
<td>1</td>
<td>6.24E-02</td>
</tr>
<tr>
<td>membrane-enclosed lumen</td>
<td>0</td>
<td>0</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>nucleoid</td>
<td>0</td>
<td>0</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>extracellular matrix</td>
<td>0</td>
<td>0</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>extracellular matrix part</td>
<td>0</td>
<td>0</td>
<td>1.00E+00</td>
</tr>
<tr>
<td>extracellular region part</td>
<td>0</td>
<td>0</td>
<td>1.00E+00</td>
</tr>
</tbody>
</table>
Table 7 (on next page)

Table 7 Up- and down-regulated Chinese fir DEGs by GO function ontology
## Table 7 Up- and down-regulated Chinese fir DEGs by GO function ontology

<table>
<thead>
<tr>
<th>GO function ontology</th>
<th>7Y vs. 15Y</th>
<th>7Y vs. 21Y</th>
<th>15Y vs. 21Y</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Up</td>
<td>Down</td>
<td>Up</td>
</tr>
<tr>
<td>catalytic activity</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>binding</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>transporter activity</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>structural molecule activity</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>nucleic acid binding transcription factor activity</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>electron carrier activity</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>molecular transducer activity</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>enzyme regulator activity</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>antioxidant activity</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>receptor activity</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>guanyl-nucleotide exchange factor activity</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>protein binding transcription factor activity</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>nutrient reservoir activity</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>metallochaperone activity</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 8 (on next page)

Table 8 Up- and down-regulated Chinese fir DEGs by GO process ontology
### Table 8 Up- and down-regulated Chinese fir DEGs by GO process ontology

<table>
<thead>
<tr>
<th>GO function ontology</th>
<th>7Y vs. 15Y</th>
<th>7Y vs. 21Y</th>
<th>15Y vs. 21Y</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Up</td>
<td>Down</td>
<td>Up</td>
</tr>
<tr>
<td>metabolic process</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>cellular process</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>single-organism process</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>response to stimulus</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>biological regulation</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>localization</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>cellular component organization or biogenesis</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>developmental process</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>multicellular organismal process</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>reproductive process</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>multi-organism process</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>signaling</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>growth</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>reproduction</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>immune system process</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>biological adhesion</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>rhythmic process</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>biological phase</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>locomotion</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>