Physiol Mol Biol Plants (July–September 2014) 20(3):393–397
DOI 10.1007/s12298-014-0230-x

SHORT COMMUNICATION

Identification and analysis of expressed sequence tags present in xylem tissues of kelampayan (*Neolamarckia cadamba* (Roxb.) Bosser)

Wei-Seng Ho · Shek-Ling Pang · Julaihi Abdullah

Received: 20 November 2013 / Revised: 16 February 2014 / Accepted: 3 April 2014 / Published online: 24 May 2014

Abstract The large-scale genomic resource for kelampayan was generated from a developing xylem cDNA library. A total of 6,622 high quality expressed sequence tags (ESTs) were generated through high-throughput 5' EST sequencing of cDNA clones. The ESTs were analyzed and assembled to generate 4,728 xylogenesis unigenes distributed in 2,100 contigs and 2,628 singletons. About 59.3 % of the ESTs were assigned with putative identifications whereas 40.7 % of the sequences showed no significant similarity to any sequences in GenBank. Interestingly, most genes involved in lignin biosynthesis and several other cell wall biosynthesis genes were identified in the kelampayan EST database. The identified genes in this study will be candidates for functional genomics and association genetic studies in kelampayan aiming at the production of high value forests.

Keywords cDNA library · Expressed sequence tags (EST) · Developing xylem · *Neolamarkia cadamba* · Wood formation · Lignin biosynthesis

Forest trees represent the majority of terrestrial biomass production and are a vital component of biodiversity. However, slow growing trees are unable to meet current global demand for wood, resulting in the loss and degradation of forests. Plantation forests of fast growing species have the potential to supply the bulk of wood needs on a long-term basis, and thus reduce the harvest pressure on natural forests for wood production to an acceptable level. In Sarawak, the state government has set a target of 1 million hectares for forest plantations to be established within 15 years. *Neolamarckia cadamba* (Roxb.) Bosser, locally known as kelampayan, has been identified as a promising fast growing species for planted forest development in Sarawak. Kelampayan is a large, deciduous and fast-growing tree species, thus with characteristics which guarantee early economic return within 8 to 10 years. Under normal conditions, it reaches a height of 17 m and a diameter of 25 cm at breast height (dbh) within 9 years. It is one of the best sources of raw material for the plywood industry, besides pulp and paper production. Kelampayan can also be used as a shade tree for dipterocarp line planting, whilst its leaves and bark have medical application (Joker 2000). The dried bark can be used to relieve fever and as a tonic, whereas a leaf extract can serve as a mouth wash (World Agroforestry Centre 2004).

Despite the high economic value of tropical wood, little is known about the genetic control of wood formation or xylogenesis for this species compared to loblolly pine (59,797 ESTs, Whetten et al. 2001), poplar (25,218 ESTs, Sterky et al. 2004) and spruce (16,500 ESTs, Pavy et al. 2005). Wood or secondary xylem is produced through the process of cell division, cell expansion and secondary cell-wall formation, the latter involving cellulose, hemicellulose, cell-wall proteins and lignin biosynthesis and deposition, and finally programmed cell death (Li et al. 2009). These processes are strongly interlinked, and the modulation of any one aspect may affect several others. Thus, the careful use of a functional genomics approach could rapidly provide information on the regulation of not just one gene, but of an entire pathway or several pathways at the same time. As of July, 2009, no kelampayan EST information was available in the NCBI GenBank. Therefore, we applied genomics approaches to explore the molecular basis of wood formation in

W.-S. Ho (✉)
Forest Genomics and Informatics Laboratory, Department of Molecular Biology, Faculty of Resource Science and Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
e-mail: wsho@frst.unimas.my

S.-L. Pang · J. Abdullah
Applied Forest Science and Industry Development Unit, Sarawak Forestry Corporation, Kuching, Sarawak, Malaysia

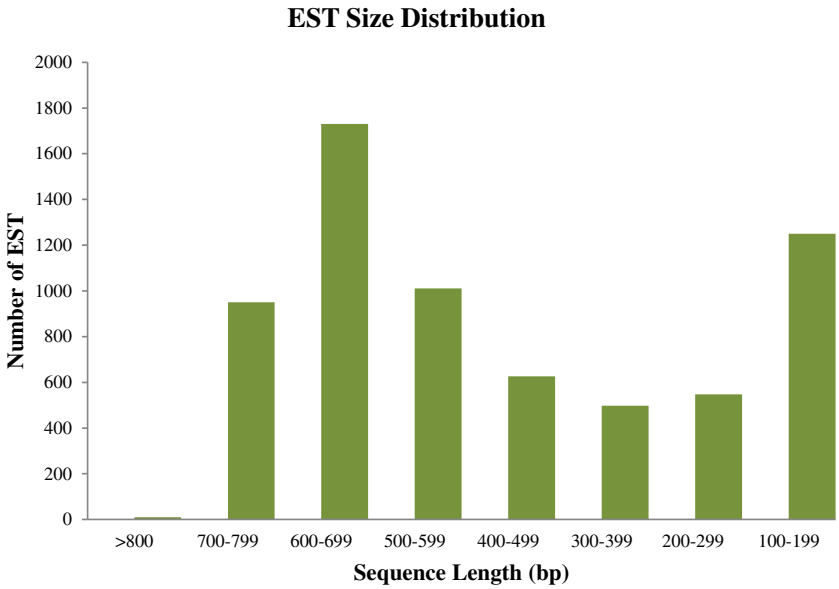

Fig. 1 Size distribution of the ESTs generated from the developing xylem tissues of kelampayan

kelampayan. Here we report on the generation and analysis of a genomic resource (expressed sequence tags, ESTs) for wood formation in kelampayan, via high-throughput DNA sequencing of cDNA clones derived from developing xylem tissues.

In order to investigate the molecular basis of wood formation in kelampayan, developing xylem tissues were collected by scraping a thin layer from the exposed xylem surface, at breast height (approximately 1.3 m above ground) and after removing the bark from a 2-year old kelampayan tree. The collected tissues were put in a clean plastic bag and immediately frozen in liquid nitrogen in the field, and then kept in -80 °C for later RNA isolation. Total RNA was extracted using RNeasy Midi Kit (Qiagen, Germany) with modification. Poly(A)$^+$ mRNA was isolated from the total RNA using using Micro-FastTrack^{TM} 2.0 Kits (Invitrogen, USA). A total of 105 μg total RNA was used for mRNA isolation. Purity and quality for both total RNA and mRNA were checked by agarose gel electrophoresis and spectrophotometry. The cDNA library was constructed using CloneMiner^{TM} cDNA Library Construction Kit (Invitrogen, USA) according to the manufacturer's protocol. About 0.6 μg mRNA was used as starting template for 1st strand cDNA synthesis. The *att*B1 adaptor was then ligated at the 5' end of the double-stranded cDNA. The cDNA was then subjected to size fractionation using cDNA Size Fractionation Columns supplied with the kit. A total of 80 ng size-fractionated cDNA and 250 ng pDONR^{TM} 222 plamid were used for BP recombination reaction. The cDNA was then transformed into ElectroMAX^{TM} DH10B T1 phage resistant cells using MicroPulser^{TM} electroporator (Bio-Rad, USA) and grown on LB-kanamycin agar plates overnight at 37 °C. cDNA clones were manually picked and cultured overnight with shaking in 96-well culture blocks. Glycerol stocks for each clone were prepared and kept in a duplicate 96-well plate format. All glycerol stocks were kept in -80 °C for later use. The titer of the cDNA library was 1.09 X 10^7 cfu, indicating that the cDNA library is comprehensive.

A total of 10,368 cDNA clones were randomly selected and employed in high-throughput plasmid preparation using a Montage Plasmid Miniprep96 and the MultiScreen Separation System (Milipore, USA). cDNA inserts were sequenced from the 5' end using an M13F primer and the ABI PRISM® Ready Reaction BigDye^{TM} Terminator Cycler Sequencing kit (Applied Biosystems, USA). High-throughput DNA sequencing was performed on an ABI 3730xl automated DNA Analyzer (Applied Biosystems, USA). Sequencing and bioinformatics analyses were undertaken at the Malaysia Genome Institute (MGI), UKM-MTDC. All the sequences were quality-checked before clustering and annotation. Raw ABI-formatted chromatogram reads were base-called using Phred (Ewing et al. 1998; Ewing and Green, 1998), with a threshold value of 20. Vector sequences were masked using Cross-Match. The trimming and removal of vectors, adaptors and low quality nucleotides was done using customized Perl scripts. Only high quality ESTs with a minimum of 100 bases and fewer than 4 % N (two or more peaks present at one position) were retained. High quality ESTs were matched against the NCBI non-redundant database using the blastx algorithm prior to clustering and assembling of the ESTs. Sequences with blastx E-value >10^{-10} were categorized as having no significant similarity. Multiple sequence alignment, clustering, assembly and the generation of consensus contigs was carried out with StackPACK (Miller et al. 1999). The StackPACK contains d2_cluster (Burke et al. 1999), PHRAP (Laboratory of PHIL GREEN, http://www.phrap.org) and CRAW (Chou and Burke, 1999). For d2_clustering, sequences were grouped together on presenting at least 96 % sequence similarity in any window of 150 bases. Loose

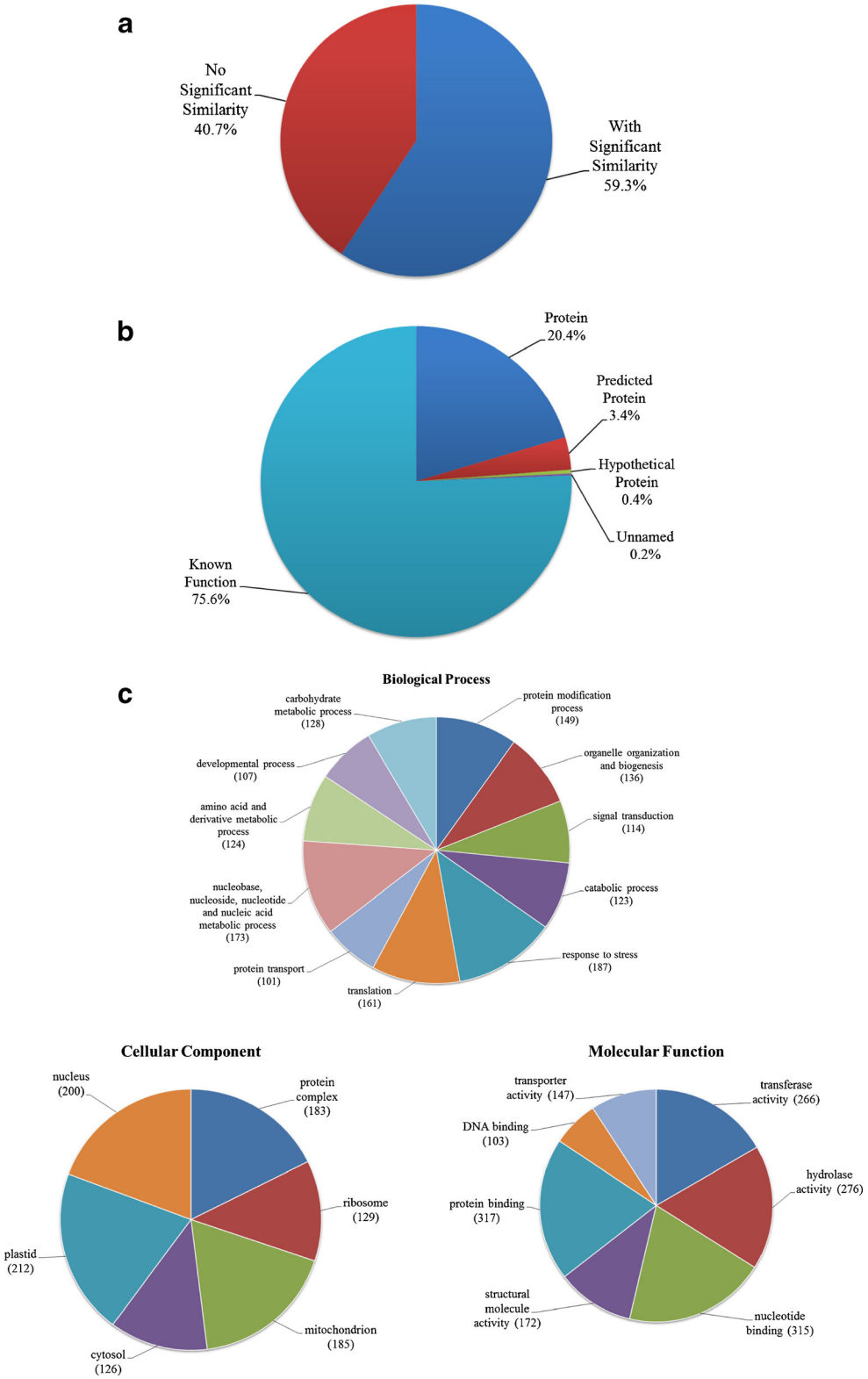

Fig. 2 EST analysis of 6,622 high quality kelampayan ESTs against the GenBank non-redundant protein database. Panel A shows the distribution of ESTs with significant and nonsignificant similarities to sequences in GenBank. Panel B shows the distribution of 3,930 ESTs with putative functions from the kelampayan developing xylem library. Panel C shows the GO-annotation results for Biological Process, Molecular Function and Cellular Component categories

clusters were then aligned using PHRAP and subsequently CRAW. Contigs and singletons generated from clustering were considered as sets of putative unique genes (unigenes).

A total of 6,622 high quality ESTs of at least 100 bp in length were generated from the kelampayan cDNA library with an average edited length of 478 bp. These ESTs were submitted to dbEST at the NCBI with a library accession number LIBEST_028358. The size distribution of the ESTs was shown in Fig. 1. The overall sequencing success rate was approximately 64 %. This was due to the presence of mono-nucleotide repeat motifs (polyA or polyT templates) in the sequences that causes sequence slippage problems. More than half (59.3 % or 3,930) of the 6,622 ESTs were assigned with putative functions using blastx analysis (E-value$\leq 10^{-10}$) against the GenBank non-redundant protein database. However, about 24.0 % of all matches with blastx were either protein (20.4 %), predicted protein (3.0 %), hypothetical protein (0.4 %) or unnamed (0.2 %). A large percentage (40.7 %) of the EST sequences was classified as no-hit or not significantly similar to sequences in GenBank (Fig. 2). This might probably have been due to the shorter average length (276 bp) of the latter sequences when compared to the average length of the EST sequences that showing similarities to the GenBank entries (616 bp). EST sequences not matching known proteins could imply that novel genes may be present in the developing xylem tissues of kelampayan.

On assembling 6,622 ESTs from 5' end sequences, 4,728 xylogenesis unigenes with an average length of 672 bp were generated. The analysis formed 2,100 consensus contig sequences, representing 3,994 or 60.3 % of all high quality ESTs, with lengths ranging from 132 bp to 2,706 bp, and an average of 621 bp. The remaining 2,628 (representing 39.7 % of the total high quality ESTs) were singletons which ranged from 104 to 839 bp, with an average length of 723 bp. Assembly analysis revealed a redundancy level of 28.5 % in the kelampayan EST database. However, the redundancy may increase during further EST sequencing (Li et al. 2009). By comparison, the EST redundancy in the kelampayan EST database is comparable to the estimated redundancy of 28 % in *Populus* (Aspeborg et al. 2005) and 28.8 % in *Pinus radiata* D. Don (Li et al. 2009).

The most abundant protein in the ESTs whose putative function was inferred from sequence comparison was 60s ribosomal protein with 92 ESTs, followed by 40s ribosomal protein with 42 ESTs. Interestingly, most genes involved in lignin biosynthesis were present in the kelampayan EST database with 1 to 21 ESTs. These included *phenylalanine ammonia-lyase* (*PAL*), *cinnamate 4-hydroxylase* (*C4H*), *coumarate 3-hydroxylase (ρ-coumaryl shikimate/quinate 3-hydroxylase)* (*C3H*), *caffeic acid O-methyltransferase* (*COMT*), *caffeoyl-CoA-3-O-methyltransferase* (*CCoAOMT*), *4-coumarate:CoA reductase* (*4CL*), *ferulate 5-hydroxylase* (*F5H*), *cinnamyl alcohol dehydrogenase* (*CAD*), *hydroxycinnamoyl-CoA: shikimate/quinate hydroxycinnamoyl transferase* (*HCT*) and *cinnamoyl-CoA reductase* (*CCR*). *COMT*, *CCoAOMT* and *C3H* are among the 30 most highly abundant genes with 18 to 21 ESTs (Table 1). Also, several ESTs exhibiting homologies to cell wall biosynthesis genes were also identified in the kelampayan EST database. The most highly abundant cell wall genes are tubulin (42 ESTs), arabinogalactan protein genes (30 ESTs) and cellulose synthase (13 ESTs). Other cell-wall related genes, including sucrose synthase, expansin, UDP-glucose dehydrogenase, xyloglucan endotransglycosylase and pectate lyase are moderately abundant with 2 to 11 ESTs in the kelampayan EST database.

Overall, this study has generated an important genomic resource for wood formation in kelampayan. The identified genes in this study will provide a useful resource for

Table 1 Thirty highly abundant genes or gene families with known functions in the 6,622 xylogenesis ESTs of kelampayan

Gene/Gene families	No. of ESTs	%
60s ribosomal protein	92	1.39
40s ribosomal protein	42	0.63
Tubulin	42	0.63
ADP-ribosylation factor	33	0.50
Histone	33	0.50
MIPs	32	0.48
Arabinogalactan protein	30	0.45
s-adenosylmethionine synthetase	29	0.44
Elongation factor	24	0.36
Actin	22	0.33
Glycine-rich RNA-binding protein	22	0.33
Caffeic acid 3-O-methyltransferase	21	0.32
Heat-shock protein	20	0.30
Ribosomal protein	20	0.30
Actin depolymerizing factor	19	0.29
Caffeoyl- CoA 3-O-methyltransferase	19	0.29
Regulatory protein rop	19	0.29
Cinnamate 4-hydroxylase	18	0.27
PVR3-like protein	18	0.27
Blue copper protein	17	0.26
Zinc finger family protein	17	0.26
GTP binding protein	15	0.23
Plasma membrane intrinsic protein	15	0.23
Transcription factor	15	0.23
Beta tubulin	14	0.21
Glyceraldehyde-3-phosphate dehydrogenase	14	0.21
5-methyltetrahydropteroyltriglutamate-homocysteine	13	0.19
Cellulose synthase	13	0.19
Translationally controlled tumor protein	13	0.19
B-tubulin	12	0.18

identifying molecular mechanisms controlling wood formation and will also be candidates for association genetic studies in kelampayan aiming at the production of high value forests (Thumma et al. 2005). Furthermore, comparison of kelampayan ESTs with sequences from angiosperms will also generate valuable information about the evolution of higher plants.

Acknowledgments The authors would like to thank the research officers from the Malaysia Genome Institute (MGI) for providing the EST sequencing services. Special thanks to Mr Mohd Noor Mat Isa for bioinformatics assistance. This work is part of the joint Industry-University Partnership Programme, a research programme funded by the Sarawak Forestry Corporation (SFC) and UNIMAS.

References

Aspeborg H, Schrader J, Coutine PM, Stam M, Kallas A, Djerbi S, Nilsson P, Benman S, Amini B, Sterky F et al (2005) Carbohydrate-active enzymes involved in the secondary cell wall biogenesis in hybrid aspen. Plant Physiol 137:983–997

Burke J, Davison D, Hide W (1999) d2_cluster: a validated method for clustering EST and full-length cDNA sequences. Genome Res 9: 1135–1142

Chou A, Burke J (1999) CRAWview: for viewing splicing variation, gene families and polymorphism in clusters of ESTs and full-length sequences. Bioinformatics 15:376–381

Ewing B, Green P (1998) Base-calling of automated sequences traces using Phred. II. Error probabilities. Genome Res 8:186–194

Ewing B, Hillier LAD, Wendl MC, Green P (1998) Base-calling of automated sequences traces using Phred. Genome Res 8:175–185

Joker D (2000) SEED LEAFLET *Neolamarckia cadamba* (Roxb.) Bosser (*Anthocephalus chinensis* (Lam.) A. Rich. ex Walp.) (http://curis.ku.dk/portal-life/files /20648324/ neolamarckia_ cadamba_int.pdf)

Li X, Wu HX, Dillon SK, Southerton SG (2009) Generation and analysis of expressed sequence tags from six developing xylem libraries in *Pinus radiata* D. Don. BMC Genomics 10:e41

Miller RT, Christoffels AG, Gopalakrishnan C, Burke J, Ptitsyn AA, Broveak TR, Hide WA (1999) A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. Genome Res 9:1143–1155

Pavy N, Paule C, Parsons L, Crow JA, Morency MJ, Cooke J, Johnson JE, Noumen E et al (2005) Generation, annotation, analysis and database integration of 16,500 white spruce EST clusters. BMC Genomics 6:144

Sterky F, Bhalerao RR, Unneberg P, Segerman B, Milsson P, Hertzberg M, Sandberg G (2004) A populus EST resource for plant functional genomics. Proc Natl Acad Sci USA 101(38):13951–13956

Thumma BR, Nolan MF, Evans R, Moran GF (2005) Polymorphisms in *Cinnamoyl CoA Reductase* (*CCR*) are associated with variation in microfibril angle in *Eucalyptus* spp. Genetics 171:1257–1265

Whetten R, Sun YH, Zhang Y, Sederoff R (2001) Functional genomics and cell wall biosynthesis in loblolly pine. Plant Mol Biol 47:275–291

World Agroforestry Centre (2004) A tree species reference and selection guide: *Anthocephalus cadamba* (http://www.worldagroforestrycentre.org/sea/Products/ AFDbases/af/asp/SpeciesInfo.asp?SpID=17933)