Plant Molecular Biology (2005) 57:203–224
DOI 10.1007/s11103-004-6969-7

Large-scale statistical analysis of secondary xylem ESTs in pine

Nathalie Pavy[1,*], Jérôme Laroche[3], Jean Bousquet[1,2] and John Mackay[1]
[1]*ARBOREA and Centre de Recherche en Biologie Forestiére,* [2]*Chaire de Recherche du Canada en Génomique Forestiére et Environnementale &* [3]*Centre de bio-informatique, Université Laval, Pavillon Charles-Eugène-Marchand, Sainte-Foy, Que., Canada G1K 7P4 (*author for correspondence; e-mail nathalie.pavy@rsvs.ulaval.ca)*

Received 20 August 2004; accepted in revised form 30 November 2004

Key words: computational analysis, EST, gene discovery, pine, xylem

Abstract

A computational analysis of pine transcripts was conducted to contribute to the functional annotation of conifer sequences. A statistical analysis of expressed sequential tags(ESTs) belonging the 7732 contigs in the TIGR *Pinus* Gene Index (PGI1.0) identified 260 differentially represented gene sequences across six cDNA libraries from loblolly pine secondary xylem. Cluster analysis of this subset of contigs resulted in five groups representing genes preferentially represented in one of the xylem samples (compression wood, plannings, root xylem, latewood) and one group containing mostly genes simultaneously present in compression and side wood libraries**.** To complement the sequence annotation, 27 cDNA clones representing selected transcripts were completely sequenced. Several genes were identified that could represent putative markers for xylem from different organs, at different stages of development. Several sequences encoding regulatory proteins were over-represented in root xylem as opposed to the other xylem samples. Some of them belonged to known families of plant transcription factors, but two genes were previously uncharacterized in plants. One transcript was homologous to the gene encoding the Smad4 interacting factor, a key co-activator in TGFβ (transforming growth factor) signalling in animals. Thus, the digital analysis of pine ESTs highlighted a putative gene function of potentially broad interest but that has yet to be investigated in plants. More generally, this study showed that the application of numerical approaches to EST databases should be helpful in establishing priorities among genes to consider for targeted functional studies. Thus, we illustrated the potential of extracting information from conifer sequences already accessible through well-structured public databases.

Introduction

To investigate gene function at the genome level in conifers, expressed sequence tags (ESTs) sequencing projects were undertaken in pine (Allona *et al.*, 1998; Kirst *et al.*, 2003) and are in progress in white spruce. With information available from the genome sequence for a number of angiosperms (*Arabidopsis*, poplar, rice) and from growing EST databases for several gymnosperms (*Cycas*, pine, spruce), it becomes possible to identify unique genes or families of genes specific to gymnosperms or with different roles than that seen in angiosperms (Brenner *et al.*, 2003). In an effort to discover genes related to wood formation and the underlying wood quality traits, large random EST libraries have been recently sequenced and analysed both in pine (Allona *et al.*, 1998; Zhang *et al.*, 2000; Whetten *et al.*, 2001; Dubos and Plomion, 2003) and in poplar (Sterky *et al.*, 1998; Hertzberg *et al.*, 2001; Mellerowicz *et al.*, 2001).

To analyse large EST collections, statistical tools are available that can reveal correlated expression among genes (Ewing *et al.*, 1999).

Although the computational approach is different from the clustering of microarray data (Eisen *et al.*, 1998), the aim is similar in attempting to cluster genes according to their expression profiles. Assuming that the number of sequences detected in a cDNA library is proportional to the transcript abundance, the variations in the sequence frequency are used to infer the differential expression of the corresponding genes. Several statistical methods using pairwise or multiple comparisons of libraries have been proposed (reviewed by Claverie, 1999) and are now broadly applied for this purpose. The *in silico* reconstruction of transcriptional profiles has been facilitated by the development of specialized sequence databases and statistical procedures (Bortoluzzi and Danieli, 1999; Bortoluzzi *et al.*, 2001). Applied to the human genome, this approach has already enabled the high throughput computational analysis of transcription patterns and it has been cross-validated with SAGE data (e.g. Bortoluzzi *et al.*, 2000). Furthermore, it helped to identify sequences of medical interest and potentially linked to cardiovascular pathologies (Mégy *et al.*, 2002) or to cancer (Stekel *et al.*, 2000; Romualdi *et al.*, 2001). In plants, studies in rice (Ewing *et al.*, 1999), wheat (Ogihara *et al.*, 2003) and potato (Ronning *et al.*, 2003) showed that this approach is helpful to assign a functional annotation to anonymous genes. Based on existing statistical procedures and computational annotation of transcript sequences, we prospected a dataset derived from six cDNA libraries prepared from pine xylem tissues in order to improve our understanding of the biological mechanisms involved in wood development.

To study pine gene functions, we explored more broadly the assigned annotations of the complete *Pinus* Gene Index (PGI1.0) dataset (Quackenbush *et al.*, 2000). This exhaustive dataset was built from all the pine cDNAs available in dbEST. Beside the xylem EST data, it included other pine sequence data that facilitated both the EST clustering and sequence annotation. The provided annotations and cross-references to the gene ontology were informative. The present study enabled the detection of differential distribution across the xylem libraries from pairwise and multiple comparisons of contig sets. It highlighted the existence of transcripts that could be differentially represented in xylem tissues and not previously characterized in any plant species. For some of these sequences, we undertook the sequencing of the complete cDNA clones to improve their annotation but no known function could be ascribed to some of these sequences. The overall statistical approach provided criteria to identify genes for further experimental study of xylem differentiation and indicated that large-scale transcript sequence analysis of gymnosperms might provide complementary information to angiosperm model plants.

Materials and methods

Sources of sequence data

Pine ESTs and contig sequences were retrieved from the TIGR *Pinus* Gene Index release 1.0 (PGI1.0). Informations relative to all the libraries used to build the PGI as well as the procedure for generating the contigs and their annotation are described on the TIGR web site (http://www.tigr.org/tbd/tgi/pgi/) (Quackenbush *et al.*, 2000). The definitions employed are conform to the PGI nomenclature (available at: http://www.tigr.org/tbd/tgi/definitions.html). TC sequences refer to Tentative Consensus derived from coding sequences and represent contigs. We extracted the contigs derived from *Pinus taeda* ESTs only. Indeed, in the PGI database, some contig sequences are made of ESTs from different pine species, which can lead to the misclustering of sequences and introduce a bias in the results. To extract the *Pinus taeda* contigs, we searched for the intersection between the 21,409 unique sequences from PGI and the set of 72,965 *Pinus taeda* sequences available from NCBI (June 2003 release). Among these 21,409 sequences, 1542 had a link to the gene ontology.

The poplar sequences used for comparisons were obtained from the DOE Joint Genome Institute (http://www.jgi.doe.gov/). The *Arabidopsis* and rice clusters were retrieved from the TIGR Gene Indices (http://www.tigr.org/tdb/tgi/plant.shtml); the *Cycas* clusters were retrieved from the Sputnik database at the Munich Information Center for Protein Sequences (http://mips.gsf.de/proj/sputnik/). The spruce clusters were derived from the CCGB group at University of Minnesota, based on a collection of in-house sequences (http://ccgb.umn.edu/biodata/spruce/).

Sequence comparisons against the pine root EST database available at University of Georgia (L. H. Pratt and M.-M. Cordonnier-Pratt laboratory) was conducted by using the blast search engine from http://fungen.org/blast/blast.html.

Statistical tests for contig differential distribution among cDNA libraries

For the numerical analysis of contigs, all matrices were derived by in-house programs written in PERL. A distribution matrix was derived by dividing the occurrence of each contig in each library by the total number of sequences in the library to take into account the different numbers of sequences from each library. The following homogeneity tests of occurrence of the contigs among cDNA libraries were conducted: the Audic and Claverie (AC) test (Audic and Claverie, 1997), Fisher's 2×2 exact test (F), χ^2 2×2 contingency tables, the R test (Stekel *et al.*, 2000), and the Greller and Tobin test (GT) (Greller and Tobin, 1999). For this purpose, the IDEG6 software was used to generate the *P*-values (Romualdi *et al.*, 2001). The significance level was set at $\alpha = 0.05$. The Bonferroni correction for multiple tests was applied, so that the selected significance threshold became 3.2·e−45 for the AC test, Fisher exact test, and the χ^2 test, and 4.8·e−45 for the GT and the R tests.

Cluster analysis of distribution patterns

The EST composition of the differentially distributed contigs among cDNA libraries was used to estimate the expression profile matrix, which consisted of N rows (the contigs) and M columns (the cDNA libraries), with $n = 800$ (the number of contigs most abundantly represented in the cDNA libraries, see Results) and $m = 6$. Thus, each value represented the relative frequency of the *i*th contig in the *j*th library. Pearson's correlation coefficient was calculated between each pair of contigs, thus generating a $N \times N$ matrix. The correlation matrix was transformed to an Euclidean distance matrix, which was submitted to unweighted pair group method with arithmetic averages (UPGMA) to estimate a phenogram using PHYLIP (Felsenstein, 1993). Phenograms were displayed with TreeView (Eisen *et al.*, 1998).

Cloning, sequencing and sequence processing

For contigs chosen for complete sequencing, the longest cDNA clone available in the libraries was sequenced to obtain the full sequence from a single insert. First, the insert length was assessed by PCR amplification using T7 and T3 primers. Then, if it was longer than the contig sequence, the clone was completely sequenced, starting with T7 and T3 primers and completed with sequence-specific internal primers to obtain complete sequence information of both DNA strands. The sequences were analysed and assembled with the *seqmerge* program (Wisconsin Package Version 10.3, Accelrys Inc., San Diego, CA). Similarity searches were performed locally by using the *blast* program (Altschul *et al.* 1997) against the protein and nucleic databases retrieved from NCBI and against the PRODOM database. Prediction of secondary structures elements and their alignments were conducted by using the PHD (Rost *et al.,* 1993) and SOPM methods (Geourjon and Deleage, 1994) (http://www.npsa-pbil.ibcp.fr/). Sequence alignments were performed with T-Coffee software (Notredame *et al.*, 2000).

RT-QPCR analysis of gene expression

Steady state RNA levels were determined in secondary xylem tissues isolated from the stem and from the root of *Picea glauca* L. trees, by quantitative RT-PCR (RT-QPCR). The tissues were collected from each of two trees growing close to each other in a 15-year-old plantation. Each tree was cut down, three 40–50 cm long bolts were cut from the basal portion of the stem and from the largest diameter roots; the bark was peeled away from the bolts, the differentiating xylem tissue following Zhang *et al.* (2000). One extraction of total RNA was carried for each sample from each tree, following the procedure of Chang *et al.* (1993); its concentration and quality were determined using a BioAnalyzer (Agilent technologies). Each RNA sample was treated with DNAse I amplification grade (Invitrogen) and reverse transcribed with Superscript II (Invitrogen) and then analysed by QPCR. Sequence information for the spruce R2R3-*myb* gene PgMYB8 is described in Bedon *et al.* (2004). It is the putative spruce ortholog to the TC2087 sequence and was amplified with the following primer pair (designed with

Primer3; Rozen. and Skaletsky, 2000): 5′-GGTG GACTCAGTTGTAATAA-3′ and 5′-GTATCT CACCTATTTACAGATCA-3′. The QPCR reaction was carried out with the DyNamo SYBR® green kit (MJ Research). The data were calibrated with a specific standard curve established with a dilution series of a linear DNA fragment encompassing the spruce sequence. The data were expressed as relative values, determined by dividing the target gene data by data obtained for a reference gene EF1-alpha, amplified with the following primers 5′-AACTGGAGAAGGAACC CAAG-3′, 5′-AACGACCCAATGGAGGATAC-3′.

Results

Detection of differentially distributed contigs in xylem cDNA libraries

We analysed the distribution of ESTs in six large non-normalized cDNA libraries based upon the sequence assembly of the *Pinus* Gene Index. PGI1.0 was a collection of unique transcripts derived from 61,886 ESTs from several pine species. It contained 21,409 unique pine sequences including 13,677 singletons and 7732 contigs. Most of the ESTs used to build the PGI1.0 dataset were derived from six cDNA libraries prepared from different *Pinus taeda* xylem tissues (Kirst *et al.*, 2003) contributed by the "Pine NSF" project (North Carolina State University). In this study, we used the total number of sequences available from each library: 10,443 in side wood (NXSI) – 7162 in compression wood (NXCI) – 8158 in plannings, comprised of partly lignified xylem (NXPV) – 8321 in latewood (NXLV) – 8713 in root xylem (NXRV) – 6339 in normal wood (NXNV).

Results of the distribution heterogeneity tests

The raw dataset was built and analysed largely following the procedure of Ewing *et al.* (1999) that has been broadly applied to cDNA datasets (eg. Megy *et al.*, 2002; Ogihara *et al.*, 2003). Based upon recommendations from these studies, only contigs containing five ESTs or more were considered. The application of this criterion resulted in the selection of 800 contigs most abundantly represented in the six xylem libraries. We used the composition of contigs to deduce a matrix of frequencies for the 800 contigs across the six xylem libraries. With this matrix, we tested whether each contig was more or less uniformly distributed among the cDNA libraries. Under the hypothesis that clones are picked at random within the libraries, the number of times a transcript is sequenced from one library should reflect the expression level of the gene in the corresponding tissue sample (Ewing *et al.*, 1999). The null hypothesis that a given gene has the same frequency in each library was tested by using pairwise and multiconditional homogeneity tests as described by Claverie (1999) and Stekel *et al.* (2000), respectively. We applied five statistical tests: AC (Audic and Claverie, 1997), R (Stekel *et al.*, 2000), Fisher's 2×2 exact test, GT (Greller and Tobin, 1999) and the conventional χ^2. For each contig, we determined which test or which combination of tests detected significant differences (Figure 1). Three tests, namely AC, R and χ^2, were sufficient to detect the statistically significant heterogeneity of distribution among the cDNA libraries. Our analyses showed that 260 sequences were differentially distributed among the six xylem libraries according to at least one statistical test (Figure 1). The most sensitive test was AC with 246 significant tests ($P < 0.05$, Bonferroni corrected), followed by χ^2 with 240 significant tests and R with 155 significant tests. The F and GT tests gave 161 and 135 significant tests, respectively, only representing a subset of those obtained with AC and χ^2. The F and GT tests have previously been reported to be less appropriate for EST datasets (Romualdi *et al.*, 2001). The results obtained by the AC test were mostly consistent with results obtained with χ^2 since 240 contigs were declared differentially distributed by both AC and the χ^2 tests. Finally, results obtained with χ^2, F and GT tests were more similar to those of the AC than the R test: among the 105 contigs detected as differentially distributed by AC but not by the R test, 81 were also detected by χ^2, F and GT tests.

Clustering map of distribution patterns

The clustering map of the 260 differentially distributed contigs resulted in five clusters (Figure 2). Three groups included contigs primarily found in one library: root xylem (46 contigs, group C), latewood (48 contigs, group D), plannings (89 contigs, group E). Group A (9 contigs) and group B (68 contigs) included contigs with correlated

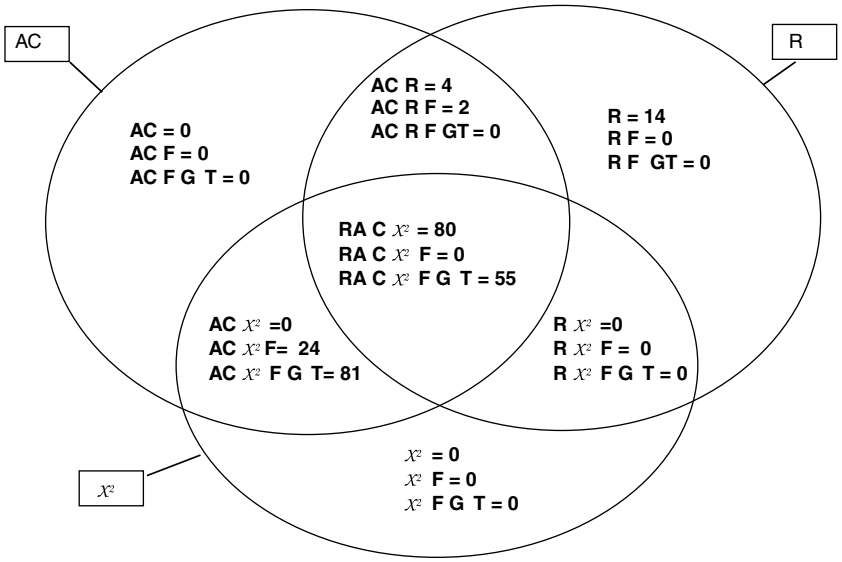

Figure 1. Number of contigs from *Pinus taeda* xylem libraries that are differentially distributed among libraries according to distribution heterogeneity tests. For each group, the names of the tests that were satisfied and the number of contigs are given. AC : Audic and Claverie test, R : Stekel test, F: Fisher test, GT : Greller and Tobin test.

distribution patterns across several libraries, predominantly compression and side wood. Group B was not as cohesive as the other ones because it included contigs with diverse distribution patterns and encoding proteins having a variety of functions. However, a subset of 37 contigs from group B (B*) were mostly found in compression wood. Almost none of the 260 contigs were found in normal xylem library.

Validation of digital expression data in independent samples

Experimental verification of digital expression data on the tissue samples or RNAs used to produce the cDNA libraries, by using different methods of analysis was not possible because the samples were no longer available for analysis. However, sequence characterizations and expression studies of R2R3-*myb* genes in *Picea glauca* L., also a softwood tree belonging to the Pinaceae family provided a validation for our findings with TC2087, with independent tissue samples (Bedon *et al.*, manuscript in preparation). Within ongoing EST sequence analyses in spruce, we studied the R2R3-*myb* family in detail and obtained complete coding sequence for 13 different genes from this family in spruce. Their steady state RNA levels were surveyed by RT-QPCR, in secondary xylem tissues of stems and roots, in two different 15-year-old trees. Six of the *P. glauca* R2R3-mybs were strongly expressed in xylem tissues, however only the *PgMyb8* sequence was more strongly expressed in root xylem compared to the shoot xylem. Its RNA levels in roots and stems were 0.1838 ± 0.0986 and 0.0567 ± 0.0314 (using relative expression units normalized against EF1-alpha), respectively. A phylogenetic analysis including all the R2R3-*myb* genes known in pine and spruce revealed that the closest spruce sequence to the pine cDNA TC2087 was *PgMyb8*. The spruce *PgMyb8* sequence shares 94.5% of sequence identity with the pine TC2087 sequence.

Annotation of the differentially distributed sequences

Annotations in PGI1.0 revealed that 52% of the 260 differentially distributed contigs had a homolog of known or unknown function in another species (Figure A, supplemental data) and 25% were associated to a gene ontology term (Table 1). Predicted functions of the corresponding proteins were consistent with the abundant literature describing the molecular mechanisms related to plant vascular development. For example, the differentially distributed contigs among xylem libraries included sequences coding for cell wall related proteins, proteolytic enzymes, transporters. Among 25 contigs differentially distributed and

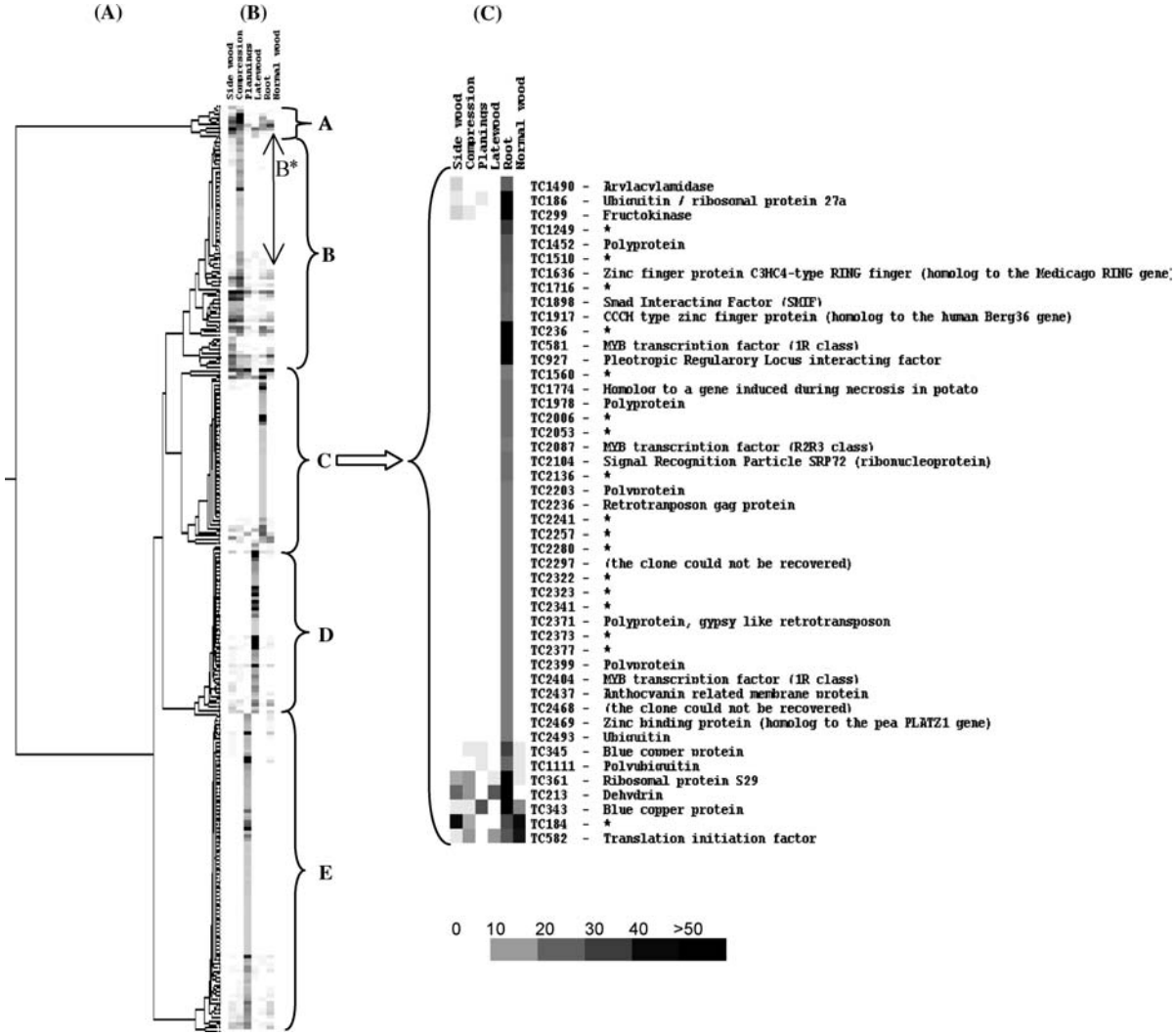

Figure 2. (A) Phenogram of the 260 differentially distributed contigs among the six xylem cDNA libraries. Clustering was based on UPGMA analysis of a pairwise Euclidean distance matrix reflecting the distribution frequency of contigs among the 6 xylem cDNA libraries. (B) Digital distribution patterns represented by shaded boxes (see scale). The five major clusters contained sequences found either in at least two libraries including compression wood and/or side wood (A and B), or predominantly in a single library, i.e. compression wood (group B*), root xylem (group C), latewood (group D), plannings (group E). (C) Curated annotations of the sequences over-represented in the root cDNA library. Significant similarities found at the protein level are reported. *stands for orphan sequences, those lacking similarity to any publicly available sequence (*blastx* reports against the non-redundant protein database with an *e*value $< 1 \cdot e{-}05$).

encoding cell wall related proteins, 12 contigs were over-represented in compression wood, and seven other contigs were more abundant in both compression and side wood, while none were preferential to the latewood library. The clustering of genes coding for proteins belonging to a common pathway or related molecular process, as observed in our dataset, has also been obtained through computational analyses of expression profiling data approaches (e.g, Ewing *et al.*, 1999).

Differential representation of related sequences including multigenic families in different cDNA libraries

The 260 differentially distributed contigs among cDNA xylem libraries were compared to each other, revealing several groups of contigs with high sequence similarity (Figure 3). For example, we searched for pairs of sequences with similarities greater than 70% or 90% of identity over a stretch of 100 nucleotides or more, and thus identified 18

Table 1. Number of contigs found as differentially distributed among the six xylem libraries and belonging to functional classes according to the gene ontology used for the PGI1.0 annotation.

Functional category	Number of contigs
Biological processes	
Aromatic amino acid family biosynthesis, shikimate pathway	1
Biological_process unknown	5
Carbohydrate metabolism	1
Cell differentiation	1
Circadian rhythm	1
Defense/immunity protein	1
Embryogenesis and morphogenesis	1
Glycine metabolism	1
Growth	2
Lignin biosynthesis	3
Mitochondrial transport	1
Mitosis	1
Peptide metabolism	1
Peroxidase reaction	2
Protein degradation tagging	6
Proteolysis and peptidolysis	4
Response to oxidative stress	2
Signal transduction	2
Translation elongationfactor	1
Translational initiation	1
Molecular functions	
Adenosinetriphosphatase	1
Aldehyde dehydrogenase (NAD(P)+)	1
Alpha-galactosidase	1
Aminopeptidase	1
Antifungal peptide	1
Calcium-dependent phospholipid binding	1
Calmodulin binding	1
Chorismate synthase	1
Copper binding	1
DNA binding	1
GTP binding	2
Heat shock protein	3
Hydrolase	2
Laccase	3
Lipoxygenase	1
Molecular function unknown	7
Oxidoreductase	5
Pectinesterase	1
Peroxidase	2
Phosphoenolpyruvate carboxykinase (ATP)	1
Pre-mRNA splicing factor	1
Protein kinase	1
Regulation of transcription, DNA-dependent	2
RNA binding	4
Subtilase	3
Cellular components	
Cell wall	1
Cellular component unknown	5
Chloroplast	1
Cytoplasm	2
Membrane	3
Nucleus	1
Structural constituent of cytoskeleton	1

and 12 groups of related contigs, respectively. Each group contained two to seven contigs that likely represent members of gene families. Several groups included contig sequences similar to genes coding for known proteins such as ubiquitins, aquaporins, tubulins, late embryogenic abundant proteins (LEA), blue copper proteins, GAPDH, cellulose synthases and heat shock proteins. Nine groups of related contigs with sequences sharing 70% of identity over at least 100 nucleotides consisted of orphan contigs, that is, these sequences lacked similarity to any publicly available sequence. Among these nine groups, we recovered and completely sequenced the longest cDNA clone available for six contigs, and did not recover cDNA clones for the three other contigs. With a longer cDNA sequence, we were able to assign a putative function to a few groups of contigs based upon sequence similarity. For some groups of contigs, the longer cDNA sequence had significant similarity with a sequence of known function in another species. Indeed, one group of four contigs showed high homology to a lipid transfer protein, one group of three contigs was similar to a putative dehydrin and one group of two contigs was highly similar to an allyl alcohol dehydrogenase. In spite of the additional information obtained from the other longer cDNA sequences, the three last groups of contigs remained orphan, suggesting that these contigs represent genes not yet reported in plants.

Functional assignment of conifer sequences based on sequence similarities

Among the 260 differentially distributed sequences, 48% showed no sequence similarity to genes

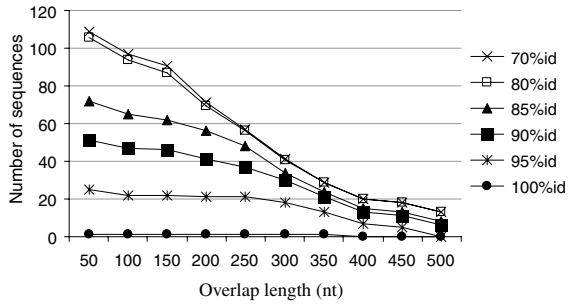

Figure 3. Sequence identity within the set of 260 differentially distributed contigs among cDNA xylem libraries. Each data series represents the number of sequences sharing the specified percentage of identity (%id) as a function of the length of the overlap between two sequences.

from other species in publicly available databases, according to the PGI1.0 annotation (Figure 4). These sequences are termed 'orphans' (Olivier, 1996). There were more orphans among shorter consensus sequences, as might be expected due to the potential lack of conserved regions in shorter sequences. Nonetheless, 52 orphans were found among the 148 contigs longer than 600 nucleotides, strongly suggesting that the lack of sequence similarity cannot entirely be ascribed to insufficient sequence information. We conducted two additional analyses to further assign putative functions to pine contigs. First, we conducted large-scale sequence similarity searches by confronting all the contigs of PGI1.0 to various large plant datasets. Second, we determined the complete sequence for a subset of 21 cDNA clones corresponding to contigs annotated as orphans and over-represented in root xylem.

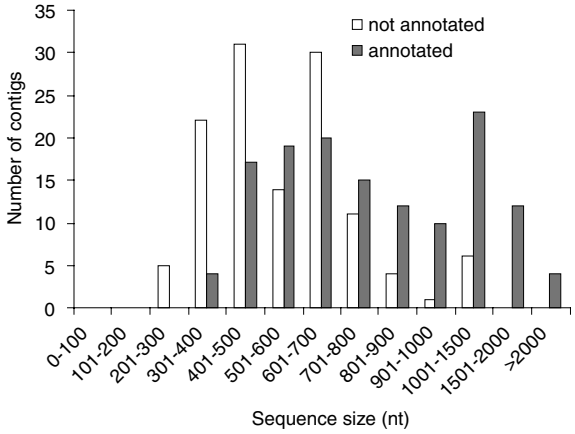

Figure 4. Size of the 260 differentially distributed contigs according to the ability to annotate them based on sequence similarity searches according to the PGI1.0 annotation.

Extended sequence similarity searches based on large datasets

For a more comprehensive overview of sequence conservation between conifers and angiosperms than that offered by the subset of 260 differentially represented sequences, we considered the complete set of 7732 pine contigs, which were of mean size of 591 nt. According to PGI1.0 annotations, 4210 (54%) contigs were orphans, 400 were similar to sequences of unknown function and 3122 were similar to sequences of known function. We conducted new homology searches with the *Arabidopsis* and rice gene indices and the recently available sequence datasets from poplar (DOE Joint Genome Institute – http://www.jgi.doe.gov/), *Cycas* (Brenner *et al.*, 2003), and spruce (Arborea project – http://www.arborea.ca), as they could bring meaningful insight to our analysis and reduce the reported proportion of orphan contigs in PGI1.0 (Figure 5A). The sequences were compared at the protein level by using *blastx* or *tblastx*, with an *e*Value $< 1 \cdot e{-}10$. Among the 7732 pine contigs, 61.5%, 59.4%, 55% and 27.7% matched a sequence from *Arabidopsis,* rice*,* poplar, and *Cycas*, respectively. As might be expected, a larger overlap was found between pine and spruce with 5510 (71%) of the pine contigs matching one of the 16,602 spruce contigs and singletons.

Only considering the data relative to the 4210 orphan contigs (Figure 5B), we found that 29% and 12% matched an *Arabidopsis* or *Cycas*, respectively (with *e*Value $< 1 \cdot e{-}10$). In contrast, 57% of the orphan pine contigs matched a spruce sequence, consistent with the view that they are either highly diverged or unique to conifers (Figure 5b). In light of these results, the proportion of orphans in the pine contig dataset decreased from 54% to 39%. Similarly, the proportion of orphans among the 260 differentially represented sequences went from 48% to 32%. As such, we could not eliminate all the orphans by using the analysis with the more extensive sequence datasets.

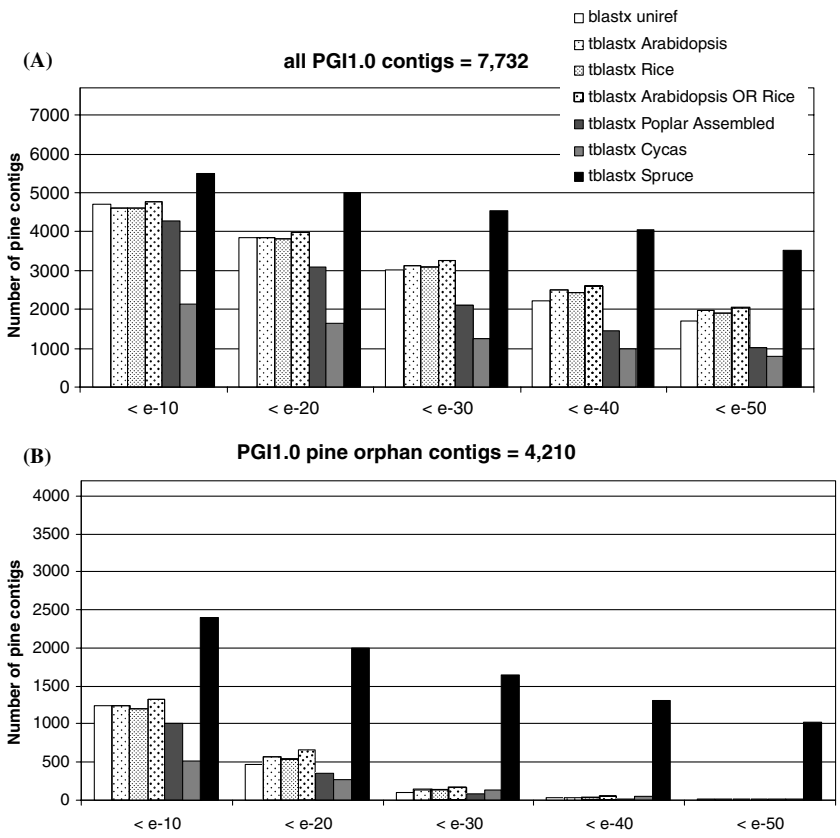

Figure 5. Frequency of positive matches from sequence comparisons for several sets of conifer sequences. (A) 7732 pine contigs from PGI1.0; (B) 4210 pine contigs of PGI1.0 previously annotated as orphans, those lacking similarity to any publicly available sequence. The sequences were compared to the contigs from spruce, *Cycas*, *Arabidopsis* and rice, to the assembled genomic sequence from poplar, and to the uniref set of protein sequences. Pairs of homologs were selected according to the *blastx* *e*Value. The *Y*-axis reports the total number of pairs of homologs found between the two taxons compared.

Further sequencing of cDNA clones to improve their functional assignment

The complete sequencing of 21 sequences annotated as orphans and over-represented in root xylem was undertaken with the longest available cDNA clones encompassing the corresponding contigs to verify whether more complete sequence information would help in ascribing putative functions. Sequence similarity searches showed that the completed sequences added substantial information for assigning putative functions to six of the 21 orphan sequences (Table 2). For each of these six sequences, we found at least one homologous sequence in the protein database and in the *Arabidopsis* genome. In spite of the supplemental sequencing of cDNAs, the other 15 contigs remained orphans. These sequences were blasted against a recently available *Pinus taeda* L. EST dataset derived from root tissues at University of Georgia (J Dean, L. H. Pratt and M-M. Cordonnier-Pratt, UGA). Four of the 15 orphans sequences were also found in the UGA root dataset, confirming that the sequences are indeed expressed in pine roots where their role remains to be determined.

Sequence analysis of contigs similar to regulatory proteins

We chose to analyse sequence annotation in the root xylem cluster in greater detail because it contained several putative regulatory proteins or transcription factors (Figure A, supplemental data), whereas other clusters contained few or no such proteins. For each sequence in this cluster, we carefully searched for similarities with proteins of experimentally known function and found unexpected domains and motifs. We scanned the

Table 2. Subset of contigs over-represented in root xylem and for which the longest available cDNA clone was completely sequenced. The size and annotation of the contigs (TC, Tentative Consensus) are reported as in PGI1.0.

Contig name	Contig size	Annotation in PGI1.0	Sequenced insert size	Best match with a gene of known function	Function	Score (*e* Value)	Match with PRODOM domains (hit location # domain accession)	Domain annotation	Number of domains in family	Score (*e* Value)
TC2404	771	Similar to GP|8809622 |dbj|BAA97173.1 Myb-related transcription activator-like {*Arabidopsis thaliana*}, partial (22%)	947	gi|24850307|gb| AAN63154.1| [*Oryza sativa*]	Transcription factor MYBS3	139 (1 · e−31)	707–868 #PD202832	DNA-binding nuclear transcription factor MYB-related binding activator I-BOX MYBST1 MYB-family	42	248 (3 · e−22)
							869–952 #PD348865	MYB-related DNA-binding nuclear transcription activator F14D16.15	5	118 (4 · e−07)
							767–907 #PD013002	MYBST1 activator DNA-binding nuclear MYB-related late 5 P1 A_IG002N01.20 hypocotyl CCA1	18	113 (1 · e−06)
TC1249	697	*	1701	*			*			
TC1510	687	*	1375	*			*			
TC1636	737	*	1591	gi|15221638|ref |NP_173809.1| [*Arabidopsis thaliana*]	Zinc finger (C3HC4-type RING finger) protein	66 (2 · e−09)	356–496 #PD000157	Zinc-finger metal-binding finger ring nuclear repeat tripartite DNA-binding similar. This family was built from an expert validated domain	1083	174 (3 · e−13)
TC1716	1498	*	1498	*			*			
TC1898	715	Similar to PIR|T52141|T52141 hypothetical protein [imported]- *Arabidopsis thaliana*, partial (25%)	1034	gb|AAH44477.1| [*Dano rerio*]	SMIF (Smad Interacting Factor)	54 (6^{e}−06)	40–225 #PD129069	Transcription factor similar ECU11_0970 YEAST CG11183 mRNA T27G7.7 SMIF complete	10	237 (7 · e−21)

TC1917	725	*	2021	emb|CAA67781.1| Berg36 [*Homo sapiens*]	CCCH type Zinc finger protein	99 (3 · e−19)	1460–1549 #PD416454	Nuclear zinc-finger metal-binding DNA-binding repeat zinc finger TIS11 factor ZFP-36	79	160 (1 · e−11)
							1466–1648 #PD304307	C3H-4 finger zinc CCCH	1	134 (1 · e−08)
							1559–1651 #PD113276	Nuclear zinc-finger metal-binding DNA-binding repeat zinc finger TIS11 factor ZFP-36	79	169 (1 · e−12)
TC236	732	*	732	*			*			
TC581	720	*	1692	gb|AAN63154.1| [*Oryza sativa*]	Transcription factor MYBS3	212 (2 · e−53)	745–933 #PD202832	DNA-binding nuclear transcription factor MYB-related binding activator I-BOX MYBST1 MYB-family	42	240 (6 · e−21)
							1084–1209 #PD330539	DNA-binding nuclear MYB-related activator	9	176 (2 · e−13)
							937–1089 #PD348865	MYB-related DNA-binding nuclear	5	139 (3 · e−09)
TC927	669	Weakly similar to GP|11139262|gb|AAG31649.1 PRLI-interacting factor A {*Arabidopsis thaliana*}, partial (25%)	1253	gb|AAG31649.1| [*Arabidopsis thaliana*]	PRLI-interacting factor A	213 (5 · e−54)	*			
TC1560	887	*	887	*			*			
TC1774	693	Weakly similar to GP|21644691|dbj|BAC01247. Hypothetical protein˜similar to *Arabidopsis thaliana* chromosome 3 At3g61750, partial (41%)	693	CAC37358.1| [*Solanum tuberosum*]	putative membrane protein, induced during cell necrosis	215 (3 · e−16)	289–585 #PD260839	Membrane 5 DNA P1 chromosome genomic clone:MXH1 MLP3.2 AT4G17280/ DL4675C AT2G30890	21	369 (2 · e−36)
							319–621 #PD014340	receptor AT5G54830/ MBG8_9 LD47639P	9	139 (1 · e−09)
TC2006	668	*	2969	*			*			
TC2053	743	*	770	*			*			
TC2087	732	Similar to GP |13486737|dbj|BAB39972. putative transcription factor (myb) {*Oryza sativa* (japonica cultivar-group)}, partial (31%)	2381		MYB transcription factor		368–529 + 2 #PD000364	DNA-binding nuclear transcription factor MYB-related MYB repeat regulation family MYB-like	258	296 (3 · e−27)

Table 2. Continued.

Contig name	Contig size	Annotation in PGI1.0	Sequenced insert size	Best match with a gene of known function	Function	Score (*e* Value)	Match with PRODOM domains (hit location # domain accession)	Domain annotation	Number of domains in family	Score (*e* Value)
							371–529 + 2 #PD000419	DNA-binding nuclear transcription factor MYB-related MYB repeat regulation MYB-like	537	172 (7 · e−13)
							533–685 + 2 #PD523092	DNA-binding nuclear transcription MYB-related MYB repeat factor regulation activator Proto-oncogene	42	174 (4 · e−13)
TC2104	748	Weakly similar to PIR|B96700|B96700 protein F12A21.17 [imported] - *Arabidopsis tha'liana*, partial (23%)	1404	Ref|NP_008878.1 [*Homo sapiens*]	Signal recognition particle 72kDa	336 (5 · e−30)	414–899 #PD017338	Particle recognition signal SRP72 Ribonucleoprotein homolog F12A21.17 kinase LP10092P	13	273 (7 · e−25)
							933–1298 #PD563822	Recognition signal particle Ribonucleoprotein SRP72 kinase II CAM CG5434 isoform	8	384 (9 · e−38)
TC2136	583	*	642	*			*			
TC2241	699	*	2335	*			*			
TC2257	675	*	1253	ref|NP_189678.1| [*Arabidopsis thaliana*]	Putative non-LTR retroelement reverse transcriptase	55 (3 · e−06)	*			
TC2280	608	*	608	*			*			
TC2322	763	*	763	*			*			
TC2323	561	*	798	*			*			
TC2341	769	*	787	*			*			

TC2371	670	*	3118	ref|NP_566407.1| [*Arabidopsis thaliana*]	polyprotein, GYPSY like Alisei family retrotransposon from *Picea abies*	47 (8 · e−11)	1493–1639 #PD472266	Polyprotein RNA-directed DNA polymerase transferase retroelement POL Reverse GAG-POL transcriptase	28	107 (3 · e−05)
							1695-1934 #PD336416	Transferase retroelement POL	43	114 (5 · e−06)
TC2373	744	*	2116	*			*			
TC2377	564	*	729	*			*			
TC2437	790	Weakly similar to GP|16416383|dbj|BAB70612. anthocyanin-related membrane protein 1 {*Arabidopsis thaliana*}, partial (36%)	1296	dbj|BAB70612.1| [*Arabidopsis thaliana*]	Anthocyanin-related membrane protein 1	494 (2 · e−48)	517-906 #PD058316	Complete proteome transmembrane membrane integral plasmid nodulin-like permease Transporter probable	1194	437 (6 · e−44)
TC2469	751	Similar to PIR |A96794|A96794 unknown protein F14G6.19 [imported]- *Arabidopsis thaliana*, partial (30%)	1143	dbj|BAB69816.1| [*Pisum sativum*]	Zinc-binding protein	237 (1 · e−18)	255–635 #PD447809	F9H16.1 zinc-binding 5 AT1G21000/F9H16_1 P1 F6N18.8 DNA B1108H10.20 B1108H10.23 Clone:MZA15	15	122 (2 · e−07)
							915–998 #PD447809	F9H16.1 zinc-binding 5 AT1G21000/F9H16_1 P1 F6N18.8 DNA B1108H10.20 B1108H10.23 clone:MZA15	15	140 (1 · e−09)
							1002–1139 #PD260020	Genomic F9H16.1 similar thaliana zinc-binding ARABIDOPSIS 5 AT1G21000/F9H16_1 BAC P1	18	207 (2 · e−17)
TC2493	900	Weakly similar to GP|21554742|gb|AAM63677.1 unknown {*Arabidopsis thaliana*}, partial (67%)	1448	gi|18408968|ref |NP_564924.1| [*Arabidopsis thaliana*]	Ubiquitin family	157 (6 · e−37)	530-742 #PD006048	Ubiquitin-like conjugation UBL pathway small SMT3 sentrin ubiquitin-related modifier	37	214 (5 · e−18)
							659–1183 #PD654373	–	1	266 (5 · e−24)
							803–1198 #PD657989	Permease or GABA amino acid	1	151 (1 · e−10)

Table 2. Continued.

Contig name	Contig size	Annotation in PGI1.0	Sequenced insert size	Best match with a gene of known function	Function	Score (*e* Value)	Match with PRODOM domains (hit location # domain accession)	Domain annotation	Number of domains in family	Score (*e* Value)
TC213	583	*	824	gi|4704603|gb|AAD28175.1|[*Picea glauca*]	Dehydrin	133 (4 · e−30)	85-306 #PD244055	Dehydrin	1	261 (9 · e−24)
							433–528 #PD006519	Dehydrin repeat cold seed 2-like group dehydrin/LEA acclimatation RAB multigene	125	112 (2 · e−06)
TC582	1120	Homologue to SP|P24922|IF52_NICPL Initiation factor 5A-2 (eIF-5A) (eIF-4D). [Leadwort-leaved tobacco]{Nicotiana plumbaginifolia}, complete	1120	gb|AAQ08198.1|[*Hevea brasiliensis*]	Eukaryotic translation initiation factor 5A	303 (4·e−81)	324–506 #PD462244	Initiation factor translation Hypusine biosynthesis EIF-5Aeukaryotic 5A multigene family	58	330 (1 · e−31)

Sequence similarity matches reported in the table are not necessarily the matches with the highest scores but the matches with experimentally characterized genes. For each *blast* output, these matches were identified by careful examination of the annotations and if needed, the corresponding publications for all the similar accessions. These matches were considered as more meaningful at the biological level than reporting matches with genes computationally predicted in genomic clones (mainly from *Arabidopsis* and rice). Stars * stand for the sequences with 'no hits found'.

PRODOM database to evaluate their abundance in other proteomes (Table 2). In the end, we found that the root xylem cluster contained seven putative regulatory proteins or transcription factors, i.e. sequences with significant homology to three MYBs, the PRL interacting factor, RING, PLATZ1, BERG6, and SMIF proteins. The putative function of four of them was discovered after complete sequencing of the longest cDNA clones encompassing contigs previously classified as orphans in PGI1.0 (Figure 2C).

MYBs

The orphan contigs TC581 and TC2404 showed similarity to MYB transcription factors upon complete sequencing of the corresponding cDNA clones. Both cDNA clones encoded single Myb-repeat (R1) proteins with the SHAQKYF motif recently found in MYBST1, CCA1 and LHY from *Arabidopsis* (Mercy *et al.*, 2003). The 45 amino acids in the N-terminal region were almost identical between the pine and the angiosperm sequences. A third contig sequence, TC2087, was annotated as a MYB transcription factor in PGI1.0 and was 88% identical to the Orchid *Dendrobium* MYB2, a member of the R2R3-myb class (Wu *et al.*, 2003).

PLATZ1

The complete sequencing of the clone corresponding to TC2469 classified as an 'unknown *Arabidopsis* protein' in PGI1.0, showed that it was similar to the plant specific PLATZ1 pea protein, experimentally shown to be a zinc dependent DNA-binding protein responsible for A/T rich sequence-mediated transcriptional repression (Nagano *et al.*, 2001). The pine sequence partially overlapped a zinc finger domain with a level of identity of 54% (41/75 AAs) and a level of similarity of 72% (54/75 AAs).

BERG6

Several pine sequences were similar to sequences found in plants, but characterized only in animal models. For example, the complete sequencing of a 2021 nucleotide-long cDNA representing the orphan TC1917 gave a match with the human BERG6 protein, which contains a DNA-binding domain. Similarities with other proteins of known function showed that the pine sequence encodes a putative CCCH-type Zn-finger protein. Numerous matches were found both in the *Arabidopsis* and rice genomes; however, none of the plant homologs have been functionally characterized. In contrast, the CCCH-type Zn-finger domain is found in several types of animal proteins documented in the SwissProt database, like the butyrate response factor1 (TIS11), which regulates the response to growth factors in human and mouse, the TTP inducible protein growth factor from mouse, and the human splicing factor U2AF 35 kDa subunit.

SMIF

We found that TC1898 matched SMIF transcription factors known in human, mouse and zebrafish, although it was annotated as similar to an *Arabidopsis* 'hypothetical protein' in PGI1.0. Moreover, we identified several homologs in both the *Arabidopsis* and rice genomes but some of them were mis-annotated as proline-rich proteins (for example, NP_563814.1 from *Arabidopsis*). The complete sequence of a 1034 nucleotide-long cDNA clone helped to identify three regions in the putative protein: a N-terminal-WH1 domain, a large internal putative transcriptional domain and a C-terminal region of unknown function.

We aligned human, mouse and zebrafish SMIF proteins with sequences from pine, *Arabidopsis* and rice and found that the plant sequences contained a region similar to the domain called EVH1/WH1, near the N-terminus. At the protein level, there was 30–40% of identity between animal and homologous plant sequences (Figure 6). The EVH1/WH1 domain is conserved in numerous proteins (in Interpro: EVH1 matches 161 proteins, IPR000697, and WH1 matches 86 proteins, IPR001960) and its structure and function are well documented (review by Ball *et al.*, 2002). We searched for key amino acids implicated in protein function or folding in structurally characterized proteins like MENA (Mouse enabled) and HOMER (rat) (Callebaut, 2002). These WH1 domain markers were conserved in the plant sequences (Figure 6A). The plant sequences contained three aromatic amino acids involved in ligand binding and which specifically interact with ligand proline residues in animal sequences (boxed amino acids in Figure 6A). Several other amino acids required for ligand binding were also conserved in pine and other plants (see arrows and stars in

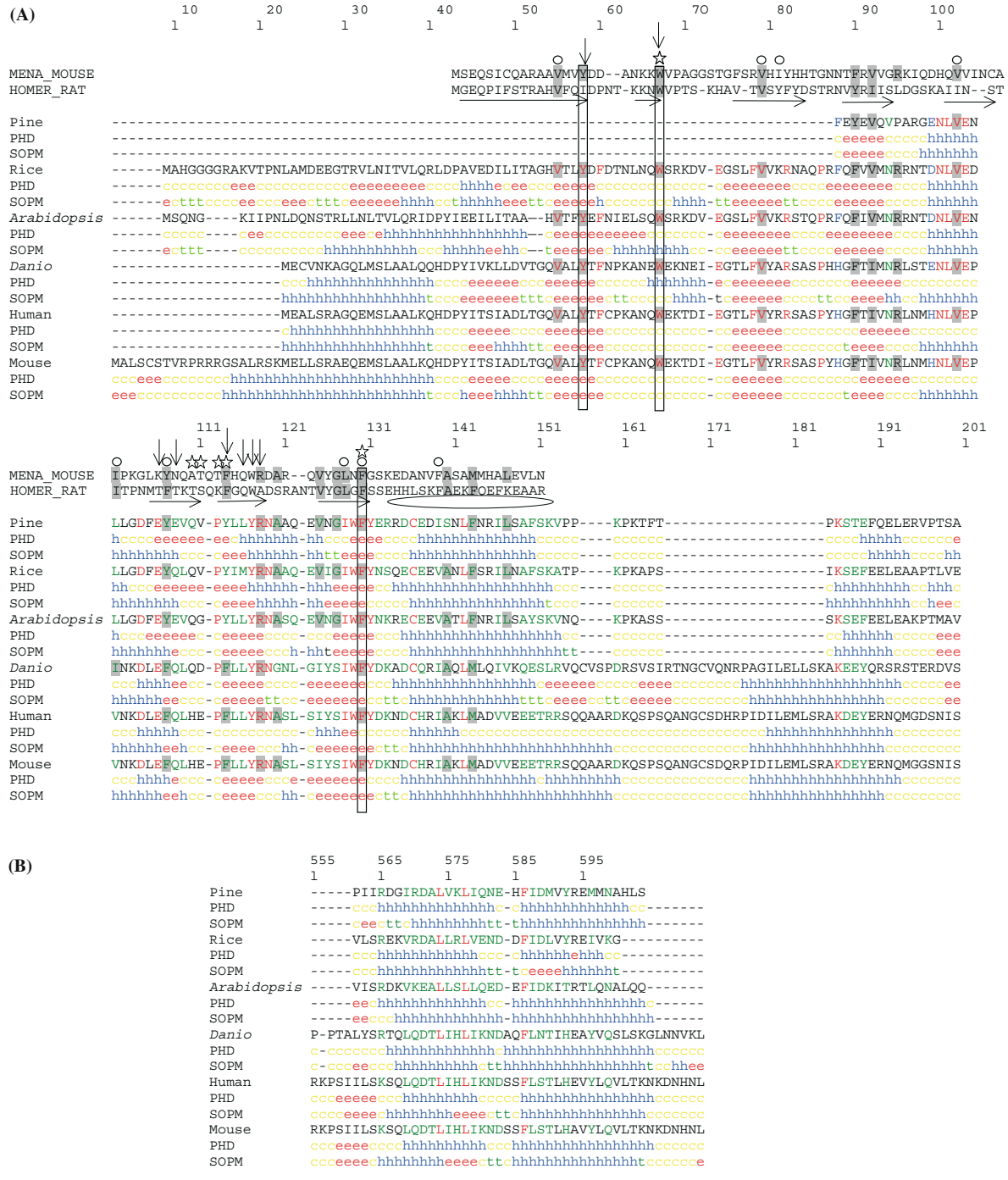

Figure 6A). In animals, these residues are involved in interactions either between the mouse MENA sequence and cytoskeleton related proteins (ActA) or through the interaction between a synaptic terminal protein (HOMER) and glutamate receptors (mGluR). Furthermore, the structure of the WH1 domain, which consisted in six beta sheets and one alpha helix, was highly conserved between plants and animals. For both plant and animal SMIF homologs, we obtained predicted secondary structures with SOPM (Geourjon and Deleage, 1994) and PHD methods

Figure 6. (A) Sequence and structure conservation of the N-terminal region of SMIF sequences from animals and homologous plant sequences. Multiple sequence alignments were performed with the program T-Coffee (Notredame *et al*., 2000). First, MENA, HOMER and SMIF sequences from mouse, human, rice, and pine were aligned. Then the six SMIF sequences presented in the figure were aligned. Finally, the results from each multiple alignment were superposed manually. The upper part of the alignment includes the MENA sequence from mouse and the HOMER sequence from rat for which structural data are available. Functionally and structurally conserved amino acids are shown, those involved in the MENA/ActA interaction ↓ (polyproline-rich domain of bacterial actin/cytoskeleton related protein), HOMER/mGluR interaction ☆ (glutamate receptors). The core residues are symbolized with ○. Secondary structures elements determined for MENA are indicated below the alignments: Beta sheets :→ Alpha helix : ⊂⊃ .Sequence accessions are : MENA (gi: 5107580), HOMER (gi: 13928988), *Danio* (gi: 28277705), human (gi: 8923767), mouse (gi:15617374), rice (gi: 33146979), *Arabidopsis* (gi: 18390886). For sequence identity of 100% or above 50%, the amino acids appear in red and green, respectively. Amino acids that are conserved between the SMIF related sequences and MENA_MOUSE and/or HOMER_RAT are shaded grey. Boxes show the three highly conserved aromatic acids involved in ligand binding. Secondary structure elements predicted by two methods (PHD stands for Profile Network from HeiDelberg and SOPM stands for 'Self-OPtimized Method') are indicated under each sequence. For each amino acid, a letter symbolizes the type of secondary element (h: helix, e: sheets, c: coil, t: turn). (B) C-terminal region of SMIF sequences from animals and homologous sequences from plants. The accessions and the symbols for secondary structure elements are the same as in (A).

(Rost and Sander, 1993) that were in agreement with the experimentally determined structure of the MENA EVH1 domain registered in pdb as 1 evh. These observations confirm that amino acids that are important for the structure or the function of the WH1 domain are also conserved in predicted plant proteins.

The transcriptional activator domain found in the internal region of the SMIF proteins is well conserved in animals. However, conservation of this domain is not obvious in the sequences from pine, *Arabidopsis* and rice, and it appeared to be partially buried in a proline-rich sequence. On the other hand, we found that the C-terminal region of SMIF is highly conserved in sequence between plants and animals. The prediction of the consensus structure suggested the presence of two hydrophobic alpha helices separated by a turn (Figure 6B).

Discussion

Prospecting extensive sets of data generated by EST projects can be undertaken in several ways. Here, we made use of the transcript consensus sequences and their annotations based upon similarity searches and the gene ontology information provided in the *Pinus* Gene Index (PGI1.0). As such, our analysis illustrated the use of large-scale sequence data already structured into public databases and its use for comparative genomics purposes. Comparison of six extensively sequenced cDNA libraries identified sets of sequences that were differentially distributed among cDNA libraries made from different xylem types, or represented at similar levels in the different xylem libraries. Among the 800 most abundantly represented transcripts in PGI1.0, we identified 260 differentially distributed gene sequences that formed five major clusters mostly specific to individual xylem libraries. We discovered several putative markers of xylem development, and our results pertaining to sequences known to be involved in wood formation were consistent with previous studies.

The statistical analysis of ESTs provided an inclusive approach to identify new genes that may play key roles in conifer xylem development, whether a predicted function could be assigned or not to the genes. A subset of 83 of the sequences we identified were orphans that were previously uncharacterized and thus represent putative novel markers of this developmental process. Complete sequencing of cDNA clones representing a targeted group of orphans helped to reveal sequence similarities to known plant or animal genes. The overall strategy described in this report appears useful to establish priorities among genes for further analysis following large-scale sequencing of cDNA libraries, especially when high-throughput mutagenesis and large-scale functional studies are not possible or restricted, such as for conifer taxa.

Distribution analysis of contigs among cDNA libraries

Comparison of results derived from the different statistical tests showed that they produced different results, indicating differences in test sensitivity,

especially between the pairwise AC test and the multiconditional R test (Figure 1). The AC test was the most sensitive, recovering most of the significant cases. In contrast, the R test alone was not sufficient to detect subtle differences in the analysed data. Moreover, out of 246 significant differences obtained by the AC test, only six cases were not indicated by the χ^2 test. This observation was consistent with results based on theoretical and observed data from Romualdi *et al.* (2001), who considered the use of AC and χ^2 as the most appropriate combination to test differential distribution in multiple tag sampling experiments with cDNA libraries. Our results showed that two test combinations (AC–χ^2 and R–χ^2 were adequate to detect all the 260 differentially distributed sequences. The sensitivity of the statistical methods used to test the differential distribution among cDNA libraries was reported to be related to the expression level of the genes (Romualdi *et al.*, 2001). We have also observed that the AC test was especially more effective at recovering contigs with a smaller number of members than the R test (Figure A, supplemental data). The Poisson distribution which is assumed in the AC and R tests distribution was shown to be the most adequate to describe the EST sampling data, resulting in a better fit of these tests (Audic and Claverie, 1997; Claverie, 1999) whereas the F test was reported as too conservative.

Several genes identified in this study could provide potential markers of different states or stages of secondary xylem differentiation. The overall relevance of the statistical procedures used herein to identify these genes is supported by the consistency of our 'digital' results with transcriptome analyses of vascular tissues in trees. For example, our findings were congruent with well documented cases of higher expression of several genes encoding cell wall related proteins in compression wood, based upon analyses of individual genes (McDougall, 2000), ESTs, and other methods such as microarrays (Whetten *et al.*, 2001). In addition to compression wood, our analysis revealed groups of proteins with similar functions or potentially related roles in the plannings library, the latewood library and the root xylem library. We also identified several cases where members of multigenic families were differentially distributed in the six libraries analysed. The utility of the statistical tests used to assess the differential representation of sequences in cDNA libraries was evaluated in simulated datasets, which showed that the detection of false positives was nearly 0 (at the significance threshold of 0.001; Romualdi *et al.*, 2001). The statistical inferences made from the comparison of libraries are a first step in the experimental demonstration of expression patterns. Thus, these putative markers of different conditions or stages of xylem development must be validated experimentally and should not be considered as an exhaustive set of markers. The technical validation of our statistical findings, by verification of gene expression levels in the same tissue samples as those used for the cDNA library synthesis, with a different method like RT-QPCR or northern blotting was not possible. As an alternative, we utilized sequence analyses and expression data obtained in spruce (*Picea glauca* L.) trees (Bedon *et al.*, manuscript in preparation) to provide an independent biological validation for the R2R3-*myb* gene that was differentially representated in the root xylem library. When conducting validation experiments for members of multigenic families like MYBs, using different methods and different species, it is critical to ensure that the sequences represent the closest homolog or a putative ortholog, and that the methods of analysis are specific. The independent RT-QPCR analysis meets these two requirements. Although, it only represented a validation for one of the gene classes, it lended further support to the usefulness of the digital expression analysis presented in this report.

Annotation of conifer sequences

To further explore the composition of the group of 260 differentially distributed sequences and to attempt to improve their annotation with additional sequence similarity searches, we followed two approaches. The first one was a computational analysis which incorporated sequence comparisons with large datasets of available angiosperms and gymnosperms sequences, where we considered all pine contigs, whether they were classified or not as orphan. The second approach was the experimental determination of a longer coding sequence for a subset of pine contigs.

We compared the pine sequences with a large in-house set of spruce sequences. As expected, we

found many more homologous pairs between the two conifer genera pine and spruce than between pine and the angiosperms *Arabidopsis* and poplar, given that genera of the Pinaceae would have diverged 140 Myr BP (Florin, 1963), much later than the early split of 300 Myr BP between the gymnosperm and the angiosperm lineages (Savard *et al.*, 1994). The comparisons of the pine transcripts to the recently sequenced poplar genome added information to relatively few pine sequences. Similar results were obtained with sequence comparisons between *Cycas* and *Arabidopsis* transcripts (Brenner *et al.*, 2003). Based on our observations, the extent of divergence between angiosperm and gymnosperm sequences appears to be a limiting factor for gene annotation in conifers and to assign putative gene functions based on similarity searches alone.

The insufficient length of many pine contig sequences appear to be a serious limitation to assign putative functions based on sequence similarities (Kirst *et al.*, 2003). Thus, we undertook the sequencing of the longuest cDNA clones representing selected contigs annotated as orphans in order to assess whether their annotation could thus be improved. Full length sequencing of selected cDNA inserts enabled to better annotate a few transcripts, but several pine transcripts remained orphan, thus indicating that they may truly represent new genes. The mean size of the 15 sequenced cDNA clones that remained orphans was 1247 nt, with sizes ranged between 608 and 2969 nt. Thus, it is not likely that the orphan status of these sequences could be due to sampling short cDNAs covering mostly UTR regions. The lack of sequence similarity of many differentially distributed xylem sequences could suggest that they represent a specialized set of genes which is less conserved across species. In fission yeast, correlations were observed between gene expression and gene conservation (Mata and Bahler, 2003). These authors showed that many genes conserved in yeast and worm were expressed at high levels. In contrast, a disproportionate number of orphans (yeast specific sequences) were expressed at low levels and their expression levels increased during more specialized processes such as cellular differentiation (Mata and Bahler, 2003). Similarly, specialized physiological processes occurring as secondary vascular growth undergoes different stages may imply a larger proportion of genes supporting more specialized functions, which are more likely to show up as orphans in similarity searches. Similarly, 40% of the sequences obtained from differentiating xylem in poplar had no significant similarity to *Arabidopsis* genes (Hertzberg *et al.*, 2001). Together with our results, these observations suggest that the development and function of the specialized cells of the plant vasculature involve a higher degree of divergence in gene sequence or expression.

Uncovering target genes based upon a digital expression analysis

A candidate gene approach focuses on genes well documented in one species, most often a model system, to propose their study in another species. Our approach, driven by the detection of 'digital expression' patterns, is complementary since it requires no previous knowledge of gene function. Sequences of unknown function thus identified may be included in a subsequent more refined functional analysis. Indeed, by inspecting sets of sequences according to their distribution patterns across cDNA libraries, we were able to sort unexplored sequences and identify putative regulons which included sequences not previously characterized in plants, potentially linking them to the biological process of interest. Thus, digital analysis of ESTs represents a useful starting point toward the functional analysis and annotation of these sequences, which constitutes the large part of the available gymnosperm sequences. Such an analysis should result in the exploration of novel research directions. In the light of knowledge gathered in other models, we briefly discuss some biological hypotheses derived from the *in silico* analysis of pine ESTs that would deserve further functional studies.

A gene encoding a putative regulatory protein with a WH1 domain

Our discovery of a plant gene coding for a protein containing a WH1 domain is an interesting finding, because this gene is homologous to animal genes that are involved in many development and differentiation processes. The SMIF protein consists of a Smad interacting domain and a transcriptional domain, separated by a linker domain. After induction by TGFβ, the Smad–SMIF complex translocates to the nucleus where it

possesses a transcriptional activity, conferring to the SMIF protein a key role in the TGFβ signaling pathway (Bai *et al*., 2002). Proteins which possess the WH1 domain are implicated in cytoskeleton dynamics. For example, the *Drosophila* Mena mutants display defects in the axonal architecture of the nervous system (Ahern-Djamali *et al*., 1998). Mutations in WASP induce cytoskeleton abnormalities in T cells and platelets, and implicate the gene in the implementation of cell fate decisions during *Drosophila* development (Ben-Yaacov *et al*., 2001). Further analyses of gene sequences from pine and other plants are required to address the biological role of SMIF in plants.

RING

Through complete sequencing of a cDNA clones, we also recovered a sequence that was similar to RING proteins, which are involved in the substrate specific degradation via the ubiquitin pathway. The similarity of the pine sequence with the RING proteins encompassed a Zn-finger domain. RING proteins are abundant in *Arabidopsis* but their functions are not well known (Kosarev *et al.,* 2002). Recent studies showed that alfafa *ring-h2* gene was predominantly expressed during the development of roots and symbiotic nodules (Karlowski and Hirsch, 2003). During the growth of the lateral root, *ring-h2* activity appeared to be restricted to the vascular tissues and its promoter had highest activity in the developing vascular bundles.

Myb genes

The expression of more than 90 different R2R3-*myb* genes was examined in more than 20 different tissues and growth conditions in *Arabidopsis* (Kranz *et al*., 1998). However, most of the MYBs harbored poorly characterized biological roles. A few R2R3 MYB transcription factors have been implicated as positive or negative regulators of lignin biosynthesis, a major constituent of wood (Tamagnone *et al*., 1998; Patzlaff *et al.,* 2003). The implication of MYBs in xylem differentiation was also indicated by microarray analysis in poplar, which showed that four out of six *myb* genes exhibited differential expression across four distinct stages of xylogenesis (Hertzberg *et al*., 2001). In contrast to the R2R3 class, R1 MYBs have only recently been linked to xylem differentiation. The *apl* gene (*altered phloem development*) of *Arabidopsis* was shown to be a positive regulator of phloem differentiation and a negative regulator of xylem differentiation (Bonke *et al.,* 2003). However, the pine sequences herein encompassing TC581 and TC2404 showed no similarity to the *apl* gene.

Retroelements

Interestingly, six out of 46 sequences over-represented in the root xylem library were similar to retroelements (Figure 2c). Although the clones encompassing TC2371 and TC2203 were completely sequenced and were of 3 and 2 kb long, respectively, the full length retroelement transcript was not recovered. This result suggests that large EST datasets derived from diversified tissue samples may be a good starting material to identify retrotransposon expression, which has not been reported so far in conifers.

Conclusion

The approach developed in this study clearly illustrates the value of analysing EST collections in more detail, especially in less well characterized genomes. ESTs are too often automatically annotated and perhaps processed too quickly; as a result, the inherent information content appears to remain under-exploited. In this context, a curation procedure such as that described here appears to be essential to uncover genes of interest and which might be missed by automated sequence analysis. The computational approach presented in this study should help to improve the annotation of several contigs of particular interest assembled from these pine libraries. Finally, a few genes not previously characterized in plant systems were highlighted by our analysis of pine genes, indicating that the investigation of distantly related genomes such as those of gymnosperms will in turn contribute to the annotation of gene in model organisms like *Arabidopsis*. Thus, the development of functional and comparative genomic approaches in gymnosperms is likely to contribute to a broader understanding of plant gene functions.

Acknowledgements

The pine xylem cDNA libraries were provided by R.R. Sederoff (NCSU, Raleigh, NC). The

assistance of R. Gauci, É. Fissette, M. Ouellet and S. Blais is acknowledged for sequencing the cDNA clones. We thank F. Larochelle for his help with computer procedures, C. Paule and L. Parsons for the assembly of the spruce ESTs, F. Bedon for sharing data and two anonymous reviewers for their constructive comments. Funding for this work was provided by Genome Canada and Genome Québec to J.M. and J.B. for the project *Arborea*. Accession Numbers: the sequences were submitted to GenBank and the following accession numbers were assigned: CK594759 – CK594760, CK604169 – CK604201.

References

Ahern-Djamali, S.M., Comer, A.R., Bachmann, C., Kastenmeier, A.S., Reddy, S.K., Beckerle, M.C., Walter, U. and Hoffmann, F.M. 1998. Mutations in *Drosophila* enabled and rescue by human vasodilator-stimulated phosphoprotein (VASP) indicate important functional roles for Ena/VASP homology domain 1 (EVH1) and EVH2 domains. Mol. Biol. Cell 9: 2157–2171.

Allona, I., Quinn, M., Shoop, E., Swope, K., St Cyr, S., Carlis, J., Riedl, J., Retzel, E., Campbell, M., Sederoff, R. and Whetten, R.W. 1998. Analysis of xylem formation in pine by cDNA sequencing. Proc. Natl. Acad. Sci. USA 95: 9693–9698.

Altschul, S.F., Maden, T.L., Schaffer, A. A. Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Audic, S. and Claverie, J.M. 1997. The significance of digital gene expression profiles. Genome Res. 7: 986–995.

Bai, R.Y., Koester, C., Ouyang, T., Hahn, S.A., Hammerschmidt, M., Peschel, C. and Duyster, J. 2002. SMIF, a Smad4-interacting protein that functions as a co-activator in TGFβ signalling. Nature Cell Biol. 4: 181–190.

Ball, L.J., Jarchau, T., Oschkinat, H. and Walter, U. 2002. EVH1 domains: structure, function and interactions. FEBS Lett. 19: 45–52.

Bedon, F., Blais, S., Roy, V., Bérubé, H., Morency, M.J. and Mackay, J. 2004. Characterization, expression and phylogeny of xylem expressed R2R3-MYB genes of spruce and pine trees. Plant & Animal Genomes XII Conference, San Diego, CA.

Ben-Yaacov, S., Le Borgne, R., Abramson, I., Schweisguth, F. and Schejter, E.D. 2001. Wasp, the *Drosophila* Wiskott–Aldrich syndrome gene homologue, is required for cell fate decisions mediated by Notch signalling. J. Cell Biol. 152: 1–13.

Bonke, M., Thitamadee, S., Mahonen, A.P., Hauser, M.T. and Helariutta, Y. 2003. APL regulates vascular tissue identity in *Arabidopsis*. Nature 426: 181–186.

Bortoluzzi, S., d'Alessi, F., Romualdi, C. and Danieli, G.A. 2001. Differential expression of genes coding for ribosomal proteins in different human tissues. Bioinformatics 17: 1152–1157.

Bortoluzzi, S., d'Alessi, F. and Danieli, G.A. 2000. A computational reconstruction of the adult human heart transcriptional profile. J. Mol. Cell Cardiol. 32:1931–1938.

Bortoluzzi, S. and Danieli, G.A. 1999. Towards an in silico analysis of transcription patterns. Trends Genet. 15(3): 118–119.

Brenner, E.D., Stevenson, D.W., McCombie, R.W., Katari, M.S., Rudd, S.A., Mayer, K.F., Palenchar, P.M., Runko, S.J., Twigg, R.W., Dai, G. *et al*. 2003. Expressed sequence tag analysis in *Cycas*, the most primitive living seed plant. Genome Biol. 4: R78.

Callebaut, I. 2002. An EVH1/WH1 domain as a key actor in TGFbeta signalling. FEBS Lett. 22: 178–180

Chang, S., Pureyear, J. and Cairney, J. 1993. A simple and efficient method for isolating RNA from pine trees. Plant Mol. Biol. Rep. 11: 113–116.

Claverie, J.M. 1999. Computational methods for the identification of differential and coordinated gene expression. Hum. Mol. Genet. 8: 1821–1832.

Dubos, C. and Plomion, C. 2003. Identification of water-deficit responsive genes in maritime pine (*Pinus pinaster* Ait.) roots. Plant Mol. Biol. 51: 249–262.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95: 14863–14868.

Ewing, R.M., Kahla, A.B., Poirot, O., Lopez, F., Audic, S. and Claverie, J.M. 1999. Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. Genome Res. 9: 950–959.

Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.6a2. Distributed by the author. Department of Genetics, University of Washington, Seattle.

Florin, R. 1963. The distribution of conifer and taxad genera in time and space. Acta Horti. Bergiani 20: 121–312.

Geourjon, C. and Deleage, G. 1994. SOPM: a self-optimized method for protein secondary structure prediction. Protein Eng. 7: 157–164.

Greller, L.D. and Tobin, F.L. 1999. Detecting selective expression of genes and proteins. Genome Res. 9: 282–296.

Hertzberg, M., Aspeborg, H., Schrader, J., Andersson, A., Erlandsson, R., Blomqvist, K., Bhalerao, R., Uhlen, M., Teeri, T.T., Lundeberg, J. *et al*. 2001. A transcriptional roadmap to wood formation. Proc. Natl. Acad. Sci. USA 98: 14732–14737.

Karlowski, W.M. and Hirsch, A.M. 2003. The over-expression of an alfalfa RING-H2 gene induces pleiotropic effects on plant growth and development. Plant Mol. Biol. 52: 121–133.

Kirst, M., Johnson, A.F., Baucom, C., Ulrich, E., Hubbard, K., Staggs, R., Paule, C., Retzel, E., Whetten, R. and Sederoff, R. 2003. Apparent homology of expressed genes from wood-forming tissues of loblolly pine (*Pinus taeda* L.) with *Arabidopsis thaliana* . Proc. Natl. Acad. Sci. USA 100: 7383–7388.

Kosarev, P., Mayer, K.F. and Hardtke, C.S. 2002. Evaluation and classification of RING-finger domains encoded by the *Arabidopsis* genome. Genome Biol. 3, research0016.1-0016.12.

Kranz, H.D., Denekamp, M., Greco, R., Jin, H., Levya, A., Meissner, R.C., Petroni, K., Urzainqui, A., Bevan, M., Martin, C. *et al*. 1998. Towards functional characterisation of the members of the R2R3-MYB gene family from *Arabidopsis thaliana*. Plant J. 16: 263–276.

Mata, J. and Bahler, J. 2003. Correlations between gene expression and gene conservation in fission yeast. Genome Res. 13: 2686–2690.

McDougall, G.J. 2000. A comparison of proteins from the developing xylem of compression and non-compression wood of branches of Sitka spruce (*Picea sitchensis*) reveals a differentially expressed laccase. J. Exp. Bot. 51: 1395–1401.

Megy, K., Audic, S. and Claverie, J.M. 2002. Heart-specific genes revealed by expressed sequence tag (EST) sampling. Genome Biol. 16: 3(9).

Mellerowicz, E.J., Baucher, M., Sundberg, B. and Boerjan, W. 2001. Unravelling cell wall formation in the woody dicot stem. Plant Mol. Biol. 47: 239–274.

Mercy, I.S., Meeley, R.B., Nichols, S.E., Olsen, O.A. 2003. Zea mays ZmMybst1 cDNA, encodes a single Myb-repeat protein with the VASHAQKYF motif. J. Exp. Bot. 54: 1117–1119.

Nagano, Y., Furuhashi, H., Inaba, T., Sasaki, Y. 2001. A novel class of plant-specific zinc-dependent DNA-binding protein that binds to A/T-rich DNA sequences. Nucleic Acids Res. 29: 4097–4105.

Notredame, C., Higgins, D.G. and Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302: 205–217.

Ogihara, Y., Mochida, K., Nemoto, Y., Murai, K., Yamazaki, Y., Shin, I.T. and Kohara, Y. 2003. Correlated clustering and virtual display of gene expression patterns in the wheat life cycle by large-scale statistical analyses of expressed sequence tags. Plant J. 33: 1001–1011.

Olivier, S.G. 1996. From DNA sequence to biological function. Nature 379: 597–600.

Patzlaff, A., McInnis, S., Courtenay, A., Surman, C., Newman, L.J., Smith, C., Bevan, M.W., Mansfield, S., Whetten, R.W., Sederoff, R.R. and Campbell, M.M. 2003. Characterization of a pine MYB that regulates lignification. Plant J. 36: 743–754.

Quackenbush, J., Liang, F., Holt, I., Pertea, G. and Upton, J. 2000. The TIGR Gene Indices: reconstruction and representation of expressed gene sequences. Nucleic Acids Res. 28: 141–145.

Romualdi, C., Bortoluzzi, S. and Danieli, G.A. 2001. Detecting differentially expressed genes in multiple tag sampling experiments: comparative evaluation of statistical tests. Hum. Mol. Genet. 10: 2133–2141.

Ronning, C.M., Stegalkina, S.S., Ascenzi, R.A., Bougri, O., Hart, A.L., Utterbach, T.R., Vanaken, S.E., Riedmuller, S.B., White, J.A., Cho, J. *et al.* 2003. Comparative analyses of potato expressed sequence tag libraries. Plant Physiol. 131: 419–429.

Rost, B. and Sander, C. 1993. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol 232: 584–599.

Rozen, S. and Skaletsky, H.J. 2000. Primer3 on the WWW for general users and for biologist programmers. In: S. Krawetz and S. Misener (Eds.), Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ, pp. 365–386

Savard, L., Li, P., Strauss, S.H., Chase, M.W., Michaud, M. and Bousquet, J. 1994. Chloroplast and nuclear gene sequences indicate late Pennsylvanian time for the last common ancestor of extant seed plants. Proc. Natl. Acad. Sci. USA 91: 5163–5167.

Stekel, G., Git, Y. and Falciani, F. 2000. The comparison of gene expression from multiple cDNA libraries. Genome Res. 10: 2055–2061.

Sterky, F., Regan, S., Karlsson, J., Hertzberg, M., Rohde, A., Holmberg, A., Amini, B., Bhalerao, R., Larsson, M., Villarroel, R. *et al.* 1998. Gene discovery in the wood-forming tissues of poplar: analysis of 5,692 expressed sequence tags. Proc. Natl. Acad. Sci. USA 95: 13330–13335.

Tamagnone, L., Merida, A., Parr, A., Mackay, S., Culianez-Macia, F.A., Roberts, K. and Martin, C. 1998. The AmMYB308 and AmMYB330 transcription factors from *Antirrhinum* regulate phenylpropanoid and lignin biosynthesis in transgenic tobacco. Plant Cell. 10: 135–154.

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

Wu, X.M., Lim, S.H., Yang, W.C. 2003. Characterization, expression and phylogenetic study of R2R3-MYB genes in orchid. Plant Mol. Biol. 51: 959–972.

Zhang, Y., Sederoff, R.R. and Allona, I. 2000. Differential expression of genes encoding cell wall proteins in vascular tissues from vertical and bent loblolly pine trees. Tree Physiol. 20: 457–466.