Electronic Journal of Biotechnology ISSN: 0717-3458
Vol.11 No.2, Issue of April 15, 2008

Received August 13, 2007 / Accepted December 6, 2007
DOI: 10.2225/vol11-issue2-fulltext-14
RESEARCH ARTICLE

Generation and analysis of an *Eucalyptus globulus* cDNA library constructed from seedlings subjected to low temperature conditions

Susana Rasmussen-Poblete
Departamento de Ciencias Biológicas
Universidad Andrés Bello
República 252, Santiago, Chile
Tel: 56 2 2383178
Fax: 56 2 2372259
E-mail: susana.rasmussen@bionova.cl

Jorge Valdés
Departamento de Ciencias Biológicas
Universidad Andrés Bello
República 252, Santiago, Chile
Tel: 56 2 2383178
Fax: 56 2 2372259
E-mail: jorge.valdes@gmail.com

Maria Cecilia Gamboa
Departamento de Ciencias Biológicas
Universidad Andrés Bello
República 252, Santiago, Chile
Tel: 56 2 2383178
Fax: 56 2 2372259
E-mail:cecy_gamboa@yahoo.es

Pablo D.T. Valenzuela
Departamento de Ciencias Biológicas
Universidad Andrés Bello
República 252, Santiago, Chile
Tel: 56 2 2398969
Fax: 56 2 2372259
E-mail:pvalenzu@bionova.cl

Erwin Krauskopf*
Departamento de Ciencias Biológicas
Universidad Andrés Bello
República 252, Santiago, Chile
Tel: 56 2 2383178
Fax: 56 2 2383178
E-mail: ekrauskopf@unab.cl

Financial support: This work was partially funded by Universidad Andrés Bello (DI Proyect: 04-05/1) and MIFAB (Proyect: P04-071-F) and by the Microsoft Joint Research Program.

Keywords: cellulose, cold-temperature, EST database, forest biotechnology, lignin.

Note: The sequences have been deposited in GenBank. Accession numbers: ES588357-ES597093

Abbreviations:
4CL: CoA ligase
AUX/IAA: auxin/indole-3-acetic acid
bZIP: basic leucine zip
C3H: p-coumarate 3-hydroxylase
C4H: 4-hydroxylase
CAD: cinnamyl alcohol dehydrogenase
CCR: cinnamoyl CoA reductase
CCoAOMT: caffeoyl-CoA 3-O-methyltransferase
ESTs: expressed sequence tags
F5H: ferulate 5-hydroxylase
GO: gene ontology
HCT: hydroxycinnamoyltransferase
PAL: phenylalanine ammonia lyase

Eucalyptus globulus **is the most important commercial temperate hardwood in the world because of its wood properties and due to its characteristics for biofuel production. However, only a very low number of expressed sequence tags (ESTs) are publicly available for this tree species. We constructed a cDNA from *E. globulus* seedlings subjected to low temperature and sequenced 9,913 randomly selected clones, generating 8,737 curated ESTs. The assembly produced 1,062 contigs and 3,879 singletons forming a *Eucalyptus* unigene set. Based on BLASTX analysis, 89.3% of the contigs and 88.5% of the singletons had significant similarity to known genes in the non-redundant database of GenBank. The *Eucalyptus* unigene set generated is a valuable public resource that provides an initial model for genes and regulatory pathways involved in cell wall biosynthesis at low temperature.**

Forests cover nearly 30% of the earth surface, nearly 4 billion hectares, serving multiple functions including conservation of biological diversity, renewing the oxygen supply of the atmosphere, preventing soil erosion and supplying pulp and wood (FAO, 2005). Forest tree breeding aims to improve the quality of trees by the selection of individuals with desirable traits that will later be used to produce trees with improved genotype. Genetic improvement programs such as controlled cross-pollination breeding have been used since the 1950s. Nevertheless, phenotype assessment is a complex process due to the long generation times of woody species (Grattapaglia, 2004). It is within the context of reducing this time-frame that functional genomics has become a powerful tool in forestry.

In the last few years functional genomics has been used extensively for gene discovery in species whose genomes have not been completely sequenced. A cost-effective and rapid way to obtain new data from an organism is through partial sequencing of randomly selected cDNA clones (Braütigam et al. 2005). The resulting collection of expressed sequence tags (ESTs) reveals a portion of genes in an organism expressed under a particular condition. Using this approach, several traits have been analyzed in trees, such as wood formation (Allona et al. 1998; Sterky et al. 1998; Israelsson et al. 2003) or cold tolerance (Nanjo et al. 2004; Sterky et al. 2004). Unfortunately, these studies have focused on gene expression profiles having a direct effect on the particular trait studied, without expanding the range of effects that the set condition might have on other metabolic pathways. In fact, cold stress in poplar cuttings (*Populus tremula* x *Populus tremuloides* cv. Mush1) has been shown to produce variations in parameters such as sucrose concentration and lignin content, illustrating the direct effect of cold conditions on wood quality (Hausman et al. 2000).

The amount and type of lignin and cellulose are important in the timber and pulp industry as they have a direct effect on the chemical properties of the wood produced by the tree (Jung and Ni, 1998; Fukushima, 2001; Plomion et al. 2001). For the production of biofuels, cellulose needs to be separated from lignin so it can be made available for enzyme hydrolysis. Therefore, several research groups have studied different ways by which to modify lignin and cellulose content on the plant cell wall. As a result, various studies have shown a co-regulation of these two compounds (Hu et al. 1999; Li et al. 2003; Rastogi and Dwivedi, 2006). For instance, the down-regulation of a single lignin biosynthetic gene resulted in a decrease of lignin production by the plant, while exhibiting an increase in cellulose production (Hu et al. 1999). Hence, the modification of plant cell wall composition in trees may provide a way to engineer wood for biofuel production.

E. globulus is considered the most important temperate hardwood plantation species in the world due to its combination of wood properties suitable for the pulp and paper industry (Jones et al. 2002; Grattapaglia, 2004). This tree species has fast growth rates and an ability to adapt to a broad range of geographic locations (ranging from latitude 35ºS to 42ºS), even though its growth rate diminishes due to frost conditions (Jones et al. 2002; Miranda and Pereira, 2002). Most importantly, *Eucalyptus* has been listed as one of the candidate biomass energy crops by the U.S. Department of Energy (U.S. Department of Energy, 2007). Nevertheless, public genomic information from *E. globulus* is limited. In fact, an analysis of publicly available *E. globulus* ESTs at the GenBank EST repository (by July 06, 2007) registered only 3,953 ESTs for *E. globulus* compared to the mostly represented tree, *Pinus taeda* (329,469 ESTs). Thus, in this study we provide and describe the first publicly available cDNA library from cold-treated *E. globulus* seedlings, paying particular attention to genes predicted to be involved in cell wall biosynthesis and the transcription factors suggested to be involved in their regulation).

MATERIALS AND METHODS

Plant material

E. globulus seeds were germinated in a soil mixture and grown in a culture cabinet with a 16 hrs day/8 hs night photoperiod at a temperature of 21ºC. The library was constructed from 3-month old *Eucalyptus globulus* plants maintained at 4ºC degrees for 30 min. After cold treatment, *E. globulus* leaves were collected and frozen in liquid nitrogen until use.

RNA extraction and cDNA library construction

Total RNA was extracted according to the method described by Chang and collegues (Chang et al. 1993). RNA integrity was confirmed by gel electrophoresis and 1

*Corresponding author

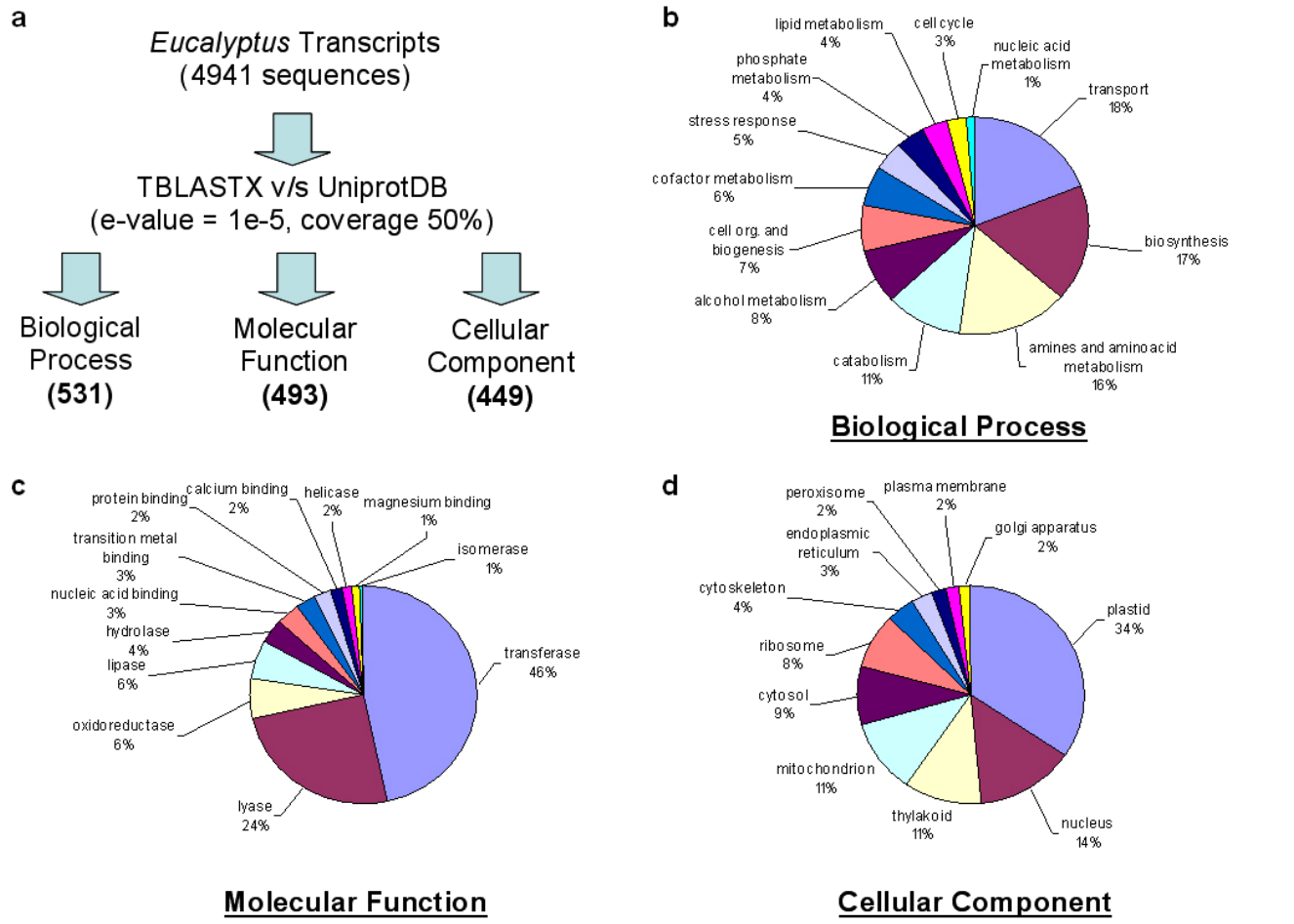

Figure 1. Functional categorization of *E. globulus* unigenes. (a) Schematic representation of the functional categorization process. (b-d) Distribution of *E. globulus* unigenes across GO categories. Parent categories and their percentages are shown in bold, sub-categories and the number of deduced proteins is shown in parenthesis.

mg was quantified using a RNA standard (Invitrogen, Cat 15620-016). Poly (A) mRNA was isolated from total RNA with the Stratagene Poli (A) Quick mRNA Isolation Kit (Stratagene, La Jolla, CA, USA). cDNA was prepared and cloned using the vector pExpress I exploiting the Not I and Eco RV restriction sites. The cDNA library was not normalized, *i.e.* no attempt was made to reduce the redundancy of highly expressed transcripts.

EST sequencing, filtering and assembly

In total, 9,913 bacterial colonies were randomly picked and single-pass sequence reactions performed. These sequences were analyzed using Phred base calling software (with Q>20) (Ewing et al. 1998). All traces were subjected to a trimming process for the removal of ribosomal RNA, poly (A) tails, low-quality sequences, vector and adapter regions. Sequences with 94% of identity over 40 or more nucleotides were assembled using the CAP3 software (Huang and Madan, 1999).

Unigene function assignment and categorization

The unigene set was classified and analyzed according to gene ontology (GO) terms (Ashburner et al. 2000) across functional categories. The complete unigene set was compared against the protein non-redundant database using BLASTX (Altschul et al. 1997) and analyzed with the InterProScan program (Zdobnov and Apweiler, 2001) to assign a putative function. GO terms were extracted from the best hits obtained from the BLASTX comparison against SwissProt-Trembl database (Fleischmann et al. 1999) (E-value < E-15 and >70% of alignment coverage) and compared to the InterProScan GO suggestions. All the GO assignments were curated manually (Ashburner et al. 2000). The unigene dataset was compared to other *Eucalyptus* cDNA libraries available in Genbank through BlastN program using an e-value cutoff of E-5.

RESULTS AND DISCUSSION

Analysis of *E. globulus* cDNA library

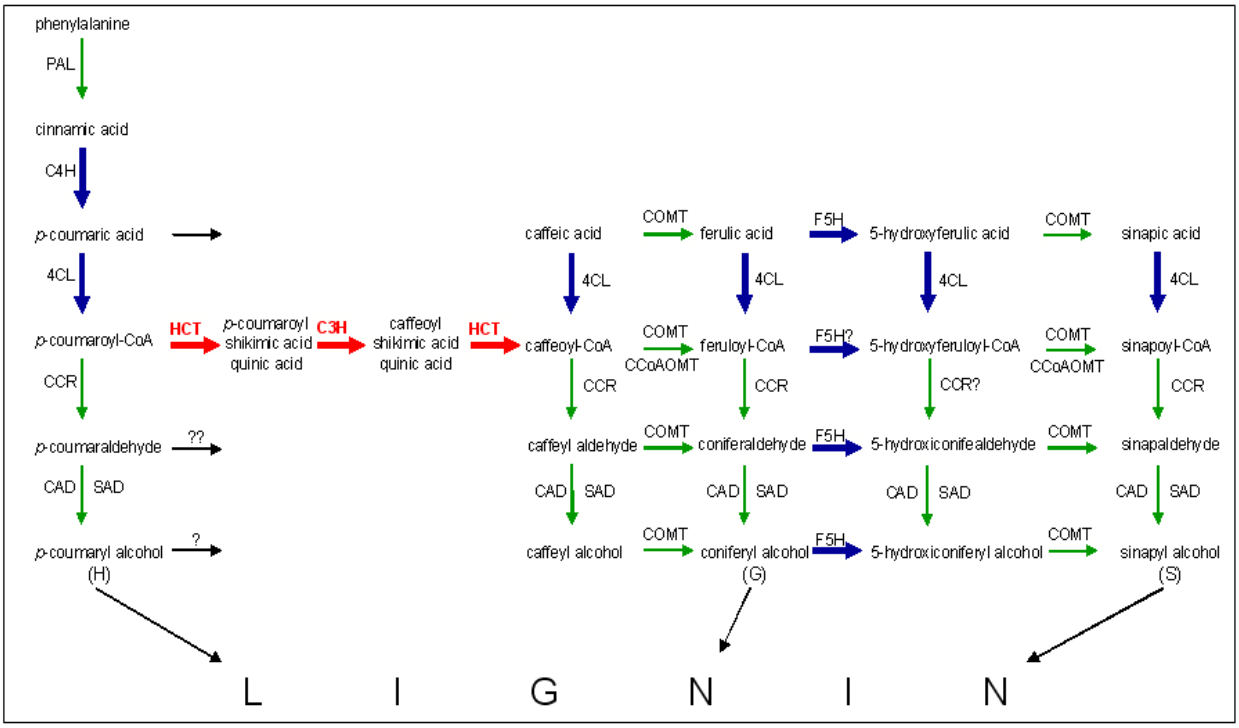

Figure 2. Reduction of total sulfide in effluent from a Figure 2. Simplified scheme of the enzymes proposed to be involved in lignin bio synthesis. Green a rrows represent predicted *E. globulus* enzy me mediated reactions w hose sequenc es are available in GenBank. Blue arrows rep resent sequences fr om other *Eucalyptus* species available in GenB ank. Red a rrows r epresent reactions catalyzed by enzymes that had not been previously described for any *Eucalyptus* species. Bold arrows correspond to sequences found in our *E. globulus* library th at were n ot previo usly d escribed f or this species. Accesion numbers for *E. globulus* enz ymes. P AL: phenylalanine a mmonia ly ase [GenBank: AF16 7487]; CCR: cinnamo yl CoA red uctase [GenBan k: DQ084797 -DQ084795, GenB ank: AH0154889]; CAD: cinnam yl alcohol dehy drogenase [GenBan k: AF109157, AF038561 an d O64969], COM T: caffeic acid *O*-methyltransferase [GenBank: Q9 SWC2] and CCo AOMT: caffeo yl-CoA 3-*O*-methyltransferase [Gen Bank: AAD50443 and AAC2619 1). Accesion numbers for Eucal yptus enz ymes, C4H : trans-cinnamat e 4-h ydroxylase [GenBank: A F149713-AF149715], F5H: ferulate 5-hydroxylase [Ge nBank: AJ24909 3; AF1497 13-AF149715] a nd 4C L: 4-coumarate: CoA ligase [GenBank: NC_0 03070, D Q147001, AJ244010, AF149715, AF149714 and AF149713]. reduction reactor by sulfur oxidizing bacteria.

The analysis of the 9,913 sequence-reads resulted in the generation of 1,062 contigs and 3,879 singletons (4,941 unigenes) (Figure 1). The fraction of sequences represented by more than one cDNA was 60.9%, providing an estimate of library redundancy. Based on BLASTX analysis, 89.3% of the contigs and 88.5% of the singletons had a significant similarity to known genes in the non-redundant database (Altschul et al. 1997). As for contigs, their composition ranged from 2 to 118 ESTs. The deepest contigs were considered highly represented unigenes. Those contigs with more than 50 ESTs are shown as Table 1 (contigs with 20 or more ESTs are included as Supplementary data 1).

Overall, 541 unigenes were assigned to biological processes, 449 to cellular component and 493 to molecular function categories. This is a low number of assignments compared to other libraries generated in different studies of trees (*Pinus*: 5474, 5064 and 5886 respectively; Poplar: 6158, 5751 and 6622 respectively; Spruce: 1697, 1467 and 2188 respectively) (Quackenbush et al. 2000). We suggest that this is due to low average similarity between our database and the uniprot sequences database, in addition to the low alignment coverage obtained (we used both parameters to make the assignments). We focused our analyses on the physiological processes (431) being the most represented process related to cellular metabolism, with 48 unigenes related to alcohol metabolism, 95 unigenes associated to amines and aminoacid derivative metabolism, 116 unigenes involved in transport and 164 related to biosynthetic processes.

The most represented molecular functions corresponded to binding and catalytic activities. The unigenes allocated to binding activity were associated with ion binding (130) and nucleic acids binding (62). Furthermore, 114 unigenes were associated with enzymes involved in redox reactions related to lignin biosynthesis and 88 with tranferase activities, including enzymes involved in lignin and cellulose biosynthesis.

Genes predicted to be involved in wood formation

The EST database was screened for sequences with significant similarity to genes involved in the biosynthesis

Table 1. Contigs with ESTs highly represented. Assigned function is indicated in contigs with more than 50 ESTs.

Number ESTs	Contigs length (nt)	Assigned function	Relative organism	Similarity (%)	Accession number (GI)
118	1604	Plastidic aldolase	*Nicotiana paniculata*	94.8	4827253
86	1509	Chloroplast latex aldolase-like protein	*Manihot esculenta*	90.9	56122688
81	979	Ribulose-1,5-bisphosphate carboxylase/oxygenase small subunit	*Panax ginseng*	93.2	77157637
80	1693	Ribulose-1,5-bisphosphate coarboxylase/oxygenase activase precursor	*Malus x domestica*	93.1	415852
73	1590	Glyceraldehydes-3-phosphate dehydrogenase A subunit	*Glycine max*	87.8	77540210
51	1721	AAA ATPase, central region; homeodomain-like	*Medicago truncatula*	89.1	92870561

of lignin monomers and cellulose. All of the genes known to participate in the lignin biosynthetic pathway are represented in our cDNA library. Two of the predicted gene products, *p-coumarate 3-hydroxylase* (C3H) and *CoA:shikimate/ quinate hydroxycinnamoyltransferase* (HCT) had not been described previously in any *Eucalyptus* species. However, genes encoding *trans-cinnamate 4-hydroxylase* (C4H), *ferulate 5-hydroxylase* (F5H) and *4-coumarate: CoA ligase* (4CL) had been described in other *Eucalyptus* species but not in *E. globulus* (Harakava, 2004). The remainder of the genes found had been previously described for *E. globulus* and published in GenBank, including *phenylalanine ammonia lyase* (PAL), *cinnamoyl CoA reductase* (CCR), *cinnamyl alcohol dehydrogenase* (CAD), *caffeic acid O-methyltransferase* (COMT) and *caffeoyl-CoA 3-O-methyltransferase* (CCoAOMT) (Figure 2) (Supplementary data 2).

The assembly of the C3H and HCT ESTs showed that two isoforms of their gene-products are represented in our cDNA library. C3H and HCT participate in the process of converting p-coumaryl CoA into caffeoyl-CoA, resulting in the production of coniferyl (G) and sinapyl (S) lignin units. Down-regulation of C3H in transgenic alfalfa plants and *Arabidopsis* mutants resulted in a significant difference in lignin composition due to an alteration in the number and nature of the monolignol monomers (Franke et al. 2002; Ralph et al. 2006). The characterization of the *Arabidopsis reduced epidermal fluorescence* (r*ef8*) mutant defective in C3H suggested that the genetic modification of this gene may not be appropriate for the reduction of lignin content in forest species because the mutant plants generated exhibited vascular collapse, developmental abnormalities and increased susceptibility to pathogen attack (Boerjan et al. 2003; Cooke et al. 2004).

Three unigenes exhibited similarity to known cellulose synthase genes. Analysis of their predicted domains by InterProScan revealed that all of them contained the cellulose synthase domain that is composed of three aspartic residues and a QXXRW motif, playing a significant role in the catalytic activity of this enzyme (Krauskopf et al. 2005). However, the zinc finger domains (IPR001841 and IPR011011) present in cellulose synthase proteins were not found in our sequences since the sequences were not full-length. The deduced *E. globulus* proteins were compared with the ones previously described for *E. grandis* (Ranik and Myburg, 2006) as no sequences were available for *E. globulus* (Supplementary data 3).

Transcription factors involved in wood formation

Of the 56 transcription factor families described in *Arabidopsis* and 63 in rice (Guo et al. 2005; Gao et al. 2006), 11 of them were represented in our library: auxin/indole-3-acetic acid (AUX/IAA) family, B3 family, basic/helix-loop-helix (bHLH) family, basic leucine zip (bZIP) family, GRAS family, homeodomain-leucine zipper (HD-Zip) (HB) family, heat shock family (HSF), MYB family, WRKY family, zinc finger homeobox (ZF-HD) family and ZIM family. Transcription factors families such as AUX/IAA, MYB and HD containing domains (zinc finger proteins and homeodomain-leucine zipper) regulate the expression of genes that participate in xylem development and secondary wall formation (lignin and cellulose biosynthesis) (Oh et al. 2003; Cánovas et al. 2004).

Many of the genes encoding the enzymes of general phenylpropanoid metabolism, such as PAL, C4H, 4CL, COMT and CAD contain conserved motifs within their promoters that are recognized by plant MYB transcription factors (Tamagnone et al. 1998). Twelve members of the MYB family were found in our library. Some of them had a best BLASTX hit with *GOLDEN2-like 1* gene, *LHY-CCA1-like* 5 gene and *DIVARICATA* gene. The coverage of the sequences with their best BLASTX hit ranged from 25% and 100%. Two *E. gunnii* MYB transcription factors sequences were found in GenBank [GenBank: AJ576023-AJ576024] (Goicoechea et al. 2005). Based on BLASTN

analysis, these sequences were different from the ones obtained in our library. Others families less represented in our library belonged to the ZF family and bZIP (with seven members each), WRKY family (five members with coverage of their best BLASTX hit between 12% and 50%) and one member of the AUX/IAA family, (Supplementary data 4).

In addition, the data gathered through these analyses was compared with the few existing *Eucalyptus* cDNA libraries currently found in GenBank. The comparison was made against *Eucalyptus gunnii* (8,538 ESTs), *Eucalyptus globulus subsp. bicostata* (2,685 ESTs), *Eucalyptus grandis* (1,574 ESTs) and *Eucalyptus globulus 'blue gum'* (1,266 ESTs). BlastN comparisons against our *E. globulus* database revealed a low level of similarity between our sequenced library and the available datasets. The number of sequences that have at least one match with E-values better than 1E-5 for each library were 1,335 ESTs for *E. gunnii* (15%), 464ESTs for *E. globulus subsp. bicostata* (17%), 267 for *E. grandis* (17%) and 261 ESTs for *E. globulus 'blue gum'* (17%).

In conclusion, a unigene set of approximately 4900 unigenes was obtained from our *E. globulus* cDNA library. Analysis of its content has provided valuable data for the future metabolic engineering of plant cell walls by identifying new potential targets that will allow future modification for biofuel production and industrial use. In addition, our results will be useful for comparative genomic studies among hardwoods and softwoods.

ACKNOWLEDGMENTS

We thank Dr. Danilo González who provided the computer cluster to generate the unigene set and Dr. David Holmes for critical reading of the manuscript.

REFERENCES

ALLONA, I.; QUINN, M.; SHOOP, E.; SWOPE, K.; ST. CYR, S.; CARLIS, J.; RIEDL, J.; RETZEL, E.; CAMPBELL, M.; SEDEROFF, R. and WHETTEN, R. Analysis of xylem formation in pine by cDNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, August 1998, vol. 95, no. 16, p. 9693-9698.

ALTSCHUL, S.; MADDEN, T.; SCHAFFER, A.; ZHANG, J.; ZHANG, Z.; MILLER, W. and LIPMAN, D. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, September 1997, vol. 25, no. 17, p. 3389-3402.

ASHBURNER, M.; BALL, C.; BLAKE, J.; BOTSTEIN, D.; BUTLER, H.; CHERRY, J.; DAVIS, A.; DOLINSKI, K.; DWIGHT, S.; EPPIG, J.; HARRIS, M.; HILL, D.; ISSEL-TARVER, L.; KASARSKIS, A.; LEWIS, S.; MATESE, J.; RICHARDSON, J.; RINGWALD, M.; RUBIN, G. and SHERLOCK, G. Gene Ontology: tool for the unification of biology. *Nature Genetics*, May 2000, vol. 25, p. 25-29.

BOERJAN, W.; RALPH, J. and BAUCHER, M. Lignin biosynthesis. *Annual Review of Plant Biology*, June 2003, vol. 54, p. 519-546.

BRAÜTIGAM, M.; LINDLÖF, A.; ZAKHRABETKOVA, S.; GHARTI-CHHETRI, G.; OLSSON, B. and OLSSON, O. Generation and analysis of 9792 EST sequences from cold acclimated oat, *Avena sativa*. *BMC Plant Biology*, September 2005, vol. 5, p. 18.

CÁNOVAS, F.; DUMAS-GAUDOT, E.; RECORBET, E.; JORRIN, J.; MOCK, H.-P. and ROSSIGNOL, M. Plant proteome analysis. *Proteomics* 2004, vol. 4, p.285-298.

CHANG, S.; PURYEAR, J. and CAIRNEY, J. A simple and efficient method for isolating RNA from pines trees. *Plant Molecular Biology Reporter*, June 1993, vol. 11, p. 113-116.

COOKE, J.; MORSE, A. and DAVIS, J. Forestry. In: KLEE and P. CRISTOU eds. *Handbook of Plant Biotechnology*. United Kingdom. John Wiley and Sons, 2004, vol. 2, p. 881-904.

EWING, B.; HILLIER, L.; WENDL, M. and GREEN, P. Basecalling of automated sequencer traces using phred I. Accuracy Assessment. *Genome Research*, March 1998, vol. 8, no. 3, p. 175-185.

FAO. Progress towards sustainable forest managment. In: FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS eds. *Global forest resources assessment*. Rome. FAO forestry paper, 2005, p. 147.

FLEISCHMANN, W.; MOELLER, S.; GATEAU, A. and APWEILER, R. *A novel method for automatic functional annotation of proteins. Bioinformatics*, March 1999, vol. 15, no. 3, p. 228-233.

FRANKE, R.; HUMPHREYS, J.M.; HEMM, M.R.; DENAULT, J.W.; RUEGGER, M.O.; CUSUMANO, J.C. and CHAPPLE, C. Changes in secondary metabolism and deposition of an unusual lignin in the *ref8* mutant of *Arabidopsis*. *The Plant Journal*, April 2002, vol. 30, no. 1, p. 33-45.

FUKUSHIMA, K. Regulation of syringyl to guaiacyl ratio in lignin biosynthesis. *Journal of Plant Research*, February 2001, vol. 114, no. 4, p. 499-508.

GAO, G.; ZHONG, Y.; GUO, A.; ZHU, Q.; TANG, W.; ZHENG, W.; GU, X.; WEI, L. and LUO, J. DRTF: a database of rice transcription factors. *Bioinformatics*, March 2006, vol. 22, no. 10, p. 1286-1287.

GOICOECHEA, M.; LACOMBE, E.; LEGAY, S.; MIHALJEVIC, S.; RECH, P.; JAUNEAU, A.; LAPIERRE, C.; POLLET, B.; VERHAEGEN, D.; CHAUBET-GIGOT, N. and GRIMA-PETTENATI, J. *Eg*MYB2, a new transcriptional activator from *Eucalyptus* xylem, regulates secondary cell wall formation and lignin biosynthesis. *The Plant Journal*, August 2005, vol. 43, no. 4, p. 553-567.

GRATTAPAGLIA, D. Integrating genomics into *Eucalyptus* breeding. *Genetics and Molecular Research*, September 2004, vol. 3, no. 3, p. 369-379.

GUO, A.; HE, K.; LIU, D.; BAI, S.; GU, X.; WEI, L. and LUO, J. DATF: a database of *Arabidopsis* transcription factors. *Bioinformatics*, February 2005, vol. 21, no. 10, p. 2568-2569.

HARAKAVA, R. Genes encoding enzymes of the lignin biosynthesis pathways in *Eucalyptus*. *Genetics and Molecular Biology*, 2004, vol. 28, no. 3, p. 601-607.

HAUSMAN, J.F.; EVERS, D.; THIELLEMENT, H. and JOUVE, L. Compared responses of poplar cuttings and in vitro raised shoots to short-term chilling treatments. *Plant Cell Reports*, October 2000, vol. 19, no. 10, p. 954-960.

HU, W.J.; HARDING, S.A.; LUNG, J.; POPKO, J.L.; RALPH, J.; STOKKE, D.D.; TSAI, C.J. and CHIANG, V.L. Repression of lignin biosynthesis promotes cellulose accumulation and growth in transgenic trees. *Nature Biotechnology*, August 1999, vol. 17, p. 808-812.

HUANG, X. and MADAN, A. CAP3 A DNA sequence assembly program. *Genome Research*, September 1999, vol. 9, no. 9, p. 868-877.

ISRAELSSON, M.; ERIKSSON, M.; HERTZBERG, M.; ASPEBORG, H.; NILSSON, H. and MORITZ, T. Changes in gene expression in the wood-forming tissue of transgenic hybrid aspen with increased secondary growth. *Plant Molecular Biology*, July 2003, vol. 52, no. 4, p. 893-903.

JONES, R.C.; STEANE, D.A.; POTTS, B.M. and VAILLANCOURT, R.E. Microsatellite and morphological analysis of *Eucalyptus globulus* populations. *Canadian Journal of Forest Research*, January 2002, vol. 32, no. 1, p. 59-66.

JUNG, H.-J. and NI, W. Lignification of plant cell walls. Impact of genetic manipulation. *Proceedings of the National Academy of Sciences of the United States of America*, October 1998, vol. 95, no. 22, p. 12742-12743.

KRAUSKOPF, E.; HARRIS, P. and PUTTERILL, J. The cellulose synthase gene PrCESA10 is involved in cellulose biosynthesis in developing tracheids of the gymnosperm *Pinus radiata*. *Gene*, May 2005, vol. 350, no. 2, p. 107-116.

LI, L.; ZHOU, Y.H.; CHENG, X.F.; SUN, J.Y.; MARITA, J.M.; RALPH, J. and CHIANG V.L. Combinatorial modification of multiple lignin traits in trees through multigene cotransformation. *Proceedings of the National Academy of Sciences of the United States of America*, April 2003, vol. 100, no. 8, p. 4939-4944.

MIRANDA, I. and PEREIRA, H. Variation of pulpwood quality with provenances and site in *Eucalyptus globulus*. *Annals of Forest Science*, 2002, vol. 59, p. 283-291.

NANJO, T.; FUTAMURA, N.; NISHIGUCHI, M.; IGASAKI, T.; SHINOZAKI, K. and SHINOHARA, K. Characterization of full-length enriched expressed sequence tags of stress-treated poplar leaves. *Plant and Cell Physiology*, 2004, vol. 45, no. 12, p. 1738-1748.

OH, S.; PARK, S. and HAN, K.-H. Transcriptional regulation of secondary growth in *Arabidopsis thaliana*. *Journal of Experimental Botany*, December 2003, vol. 54, no. 393, p. 2709-2722.

PLOMION, C.; LEPROVOST, G. and STOKES, A. Wood formation in trees. *Plant Physiology*, December 2001, vol. 127, p. 1513-1523.

QUACKENBUSH, J.; LIANG, F.; HOLT, I.; PERTEA, G. and UPTON, J. The TIGR Gene Indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Research*, 2000, vol. 28, no. 1, p. 141-145.

RALPH, J.; AKIYAMA, T.; KIM, H.; LU, F.; SCHATZ, P.F.; MARITA, J.M.; RALPH, S.A.; SRINIVASA-REDDY, M.S.; CHEN, F. and DIXON, R.A. Effects of coumarate 3-hydroxylase down-regulation lignin structure. *Journal of Biological Chemistry*, March 2006, vol. 281, no. 13, p. 8843-8853.

RANIK, M. and MYBURG, A. Six new cellulose synthase genes from *Eucalyptus* are associated with primary and secondary cell wall biosynthesis. *Tree Physiology*, May 2006, vol. 26, no. 5, p. 545-556.

RASTOGI, S. and DWIVEDI, UN. Down-regulation of lignin biosynthesis in transgenic Leucaena leucocephala harboring O-methyltransferase gene. *Biotechnology Progress*, 2006, vol. 22, p. 609-616.

STERKY, F.; REGAN, S.; KARLSSON, J.; HERTZBERG,M.; ROHDE, A.; HOLMBERG, A.; AMINI, B.; BHALERAO, R.; LARSSON, M.; VILLAROEL, R.; VAN MONTAGU,M.; SANDBERG, G.; OLSSON, O.; TEERI, T.; BOERJAN, W.; GUSTAFSSON, P., UHLÉN, M.; SUNDBERG, B. and LUNDEBERG, J. Gene discovery in the wood forming tissues of poplar: analysis of 5,692 expressed sequence tags. *Proceedings of the National Academy of Sciences of the United States of America*, October 1998, vol.95, no. 22, p. 13330-13335.

STERKY, F.; BHALERAO, R.; UNNEBERG, P.; SEGERMAN, B.; NILSSON, P.; BRUNNER, A.;

CHARBONNEL-CAMPAA, L.; LINDVALL, J.; TANDRE, K.;STRAUSS, S.; SUNDBERG, B.; GUSTAFSSON, P.; UHLÉN, M.; BHALERAO, R.; NILSSON, O.; SANDBERG, G.; KARLSSON, J.; LUNDEBERG, J. and JANSSON, S. A populus EST resource for plant functional genomics. *Proceedings of the National Academy of Sciences of the United States of America*, September 2004, vol. 101, no. 38, p. 13951-13956.

TAMAGNONE, L.; MERIDA, A.; PARR, A.; MACKAY, S.; CULIANEZ-MACIA, F.; ROBERTS, K. and MARTIN, C. The AmMYB308 and AmMYB330 transcription factors from Antirrhinum regulate phenylpropanoid and lignin biosynthesis in transgenic tobacco. *Plant Cell*, February 1998, vol. 10, no. 2, p. 135-154.

U.S. DEPARTMENT OF ENERGY. OIT Agriculture Plants/crop-based renewable resources 2020. [cited date July 2007], 2007. Available from Internet: http://www.energy.gov.

ZDOBNOV, E.M. and APWEILER, R. InterProScan-an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, 2001, vol. 17, no. 9, p. 847-848.

APPENDIX
SUPLEMENTARY DATA

Supplementary data 1. Contigs with ESTs highly represented. Assigned function is indicated in contigs with more than 20 ESTs.

Nº ESTs	Contigs Length (nt)	Assigned Function	Relative Organism	Similarity (%)	Accession Number (gi)
118	1604	plastidic aldolase	*Nicotiana paniculata*	94.8	4827253
86	1509	chloroplast latex aldolase-like protein	*Manihot esculenta*	90.9	56122688
81	979	ribulose-1,5-bisphosphate carboxylase/oxygenase small subunit	*Panax ginseng*	93.2	77157637
80	1693	ribulose-1,5-bisphosphate carboxylase/oxygenase activase precursor	*Malus x domestica*	93.1	415852
73	1590	glyceraldehyde-3-phosphate dehydrogenase A subunit	*Glycine max*	87.8	77540210
51	1721	AAA ATPase, central region; Homeodomain-like	*Medicago truncatula*	89.1	92875540
49	1320	glyceraldehyde-3-phosphate dehydrogenase B subunit	*Glycine max*	79.5	77540212
47	1051	Ferritin-related	*Medicago truncatula*	86.7	92870561
45	951	glycolate oxidase	*Mesembryanthemum crystallinum*	97.4	1773330
43	1085	unnamed protein product type I (CAB-21) (LHCP)	*Nicotiana tabacum*	96.7	19823
41	940	serine-glyoxylate aminotransferase	*Spirodela polyrrhiza*	84.6	74229863
39	1068	thi	*Citrus sinensis*	85.1	2582665
36	1468	putative galactinol synthase	*Pisum sativum*	83.2	541885
35	916	myo-inositol 1-phosphate synthase (MI-1-P) synthase (IPS)	*Sesamum indicum*	94.0	9858816
33	1346	catalase	*Soldanella alpina*	97.9	2462661
32	1368	oxygen evolving enhancer protein 1 precursor	*Bruguiera gymnorrhiza*	88.8	9229957
32	1261	glycine hydroxymethyltransferase methylase	*Solanum tuberosum*	95.6	438247
31	1286	galactinol synthase, isoform GolS-1	*Ajuga reptans*	83.6	5608497
31	961	BURP domain-containing protein	*Bruguiera gymnorrhiza*	59.0	14422444
28	1118	ferredoxin--NADP(+) reductase	*Nicotiana tabacum*	90.3	2225993
27	1761	S-adenosylmethionine decarboxylase	*Ipomoea nil*	82.9	1498080
27	2319	PAPS-reductase-like-protein	*Catharanthus roseus*	85.7	12831474
26	1256	Phosphoribulokinase, chloroplast precursor	*Mesembryanthemum crystallinum*	92.7	125578
25	1192	AlaT1	*Vitis vinifera*	95.4	71842524
24	975	polyubiquitin	*Pinus sylvestris*	96.1	1332579
23	843	S-adenosylmethionine decarboxylase	*Prunus persica*	58.0	47232488
23	1119	putative RNA binding protein	*Arabidopsis thaliana*	84.7	3850621
23	1320	translation elongation factor 1A-9	*Gossypium hirsutum*	95.4	74486744
23	1214	AT4g04640	*Arabidopsis thaliana*	85.4	7267222
22	1074	OJ000223_09.15	*Oryza sativa* (japonica cultivar-group)	91.1	38346061
21	1038	aminomethyltransferase system T protein	*Arabidopsis thaliana*	95.3	15221119
21	1066	NADPH-protochlorophyllide oxidoreductase	*Cucumis sativus*	90.5	2244614
21	1288	carbonic anhydrase	*Vigna radiata*	79.5	8954289
21	956	NAD-dependent epimerase/dehydratase family protein-like protein	*Solanum tuberosum*	93.7	82400136
20	1979	P-Protein-like protein precursor	*Arabidopsis thaliana*	84.5	7270248
20	934	S-adenosylmethionine decarboxylase	*Prunus persica*	86.7	47232488
20	1155	sedoheptulose-1,7-bisphosphatase precursor	*Oryza* sativa (indica cultivar-group)] cultivar-group)	91.8	27804772
20	1238	Malate dehydrogenase, glyoxysomal precursor	*Citrullus lanatus*	83.3	126894
20	1857	Transketolase, C-terminal-like	*Medicago truncatula*	88.4	92892666

[a]Temperature incubation at 30ºC, [b]Initial pH at 7.
μ: mean value and standard deviation of three determinations.

Supplementary data 2. Analysis of *E. globulus* unigenes corresponding to enzymes involved in wood formation.

Gene Name	Best Blastx Hit	Score	E-value	Coverage	Domains
EgPAL1	PAL *Daucus carota*	387	8e-113	29.09%	IPR001106 IPR008948 PF00221
EgPAL2	PAL *Camellia sinensis*	306	8e-82	29.13%	IPR001106 IPR008948 PF00221
EgPAL3	PAL *Citrus limon*	278	2e-73	25.76%	IPR001106 IPR008948 PF00221
EgC4H	C4H *Populus kitakamiensis*	482	3e-142	53.07%	IPR001128 IPR002401 PF00067
Eg4CL1	4CL *Populus balsamifera subsp. trichocarpa x Populus deltoides*	246	6e-64	33.16%	IPR000873 PF00501
Eg4CL2	4CL *Eucalyptus camaldulensis*	260	4e-68	36.95%	IPR000873 PF00501
EgHCT1	HCT *Nicotiana tabacum*	331	4e-89	49.66%	IPR003480 IPR02458
EgHCT2	HCT *Oryza sativa (japonica cultivar-group)*	186	1e-45	45.13%	IPR003480 IPR02458
EgC3H1	C3H-1 *Ocimum basilicum*	249	3e-72	29.49%	IPR001128 IPR002401 PF00067
EgC3H2	C3H *Ocimum basilicum*	375	1e-102	44.41%	IPR001128 IPR002401 PF00067
EgCOMT	COMT *Eucalyptus gunnii*	451	1e-125	63.66%	IPR001077 PF00891 IPR011991
EgCCoAOMT1	CCoAOMT *Plantago major*	96.7	6e-19	47.27%	IPR002935 PF01596
EgCCoAOMT2	CCoAOMT *Ammi majus*	121	2e-26	42.32%	IPR002935 PF01596
EgCCR1	CCR *Eucalyptus gunnii*	443	1e-123	63.00%	IPR001509 PF01370
EgCCR2	CCR *Arabidopsis thaliana*	256	5e-67	61.68%	IPR001509 PF01370
EgF5H1	F5H *Camptotheca acuminata*	539	1e-151	69.26%	IPR001128 IPR002401 PF00067
EgF5H2	F5H *Camptotheca acuminata*	126	8e-28	14.00%	IPR001128 IPR002401 PF00067
EgCAD1	CAD *Eucalyptus globulus*	530	1e-149	88.76%	IPR002085 IPR002328 IPR011032
EgCAD2	CAD *Arabidopsis thaliana*	220	4e-56	57.91%	IPR001509

Supplementary data 3. Analysis of *E. globulus* cellulose synthase unigenes.

Gene Name	Best Hit with *E. grandisi*	Best Blastx Hit	Score	E-value	Coverage	Domains
EgCesA1	EgrCesA5	CesA *Populus tremula x Populus tremuloides*	400	1e-110	24.11%	IPR005150 PF03552
EgCesA2	No EgrCesA related	CesA *Medicago truncatula*	238	2e-62	28.34%	IPR005150 PF03552
EgCesA3	EgrCesA4	CesA4 *Eucalyptus grandis*	309	9e-83	17.4%	IPR005150 PF03552

Supplementary data 4. Analysis of *E. globulus* unigenes corresponding to transcription factors.

Best Blast Hit	Related Organism	Score	E-value	Coverage	Domains
putative MYB transcription factor	*Oryza sativa* (japonica cultivar-group)] cultivar-group)]	130	1e-28	30.8%	IPR001005 IPR006447 IPR009057 IPR012287
MYB transcription factor LHY-CCA1-like5	*Arabidopsis thaliana*	241	5e-62	90.1%	IPR001005 IPR006447 IPR009057 IPR012287
GPRI1 (GOLDEN2-LIKE 1); transcription factor	*Arabidopsis thaliana*	239	2e-61	77.9%	IPR000183 IPR001005 IPR006447 IPR009057
MYB-like transcription factor DIVARICATA	*Antirrhinum majus*	328	2e-88	73.9%	IPR001005 IPR006447 IPR009057 IPR012287
MYB-like transcription factor 6	*Gossypium raimondii*	288	2e-76	103.1%	IPR001005 IPR009057 IPR012287
MYB11	*Malus x domestica*	230	4e-59	50.3%	IPR001005 IPR009057 IPR012287
MYBR2	*Malus x domestica*	75.9	2e-12	41.6%	No related InterPro
MYB-like DNA-binding protein	*Catharanthus roseus*	95.5	2e-18	24.4%	No related InterPro
transcription factor MYB1	*Malus xiaojinensis*	88.6	2e-16	47.0%	No related InterPro
MYB, DNA-binding	*Medicago truncatula*	78.6	2e-13	28.3%	No related InterPro
MYB transcription factor LHY-CCA1-like5	*Arabidopsis thaliana*	147	5e-34	35.5%	IPR001005 IPR006447 IPR009057 IPR012287
MYB-related protein	*Arabidopsis thaliana*	218	2e-55	48.8%	IPR001005 IPR009057 IPR012287
IAA18; transcription factor	*Arabidopsis thaliana*	55.5	2e-06	33.3%	No related InterPro
Transcriptional factor B3	*Medicago truncatula*	102	2e-20	32.7%	IPR003340
Transcriptional factor B3	*Medicago truncatula*	188	2e-46	22.7%	No related InterPro
bZIP transcription factor protein	*Capsicum annuum*	75.1	3e-12	45.5%	IPR004827 IPR008917 IPR011616
Putative ripening-related bZIP protein	*Vitis vinifera*	171	3e-41	38.7%	No related InterPro
bZIP transcription factor ZIP-2	*Nicotiana tabacum*	80.5	7e-14	32.9%	IPR004827 IPR008917
Putative ripening-related bZIP protein	*Vitis vinifera*	123	7e-27	37.4%	No related InterPro
ATBZIP60	*Arabidopsis thaliana*	82	2e-14	61.7%	IPR004827 IPR008917 IPR011616
Putative ripening-related bZIP protein	*Vitis vinifera*	176	9e-43	45.2%	No related InterPro
Putative ripening-related bZIP protein	*Vitis vinifera*	84.3	6e-15	16.3%	No related InterPro
GRAS transcription factor	*Medicago truncatula*	67.8	5e-10	10.2%	No related InterPro
GRAS transcription factor	*Medicago truncatula*	252	9e-66	33.4%	IPR005202
ATHB-7	*Arabidopsis thaliana*	142	2e-32	33.3%	IPR000047 IPR001356 IPR003106 IPR009057 IPR012287
Heat shock transcription factor	Phaseolus acutifolius	272	1e-71	60.7%	IPR000232 IPR002341 IPR011991
WRKY9; transcription factor	*Arabidopsis thaliana*	127	1e-27	75.4%	IPR003657
Putative WRKY-type DNA binding protein	*Glycine max*	380	1e-104	50.5%	IPR003657 IPR000583
Putative WRKY4 transcription factor	*Vitis aestivalis*	74.3	4e-12	34.4%	No related InterPro
DNA-binding WRKY	*Medicago truncatula*	168	2e-40	38.3%	IPR003657

Putative WRKY4 transcription factor	*Vitis aestivalis*	240	5e-62	66.6%	IPR003657
Putative zinc finger transcription factor	*Oryza sativa* (japonica cultivar-group)cultivar-group	242	1e-62	35.1%	IPR000571
Putative zinc finger transcription factor	*Oryza sativa* (japonica cultivar-group)cultivar-group	243	4e-63	35.1%	IPR000571
Zinc finger protein	*Malus x domestica*	134	3e-30	26.1%	IPR007087
Zinc finger protein, putative	*Arabidopsis thaliana*	137	1e-30	52.4%	IPR000315
putative zinc finger protein	*Oryza sativa* (japonica cultivar-group)cultivar-group	133	9e-30	73.5%	IPR007087
Putative zinc finger transcription factor	*Oryza sativa* (japonica cultivar-group)cultivar-group	225	2e-57	31.3%	IPR000571
Zinc finger protein-like	*Arabidopsis thaliana*	123	6e-27	5.4%	IPR003349
Zinc finger protein, putative	*Plasmodium falciparum* 3D7	60.1	8e-08	28.8%	IPR002653
Putative zinc finger transcription factor	*Oryza sativa* (japonica cultivar-group)cultivar-group	239	9e-62	34.8%	IPR000571
Zinc finger protein, putative	*Arabidopsis thaliana*	109	2e-22	21.6%	IPR000315 IPR002906
Zinc finger protein, putative	*Arabidopsis thaliana*	152	1e-35	36.3%	IPR000315
Putative zinc finger transcription factor	*Oryza sativa* (japonica cultivar-group)cultivar-group	218	2e-55	28.4%	IPR000571