



## Functional genomics and cell wall biosynthesis in loblolly pine

Ross Whetten\*, Ying-Hsuan Sun, Yi Zhang and Ron Sederoff

Forest Biotechnology Group, 2500 Partners II, Centennial Campus, Campus Box 7247, North Carolina State University, Raleigh, NC 27695, USA (\*author for correspondence; e-mail ross\_whetten@ncsu.edu)

**Key words:** EST sequencing, microarrays, *Pinus taeda*, xylogenesis, wood formation

### Abstract

Loblolly pine (*Pinus taeda* L.) is the most widely planted tree species in the USA and an important tree in commercial forestry world-wide. The large genome size and long generation time of this species present obstacles to both breeding and molecular genetic analysis. Gene discovery by partial DNA sequence determination of cDNA clones is an effective means of building a knowledge base for molecular investigations of mechanisms governing aspects of pine growth and development, including the commercially relevant properties of secondary cell walls in wood. Microarray experiments utilizing pine cDNA clones can be used to gain additional information about the potential roles of expressed genes in wood formation. Different methods have been used to analyze data from first-generation pine microarrays, with differing degrees of success. Disparities in predictions of differential gene expression between cDNA sequencing experiments and microarray experiments arise from differences in the nature of the respective analyses, but both approaches provide lists of candidate genes which should be further investigated for potential roles in cell wall formation in differentiating pine secondary xylem. Some of these genes seem to be specific to pine, while others also occur in model plants such as *Arabidopsis*, where they could be more efficiently investigated.

**Abbreviations:** AGP, arabinogalactan protein; APRP, adhesive proline-rich protein; EST, expressed sequence tags; GRP, glycine-rich protein; OMT, *O*-methyltransferase; PHY, phytoeyanin; PRP, proline-rich protein; XET, xyloglucan endotransglycosylase

### Introduction

Mankind has long relied on wood from forest trees and woody stems of grasses for structural materials. More recently, pulp and paper production has taken on increased importance in the forest products industry; about 30% of the US wood harvest goes to pulp and paper manufacturing (Saltman *et al.*, 1998). The properties of wood are functions of the size, shape and arrangement of cells in wood, as well as the structure and chemistry of the cell walls. The importance of wood as an industrial raw material has motivated extensive analysis of its structure and chemistry (Lewin and Goldstein, 1991; Biermann, 1993; Higuchi, 1997).

Wood formation involves the specialized biosynthesis of a secondary cell wall in xylogenesis of woody

plants. Specialized cell walls are found in many plant organs; however, the formation of secondary thickenings in vascular tissue is particularly important in the evolution of higher land plants. Well-developed vasculature was apparent in fossil trees as early as the Devonian (Meyer-Berthaud *et al.* 1999), and has been the basis for structural support and for water transport, both essential for the large size of woody plants. In gymnosperms such as loblolly pine (*Pinus taeda* L.), only a few cell types are present in secondary xylem (Harada and Côté, 1985). Wood is formed from the terminal differentiation of cells in xylem, a process that typically continues throughout the annual growing season of a woody plant. The specialized nature of the cell walls in woody tissues provides favorable material for the investigation of many aspects of cell wall structure and biogenesis. Wood is almost entirely

composed of cell wall material, and differentiating secondary xylem is rich in cell wall biosynthetic enzymes (Sederoff *et al.*, 1994; Allona *et al.*, 1998; Sterky *et al.*, 1998).

A relatively small number of forest tree species have been subjected to intensive molecular genetic analysis. Trees in general are difficult experimental organisms, due to their large size and long generation times, and so attention has been focused on those species of greatest commercial importance. In the USA, loblolly pine is the most widely planted species, with over  $10^9$  seedlings planted per year. Pines have the additional disadvantage, as experimental organisms, of extremely large genomes, ranging from 20 to 40 pg DNA per haploid genome equivalent (Wakiyama *et al.*, 1993) or about 200–400 times larger than the genome of *Arabidopsis thaliana* (Somerville and Somerville, 2000). The pine genome is rich in repetitive DNA (Kriebel, 1985), at least some of which is due to the presence of abundant retrotransposon (Kamm *et al.*, 1996; Kossack and Kinlaw, 1999). The haploid megagametophyte of pines and other conifers does, however, provide unique advantages for genetic analysis of forest trees in natural and managed ecosystems (O'Malley *et al.* 1996).

The properties of wood, and the process of wood formation, are due to the action of the genes and proteins in differentiating secondary xylem. These genes and proteins need to be studied to understand wood formation, and are potential targets for the directed modification of wood properties. The regulation of these genes in response to developmental and environmental cues is likely to determine variation in wood properties. The tracheid length, diameter and wall thickness all affect the strength and density of wood. When lignin is removed, these properties also determine the strength, coarseness, and density of pulp and paper products.

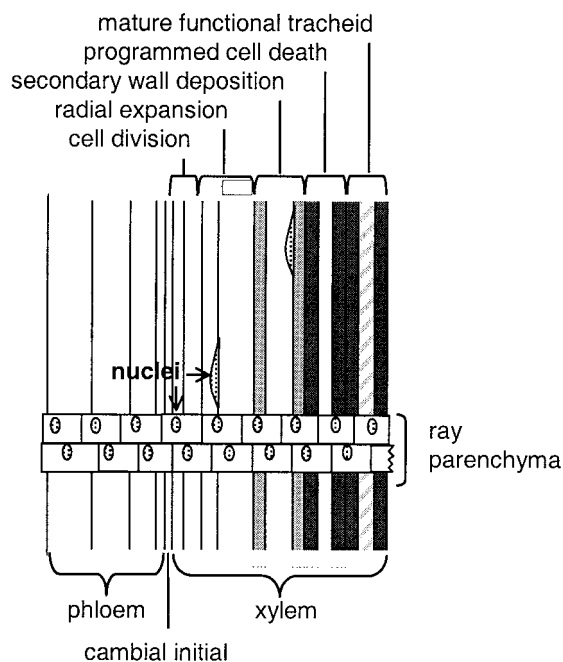
Wood in many trees varies greatly in structure, composition or both, during ontogeny, during the growing season, and under mechanical stress (Megraw, 1985; Zobel and van Buitenen, 1989). During the growing season there are major changes in the structure of the tracheids (springwood versus summerwood), which affect their ability to transport water under wet or dry conditions. In pines, the composition and morphology changes as the vascular cambium ages (the juvenile/mature transition), and in response to mechanical stress (reaction wood). Springwood (or earlywood) is characterized by large lumens and thinner walls and is lower in density than latewood, which

has smaller lumens and thicker walls. Juvenile wood has lower density, shorter fibers and a higher lignin content than mature wood (Zobel and Sprague, 1998).

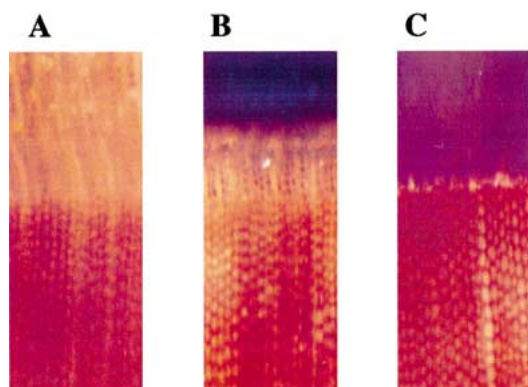
A change in the orientation of a tree stem with respect to gravity frequently stimulates the formation of a specialized type of wood termed reaction wood. The reaction wood formed in pines and other conifers is called compression wood, and is formed on the underside of a bent stem, serving to reorient the stem or branch to a vertical position. Compression wood differs in morphology and composition from normal wood, and from wood on the lateral and upper sides of the bent stem. Tracheids in compression wood are characterized by a round rather than rectangular cross-sectional profile, a higher ratio of secondary wall thickness to cell diameter, a decrease in cell length, an increase in the angle of cellulose microfibrils in the secondary wall to the long axis of the cell, an increase in the fraction of *p*-hydroxyphenyl subunits in lignin and an increase in lignin content (Timell, 1986). Changes in the relative abundance of specific transcripts between normal and compression wood provide insight into the developmental transitions that underlie the modifications in cell wall structure and composition induced by mechanical stress. Side wood does not show these changes and serves as an internal control.

Xylogenesis in pine begins with cell division of two different types of cambial cells, known as fusiform initials and ray initials, to give rise to mother cells which will eventually become tracheids and ray parenchyma cells, respectively. After cell division of fusiform initials and mother cells, the daughter cells undergo radial expansion, secondary wall deposition, lignification and (in the case of tracheids) programmed cell death (Figure 1). Trees have advantages for the study of xylem differentiation because of the ability to obtain large amounts of differentiating xylem for biochemical studies from field-grown trees, allowing genomic approaches to understanding the molecular events in wood formation. During active stem growth of pine, the bark can be removed leaving the immature xylem attached to the wood. Scraping the surface removes the non-lignified differentiating xylem, while deeper planing removes tracheids undergoing lignification and programmed cell death, as well as some mature tracheids and fully differentiated ray cells (Figure 2).

We have used a genomic approach to identifying genes and proteins involved in cell wall biosynthesis during xylogenesis in loblolly pine (Allona *et al.*,



**Figure 1.** A schematic diagram of differentiating pine secondary xylem. The cambium is a meristem comprised of a single cell layer. This layer contains both elongated fusiform initials that give rise to phloem precursors to the outside and xylem precursors to the inside, as well as ray initials that give rise to parenchyma cells that form continuous rays across the xylem and phloem.



**Figure 2.** Photographs of pine stem cross sections stained with phloroglucinol to show lignified cells walls as red. A. A stem section taken before removal of phloem and external tissues. B. A stem section taken after removal of phloem and external tissues, leaving the differentiating secondary xylem on the surface of the wood. C. A stem section taken after harvest of non-lignified differentiating secondary xylem by scraping the surface of the wood with a vegetable peeler. Lignified differentiating xylem remains, but can be harvested with a carpenter's plane.

1998; Zhang *et al.*, 2000). In this report, we present results from identification of large numbers of expressed genes based on cDNA sequencing, and from preliminary microarray analysis of relative expression levels of a subset of these genes. In addition, we have identified and partially characterized several cell wall-associated proteins by this approach.

## Materials and methods

The methods used to obtain differentiating pine secondary xylem, prepare cDNA libraries, and carry out partial DNA sequence determination have been described (Allona *et al.*, 1998). The libraries from which pine ESTs have been sampled to date were constructed from RNA obtained from a variety of tissues and organs, including several different types of differentiating xylem. Non-xylem tissues and organs include immature male strobili, or pollen cones, and shoot tips, or the terminal 2 cm of elongating primary growth from branches. Differentiating xylem samples include compression wood, side-wood (morphologically normal tissue from the side of the same stems, serving as a control), and normal vertical xylem. Pollen cones were collected from lower branches of a single individual mature (ca. 30-year old) tree in the early spring after reaching full size, but prior to dehiscence, and represent late stages of pollen differentiation. Shoot tips were collected from lower branches of a different individual tree (ca. 10 years old) in late spring during the period of active shoot elongation. The pollen cone and shoot tip libraries should therefore contain no more than two alleles for any single gene, because they are each derived from a single diploid individual. The alleles present in the two libraries may differ, however. Compression wood and side-wood libraries were made from pooled RNA samples obtained from three trees (6 years old), as previously described (Allona *et al.*, 1998). The normal vertical xylem library was made from RNA from a single individual tree (ca. 35 years old). Differentiating xylem samples collected from an individual tree were all pooled prior to RNA isolation, i.e. no separation of juvenile wood from mature wood was attempted during library construction. The compression wood and side wood libraries may contain up to six different alleles of a single gene, while the normal vertical wood library should contain no more than two alleles. Sampling from these libraries has largely been random, although an early project did construct subtracted libraries and sequence

relatively small numbers of clones from them (Allona *et al.*, 1998). The fact that most cDNAs were sampled at random from the different libraries allows statistical analysis of the frequency with which particular sequences appear in the resulting data set.

Annotation of the pine EST sequences obtained at North Carolina State University has been carried out in collaboration with Ernest Retzel and his colleagues at the Computational Biology Center of University of Minnesota, and the results are displayed at <http://web.ahc.umn.edu/biodata/doepine/> and <http://web.ahc.umn.edu/biodata/nsfpine/>. One important analysis carried out on the pine EST data sets was assembly of 'contigs', or clusters of overlapping EST sequences all apparently derived from the same mRNA. Contigs were assembled by PHRAP (Ewing and Green, 1997; see also <http://www.phrap.org>). The PHRAP parameters used for contig assembly were a minimum match of 40 and a minimum score of 80; as a result, sequences with more than 98% identity are sometimes placed into different contigs. These sequence differences may be allelic variation, or variation between members of gene families. Sequencing errors are unlikely to result in formation of multiple contigs from otherwise identical ESTs, because the PHRAP program uses error probability scores produced by the PHRED base-calling program in evaluating the probability that two ESTs are representatives of the same sequence. The assembly of contigs used in the analysis presented here used 4557 sequences from the compression wood library and 8490 sequences from the normal vertical wood library.

Contig assembly reduces the numbers of sequences to be analyzed, and can also aid in identifying allelic variation and defining members of multigene families. These contigs represent the more abundantly expressed genes in differentiating pine secondary xylem, and have better coverage of the coding sequences than single ESTs. Analysis of abundantly expressed genes is likely to be more informative regarding function in tissues than trying to analyze all ESTs, because the most abundantly expressed genes in a tissue are good candidates to have important functions in that tissue. Using contigs, we are also less likely to misidentify a gene because only a small amount of sequence has been obtained for that gene. We have restricted our analysis to those ESTs that have at least 200 nucleotides of high-quality sequence (PHRED score greater than 20, equivalent to error probabilities of less than  $10^{-2}$ , per nucleotide) and occur in contigs of 4 or more sequences.

The numbers of specific ESTs found in samples of cDNAs sequenced from different libraries represent an electronic measure of gene expression for many genes at the same time, if analyzed with appropriate statistical methods (Audic and Claverie, 1997). The number of loblolly pine EST sequences similar to a specific pine contig sequence was determined by using the pine contig as a BLASTN query to search the 'otherests' section of GenBank, with 'Pinus taeda' specified in the organism field (Altschul *et al.*, 1997). The numbers of ESTs similar to the query at expect (E) values less than  $10^{-5}$  were counted, and these values used to calculate the probability of differential expression using the method of Audic and Claverie (1997). Briefly, these authors describe the probability of the same rare event occurring twice in two independent trials, based on the Poisson distribution. The frequency of any given cDNA in a non-normalized library is generally sufficiently low that detection of that cDNA by sequencing during an EST project is a rare event. Sequencing from two different libraries constitutes independent trials, and identification of the same cDNA by assembly of EST contigs from different libraries is an example of the same rare event occurring twice. The difference in the number of ESTs corresponding to a particular cDNA between the two libraries may be due to a difference in the representation of that cDNA in the libraries, or it could be due to random chance in sampling. The total number of ESTs obtained from each library, and the number obtained corresponding to the cDNA of interest, are the key variables in calculating the probability that the difference observed is due to differential representation rather than to chance fluctuations. The equation used to calculate the probability of  $y$  occurrences of ESTs derived from a gene of interest in a population of  $N_2$  total sequences from library 2, given the occurrence of  $x$  ESTs derived from the same gene in a population of  $N_1$  total sequences from library 1, and assuming only random fluctuation, is  $p(y|x) = (N_2/N_1)^y [(x+y)! / \{x!y!(1+N_2/N_1)^{(x+y+1)}\}]$ .

Searches for similarity between pine EST sequences and ESTs from poplar were carried out with TBLASTX (Altschul *et al.*, 1997). This program translates the query sequence in all six reading frames, and searches a six-frame translation of the EST collection for similarity between predicted polypeptides. This provides a greater level of sensitivity for detection of diverged coding regions than do similarity searches at the nucleotide sequence level. Twenty pine contigs lacking significant similarity to sequences in GenBank, ten from the 'vertical' collection and

ten from the 'compression wood' collection, were used as queries. These contigs correspond to transcripts present at a frequency of greater than 0.1% in these two cDNA libraries, based on the numbers of sequences comprising each contig. Use of contigs as queries reduces the probability that similarity would fail to be detected simply because the query sequence is too short to overlap with possible similar ESTs in the database, and the use of the TBLASTX search tool provides a sensitive search for similarity at the protein level. The ten contigs from the vertical EST collection were 1460, 1463, 1493, 1496, 1500, 1507, 513, 1538, 1549, and 1553, from the 23 August 2000 contig set presented at [http://web.ahc.umn.edu/biodata/nsfpine/contig\\_dir0](http://web.ahc.umn.edu/biodata/nsfpine/contig_dir0). The compression wood contigs were 629, 634, 644, 648, 656, 661, 668, 679, 694, and 723, from the September 19, 2000 contig set presented at [http://web.ahc.umn.edu/biodata/nsfpine/contig\\_dir1](http://web.ahc.umn.edu/biodata/nsfpine/contig_dir1).

Comparisons of relative levels of gene expression between pine and *Arabidopsis* were carried out with the full-length GenBank sequences most similar to individual pine EST contigs as TBLASTN queries to search both *Arabidopsis* and pine EST collections in dbEST. The numbers of ESTs from each organism similar to the query sequence at E values of less than  $10^{-5}$  were counted for use in the statistical comparisons. This approach provides a first approximation of the relative abundance of entire classes of mRNAs, but does not discriminate well between different members of gene families. The method does provide an overview of the relative abundance of RNAs encoding particular metabolic functions. Metabolic function is typically an important criterion for interpreting the biological significance of any predicted differential abundance of mRNAs based on relative frequency of ESTs, so the fact that the estimates of differential abundance presented here represent entire gene families rather than individual genes focuses attention on metabolism rather than on specific genes. Complex gene families are relatively common in pine (Kinlaw and Neale, 1997), and high-throughput methods for accurately determining the contributions of individual members of gene families to overall levels of gene expression will be needed to explore questions of differential gene regulation more fully.

#### *Microarray methods*

Methods for printing PCR-amplified cDNA inserts from plasmid clones onto glass slides in high-density

arrays have been described (Winzeler *et al.*, 1999). The first-generation pine microarrays described here were printed at the Carnegie Institute of Washington, in collaboration with Shauna Somerville's group. These arrays consist of about 3000 pine cDNAs; 1100 of these clones were derived from a suppression subtractive PCR experiment between differentiating juvenile and mature secondary xylem (Sun, unpublished work based on the method of Diatchenko *et al.*, 1996) and the remaining clones were those available from early EST sequencing projects (Kinlaw *et al.*, 1996; Allona *et al.*, 1998). No selection of cDNA clones from the available pool was done; instead, all available cDNAs from differentiating pine xylem were arrayed. Two different pairs of samples were used in these experiments, representing one pair of developmental stages and one environmental stimulus. The developmental stages examined were differentiating juvenile wood, from the apical two meters of three 20-year old trees (of different genotypes), and differentiating mature wood from the basal two meters of the same trees. The environmental stimulus examined was mechanical stress, applied to one ramet of a pair of clonal individuals, 6 years of age, by bending one tree away from a vertical position and tying it to a stake driven into the ground so that the apical portion of the stem was parallel to the ground. The control ramet remained in a vertical orientation for the 6 days of treatment, and both trees were then harvested. The final data set used for most of the analyses presented here includes data from three complete replicates for all three genotypes in the juvenile-mature comparison and four replicates for the compression wood versus normal wood comparison. After removal of data points for which various quality control measures indicated potential problems, a final set of about 2400 elements remained suitable for analysis.

Raw images acquired from the arrays were analyzed with ScanAlyze (M.B. Eisen; <http://rana.lbl.gov/>), and the two signal channels were balanced to yield equivalent total signal summed across all elements. Background was adjusted by a local regression method to be described in detail elsewhere (Sun, manuscript in preparation). The raw signal intensity data and background-adjusted normalized data are available by anonymous ftp from <ftp.ncsu.edu> in the directory [/pub/unity/lockers/ftp/rosswhet](ftp://pub/unity/lockers/ftp/rosswhet). Testing for differential gene expression was carried out by several approaches. The first compared ratios of the signal intensity from a 'treatment' cDNA preparation to that obtained from a 'control' cDNA preparation for each

element on the same array. This method of comparison of signal intensity ratios within the data set derived from each individual array will be referred to as 'within-slide' comparison. This analysis incorporated data from replicated arrays by testing the consistency with which a particular element yielded a ratio different from the mean ratio of all elements on each array. This method identified genes that changed in relative level of expression by a greater amount than the other genes tested on the array, focusing attention on the largest changes in relative expression levels. For the 'within-slide' test, array elements were designated as differentially expressed when the natural logarithm of the ratio ( $\ln$  ratio) of background-corrected signal intensities on a given slide differed from the mean  $\ln$  ratio for that slide by a set amount on at least 50% of the replicate slides. The criterion for deviation from the mean was 1.6 times the slide standard deviation for comparison of transcript levels between vertical wood RNA and compression wood RNA, and 1.0 times the slide standard deviation for comparison of transcript levels in juvenile wood RNA and mature wood RNA. This difference in criterion was determined subjectively, based on the variability of replicated slides within each set of treatments.

An alternative approach uses statistical methods to determine the reproducibility of each measurement, both within an individual array and across duplicate arrays. This recognizes that the largest magnitude of change in relative gene expression level is not necessarily the most interesting, and a reproducible change may be interesting regardless of its relative magnitude. One relatively simple version of this statistical approach is to compare the ratio of channel intensities for replicates of each element (across multiple arrays) to that of a set of replicated 'control' elements that do not show significant differences in signal between the RNA samples being compared. This requires some means of normalizing relative intensities between arrays, and so positive control elements on the array and corresponding synthetic mRNAs added to each labeling reaction are essential to this simple approach. It is also possible to use statistical methods to account for experimental variation, for example, in labeling efficiency. Appropriate models can be used to assign components of the observed variation in signal intensities to the various experimental sources, and variation not accounted for by noise is used as an estimate of differences in transcript levels. Several groups are pursuing this approach in slightly different ways, ranging from Bayesian methods (Friedman *et al.*, 2000; New-

ton *et al.*, 2000) to ANOVA methods based on those used in agricultural field trials (Kerr *et al.*, 2000; Kerr and Churchill, 2000; Wolfinger *et al.*, 2000).

The statistical test (Dunnett, 1955) used in this study yields an experiment-wide significance level, so that the chosen significance level (e.g.  $P < 0.05$ ) signifies the probability of Type I error for the entire set of genes designated as differentially expressed, rather than for each gene individually. This is particularly important for microarrays with hundreds or thousands of elements, because of the many comparisons that must be made. A probability of 0.01 of Type I error in each of 2000 comparisons allows an average of 20 false-positive errors, which can be a major problem if only 200 genes are determined to be differentially expressed. Dunnett's test was conducted by identifying a set of 68 elements as a 'control set', defined as elements whose ratios varied from the mean by greater than 0.5 standard deviation no more than once in the four slides used for the mechanical stress experiment. None of the control set ratios deviated from the mean by 0.5 standard deviation in any of the ten slides used for juvenile-mature comparisons. The ratios from all replicates of each element were tested for significant difference from the ratios of the control set.

## Results

### *EST sequencing in loblolly pine*

In October 2000, the dbEST section of GenBank contained 22 233 loblolly pine ESTs, a substantial increase from the 750 ESTs present in February 1998. About 60% of these were from differentiating xylem, 5624 from a shoot tip library, and 1507 from a pollen cone library. Several libraries have been made from wood forming tissues under different conditions, for comparison to normal vertical wood formed in the spring. Sequencing of pine ESTs is ongoing, and the numbers of pine sequences in GenBank will continue to increase. These EST collections have been used to address questions of unique genes in pines, relative gene expression in different types of differentiating xylem, and relative levels of gene expression in pine versus *Arabidopsis*.

### *Are there genes unique to gymnosperms or to differentiating secondary xylem?*

Only recently have EST projects in pine and poplar begun to provide information on the genes expressed

in woody plants in wood-forming tissues, to address the question of whether there may be unique genes expressed during secondary xylem formation (Allona *et al.*, 1998; Sterky *et al.*, 1998; Mellerowicz *et al.*, 2001, this issue). There has also been little information on the possible number of unique genes present in pines and other gymnosperms that are not found in herbaceous or woody angiosperms. One approach to the first question is to ask about the number of expressed genes not found in herbaceous annuals, but common to pines and woody angiosperms. The current EST collections from poplar and pine are not sufficient to allow a comprehensive answer to this question, but do provide some insight into the relationships between abundant transcripts in differentiating secondary xylem of these two tree species (see Mellerowicz *et al.*, 2001, this issue).

About 7% of contigs from loblolly pine show no similarity to DNA or protein sequences in GenBank, as of August 2000. These are candidates for genes that are either specific to gymnosperms, to woody tissues, or to both. Comparison of the most abundant of these pine sequences with poplar ESTs was carried out to test the hypothesis that some of these genes are common to woody tissues of both gymnosperms and woody dicots. Sterky *et al.* (1998) described a collection of EST sequences from differentiating poplar secondary vascular tissues. These authors noted that almost 350 of the set of 5692 ESTs showed no similarity to any sequences in the public DNA and protein sequence databases, and that three of these unknown genes were among the most common transcripts detected. Only one of the twenty pine contigs tested, contig 648 from compression wood, was similar to a poplar EST; the other 19 pine contigs showed no significant similarities to any poplar EST.

#### *Relative levels of gene expression in compression wood and normal wood*

Contigs encompassing multiple ESTs can be used to estimate significant differences in relative abundance of cDNAs corresponding to a particular gene in different libraries. It is not possible to attribute statistical significance to the presence or absence of single ESTs (Audic and Claverie, 1997). Comparisons of sequence similarity among the 50 largest contigs of the compression wood and vertical wood libraries (i.e. the 50 contigs from each library comprised of the most sequences) show that a total of 69 different sequences are present, after removing redundancy within and

between the contig sets (Table 1). Of these 69 most abundant transcripts, 33 are likely to be more abundant in the compression wood library than in the normal wood library ( $P < 0.05$  for difference due to random fluctuation, as calculated by the method of Audic and Claverie, 1997), and only three are likely to be more abundant in the normal wood library than in the compression wood library by the same criterion. The relevance of these putative differentially expressed genes to cell wall formation remains to be proven, but the identities assigned by BLAST searches suggest roles for many of them in cell wall biosynthesis. Among the cDNAs likely to be more abundant in the compression wood library are several related to monolignol biosynthesis (4-coumarate CoA ligase, caffeoyl CoA *O*-methyltransferase, glycine hydroxymethyltransferase, and *S*-adenosyl methionine synthetase) and several putative cell wall proteins (described in more detail in Table 2). Two pine cDNAs with no similarity to any protein or DNA sequence, and three pine cDNAs similar to unknown *Arabidopsis* genes predicted from genomic and EST sequences are also more abundant in the compression wood library. The three cDNAs found to be more abundant in the vertical wood library are predicted to encode a Skp1-like protein, a xyloglucan endotransglycosylase (XET)-like protein, and a protein similar to pollen allergens. Skp1 of yeast is involved in the ubiquitin-proteasome protein turnover pathway (Bai *et al.*, 1996) and the *Arabidopsis* homologue is expressed in meristematic tissues (Porat *et al.*, 1998). *Arabidopsis* Skp1-like proteins have also been implicated in ubiquitin-mediated proteolysis (Schouten *et al.*, 2000), but a strong connection with cell wall formation is not clear. A relationship of an XET to cell wall restructuring seems clear, however, and the change in abundance of the pollen-allergen-like cDNA seems to implicate it also in some aspect of the morphological or chemical differences between compression wood and vertical wood.

Contigs are also useful to identify allelic variation and to identify gene families. The PHRAP program parameters used to assemble pine EST sequences into contigs resulted in creation of different contigs if otherwise identical sequences contained a few single nucleotide differences in DNA sequence. There are several cases of groups of pine EST contigs which show considerable similarity in GenBank hits among the members of the group (Table 1). Such groups may represent allelic variation or duplicate members of a multigene family. Members of multigene families would be expected to show higher levels of sequence

*Table 1.* Contigs of abundant ESTs analyzed for differential expression. The probabilities shown in column four are calculated by the method of Audic and Claverie (1997). Values shown are the upper limits of the probability of observed numbers of ESTs arising by chance alone if the cDNA is equally represented in both libraries being compared. The table shows the presence of a contig in both libraries only when it was among the top 50 contigs ranked by size (number of ESTs) in both libraries. The contigs are ordered in the table from rank 50 of normal wood to rank 1 of normal wood, then rank 50 of compression wood to rank 1 of compression wood, with appropriate cross-references to other contigs similar to the same GenBank record. Contigs shown as present only in the top 50 contigs of one library are still similar to ESTs from the other library, but the corresponding contig was not large enough to rank in the top 50.

Library, line and contig number	Hit in GenBank	Accession number	Differentially expressed?	Number of ESTs in contig
Normal xylem: 1509	probable aquaporin	O65045	no	11
1510	probable thioredoxin H	O65049	no	11
1511	3-deoxy-D-arabino-heptulosonate 7-phosphate synthase	O24051	no	11
1512, 1559	allergen-like protein	AAF16869	$P < 0.001$ , normal	11+48
1513	no hits		no	11
1514	calmodulin 3	CAA09302	no	11
1515, 1541	ADP-ribosylation factor	D17760	no	12+17
1516, comp. wood 748	4-coumarate CoA ligase	T09775	$P < 0.001$ , compression	12+17
1517	GA-regulated protein	O82328	no	12
1518	T19F11.6 protein	AC009918	no	12
1519, comp. wood 732, 751	AGP-like protein	S52995	$P < 0.001$ , compression	12+9+18
1520, 1525, 1531, 1532	S-adenosylmethionine synthase	P50300	$P < 0.001$ , compression	12+13 +14+14
1521, comp. wood 731	Low-MW heat shock protein	S71768	$P < 0.001$ , compression	12+8
1522	cellulase	T07612	no	13
1523, comp. wood 718	R40g2 from rice	T03960	no	13+7
1524	ELI-3	O82550	no	13
1526, 1527, comp. wood 719	methionine synthase	Q42699	no	13+13+7
1528, comp. wood 734	putative laccase	Q9SIY8	no	13+9
1529	fructose-bisphosphate aldolase	T12416	no	13
1530	MJK13.14 protein	AAF35414	$P < 0.001$ , compression	14
1533	unknown protein	BAA92731	no	14
1534	pine LP6 protein	Q41083	no	14
1535	putative UDPG-glucosyltransferase	Q9SK82	$P < 0.01$ , compression	14
1536, comp wood 728	actin-depolymerizing factor	P30175	nNo	15+8
1537	Skp1-like protein	AF135596	$P < 0.02$ , normal	15
1538	no hits		$P < 0.001$ , compression	16
1539	caffeoyl-CoA OMT	AAD02050	$P < 0.02$ , compression	16
1540	aquaporin	O81186	$P < 0.001$ , compression	17
1542	dTDP-glucose 4,6-dehydratase	CAB61752	no	18
1543	expansin	Q9XGI6	see Table 2	18
1544, comp wood 755	glycine hydroxymethyltransferase	B71400	$P < 0.001$ , compression	19+22
1545, comp wood 754, 756	S-adenosylhomocysteine hydrolase	O23255	$P < 0.001$ , compression	20+21+23
1546	actin	AF172094	no	21
1547, comp wood 741	translationally controlled tumor protein	AJ012484	no	22+13

Table 1 continued.

Library, line and contig number	Hit in GenBank	Accession number	Differentially expressed?	Number of ESTs in contig
1548, comp wood 735	phenylcoumaran benzylic ether reductase	AAF64176	no	22+9
1549	no hits		no	22
1550	dicyanin	AAF66242	no	24
1551, 1554, 1555, comp wood 717, 744, 757	$\alpha$ -tubulin	P33629	$P < 0.001$ , compression	24+28+30 +7+13+28
1552	unknown protein	O04324	no	24
1553	no hits		no	27
1556	XET	S61555	$P < 0.01$ , normal wood	30
1557, comp wood 713	polyubiquitin	CAB81047	no	31+7
1558, comp wood 745	elongation factor 1- $\alpha$	AAD56020	$P < 0.01$ , compression	35+14
Compression wood 708	60S ribosomal protein L12	AB005246	$P < 0.002$ , compression	7
709	allyl alcohol dehydrogenase	AB036735	$P < 0.001$ , compression	7
710	histone H2A	P35063	$P < 0.004$ , compression	7
711	glutamine synthetase	AJ005119	$P < 0.03$ , compression	7
712	collagen 1	A48295	repetitive sequence <sup>a</sup>	7
714	glycine decarboxylase H-protein	AC004667	$p < 0.05$ , compression	7
715	histone H3.2	P11105	$P < 0.03$ , compression	7
716	glyceraldehyde 3-phosphate dehydrogenase	S51836	$P < 0.001$ , compression	7
720	$\beta$ -tubulin	U76746	$P < 0.04$ , compression	7
721	nucleotide translocator	444790	$P < 0.03$ , compression	8
722	acidic ribosomal protein P2a-2	U62748	$p < 0.001$ , compression	8
723	no hits		$P < 0.001$ , compression	8
724	cytosolic malate dehydrogenase	T02935	$p < 0.02$ , compression	8
726	unknown protein	AC001645	no	8
727	unknown protein	AC007017	$p < 0.001$ , compression	8
729, 738	AGP4	AF101790	see Table 2	8
730, 747	glycine-rich RNA binding protein	AF109917	$p < 0.001$ , compression	8+16
733	methionine synthase	P93263	$p < 0.04$ , compression	9
736	proline-rich protein	AF101789	see Table 2	9
737, 746	peptidyl-prolyl isomerase	S54833	$P < 0.001$ , compression	9+14
739	unknown protein	AC009918.6	no	10
740	porin Mip1	T14863	$P < 0.002$ , compression	12
742, 750	AGP6	AF101785	see Table 2	13+17
743	RNA-binding protein 3	T15047	$P < 0.001$ , compression	13
749	unknown protein	AAF30339	$P < 0.001$ , compression	13
753	tonoplast intrinsic protein	T10804	$p < 0.001$ , compression	20

<sup>a</sup>The sequence of this contig was so repetitive as to prevent meaningful comparisons of numbers of similar sequences in the two EST collections.

variation than allelic variants, but this assumption has not yet been extensively tested in gymnosperms. In pines it is possible to distinguish between these alternatives by analysis of the haploid megagametophyte, which is derived from the same meiotic product as the maternal contribution to the zygote. Any individual megagametophyte should contain only one allele from a single locus, but will contain alleles from all members of a gene family (O'Malley *et al.*, 1996). The challenge will be to carry out segregation analysis on thousands of different sequence variants identified in EST collections to determine which represent allelic variation and which are different gene family members. A high-throughput, cost-effective method for parallel identification of thousands of sequence variants in small amounts of genomic DNA from a few dozen individuals will be essential to gaining a more complete understanding of the size and complexity of gene families in pine.

*What types of genes are abundantly expressed in wood-forming tissues?*

Among the most abundantly expressed genes in wood-forming tissues are many genes expected to be involved in the formation of the wood cell wall. Genes involved in monolignol precursor biosynthesis are found, including genes encoding 4-coumarate CoA ligase (4CL, contig 1516 in Table 1) and caffeoyl-CoA *O*-methyltransferase (CCoAOMT, contig 1539 in Table 1). Other genes of less certain relationship to lignin biosynthesis are also highly expressed, such as a gene (contig 1524) very similar to ELI3 of parsley, which encodes a protein with benzyl alcohol dehydrogenase activity (Somssich *et al.*, 1996). At least one transcript encoding a protein similar to laccase is highly abundant, while none of the abundant transcripts encode peroxidases, consistent with findings of pilot-scale EST studies of pine and poplar (Allona *et al.*, 1998; Sterky *et al.*, 1998; Mellerowicz *et al.*, 2001, this issue). Also consistent with these previous studies is the finding of abundant transcripts related to methyl-transfer reactions involving *S*-adenosyl methionine (SAM), such as methionine synthetase, *S*-adenosylmethionine synthetase, glycine hydroxymethyltransferase, and *S*-adenosylhomocysteine hydrolase. These results suggest that the supply of methyl groups for lignin biosynthesis is a significant factor for the high level of carbon flux into lignin and that one-carbon metabolism could affect the ratio of methylated and unmethylated lignin

precursors. SAM is also a precursor for other biosynthetic pathways, as well as a methyl group donor, and the prevalence of enzymes involved in SAM formation and turnover may be a signal that some of these other pathways are important in xylem formation as well.

It might also be expected that genes involving formation of polysaccharides would be abundant, and several such genes are present, including genes homologous to XET (contig 1556, vertical wood), cellulase (contig 1522, vertical wood), and RGP1 (contig 1153, vertical wood), a gene thought to be involved in xyloglucan biosynthesis (Dhugga *et al.*, 1997). Transcripts encoding expansin (EXP), proline-rich protein (PRP), glycine-rich protein (GRP), adhesive proline-rich protein (APRP) and phytoecyanin (PHY) are also abundant in differentiating xylem (Zhang *et al.*, 2000). It is less expected to find that several transcripts predicted to encode arabinogalactan (AGP) proteins are highly expressed.

Loopstra and Sederoff (1995) identified two abundant and specifically expressed AGPs in differentiating xylem. Zhang *et al.* (2000) and Loopstra *et al.* (2000) have now identified a total of six different genes encoding AGP-like proteins that are abundantly expressed in differentiating xylem. AGPs are unusual in the repeated motif protein structure, and in the high level of glycosylation. Many AGPs contain a sequence indicating the presence of a glycerophosphatidylinositol (GPI) anchor, for attachment to the cell membrane. All are abundant in the immature xylem cDNA libraries, compared to shoot tips. Some increase in expression significantly in the formation of compression wood compared to normal wood (Table 2). These results suggest an important but as yet undefined role of AGPs in the differentiation of pine secondary xylem, perhaps in formation of the secondary cell wall. In one case, the immunolocalization of AGP6 is restricted to radially expanded cells just preceding the thickening of their secondary walls (Y. Zhang, unpublished results). Sequences encoding phytoecyanin-like proteins are of interest because they contain both a domain similar to proteins associated with cell walls, and a laccase-like domain, both associated with activities thought to have a role in wood formation (O'Malley *et al.*, 1993; Zhang *et al.*, 2000).

*Comparing gene expression in differentiating pine xylem with gene expression in Arabidopsis*

Database searches can contribute ideas about possible gene functions, by allowing researchers to compare

Table 2. Significance testing of digital differential gene expression profiles. The estimates of relative abundance of mRNA for AGPs and some other cell wall-associated proteins are compared based on numbers of corresponding ESTs found in pine cDNA libraries from differentiating normal xylem, differentiating compression xylem, shoot tips, and immature pollen cones. Values in the table are probabilities as in Table 1. Values in bold indicate significantly higher gene expression in the experimental condition on the corresponding row than in the condition in the corresponding column. Values in italics indicate higher gene expression in the experimental condition given in the column relative to the row.

	Compression	Shoot tip	Pollen cone		Compression	Shoot tip	Pollen cone
<b>AGP1</b>				<b>PRP1</b>			
vertical	<i>P</i> < 0.3	<b><i>P</i> &lt; 0.002</b>	<i>P</i> < 0.2	vertical	<i>P</i> < 0.007	<i>P</i> < 0.3	<i>P</i> < 0.7
compression		<i>P</i> < 0.07	<i>P</i> < 0.6	compression		<b><i>P</i> &lt; 0.001</b>	<i>P</i> < 0.06
Shoot tip			<i>P</i> < 0.5	shoot tip			<i>P</i> < 0.8
<b>AGP2</b>				<b>GRP2</b>			
vertical	<i>P</i> < 0.001	<i>P</i> < 0.006	<i>P</i> < 0.2	vertical	<i>P</i> < 0.7	<i>P</i> > 0.9	<i>P</i> < 0.8
compression		<i>P</i> < 0.001	<b><i>P</i> &lt; 0.03</b>	compression		<i>P</i> < 0.7	<i>P</i> < 0.6
Shoot tip			<i>P</i> < 0.09	shoot tip			<i>P</i> < 0.8
<b>AGP3</b>				<b>EXP1</b>			
vertical	<b><i>P</i> &lt; 0.05</b>	<b><i>P</i> &lt; 0.001</b>	<b><i>P</i> &lt; 0.03</b>	vertical	<i>P</i> < 0.9	<i>P</i> < 0.3	<i>P</i> < 0.3
compression		<i>P</i> < 0.06	<i>P</i> < 0.4	compression		<i>P</i> < 0.3	<i>P</i> < 0.3
Shoot tip			<i>P</i> < 0.8	shoot tip			<i>P</i> < 0.7
<b>AGP4</b>				<b>APRP1</b>			
vertical	<i>P</i> < 0.001	<b><i>P</i> &lt; 0.006</b>	<i>P</i> < 0.6	vertical	<i>P</i> < 0.5	<i>P</i> > 0.02	<i>P</i> < 0.5
compression		<b><i>P</i> &lt; 0.001</b>	<b><i>P</i> &lt; 0.004</b>	compression		<i>P</i> < 0.2	<i>P</i> < 0.8
Shoot tip			<i>P</i> < 0.3	shoot tip			<i>P</i> < 0.5
<b>AGP5</b>				<b>PHY2</b>			
vertical	<i>P</i> < 0.2	<b><i>P</i> &lt; 0.001</b>	<b><i>P</i> &lt; 0.05</b>	vertical	<i>P</i> < 0.2	<b><i>P</i> &gt; 0.009</b>	<i>P</i> < 0.08
compression		<b><i>P</i> &lt; 0.001</b>	<b><i>P</i> &lt; 0.007</b>	compression		<i>P</i> < 0.5	<i>P</i> < 0.5
Shoot tip			<i>P</i> < 0.9	shoot tip			<i>P</i> < 0.8
<b>AGP6</b>							
vertical	<i>P</i> < 0.001	<b><i>P</i> &lt; 0.002</b>	<i>P</i> < 0.2				
compression		<b><i>P</i> &lt; 0.001</b>	<b><i>P</i> &lt; 0.001</b>				
Shoot tip			<i>P</i> < 0.5				

apparent expression levels between organisms. There are over 111 000 *Arabidopsis* ESTs in the dbEST division of GenBank (as of August 2000), derived from a variety of different libraries. TBLASTN searches were conducted using full-length protein sequences to which abundant pine ESTs show similarity (angiosperm protein sequences were used when possible, to avoid biasing search results against *Arabidopsis*). Some genes expressed in both pine and *Arabidopsis* are far more abundant in the pine EST data set, largely derived from wood-forming tissues, than in the *Arabidopsis* EST data set, derived from many different tissues (Table 3). These differences in abundance provide evidence for a potential role of these gene products in secondary cell wall biosynthesis or some other aspect of pine secondary xylem differentiation. Linking the pine ESTs to databases which organize and curate the growing volume of information about

other plant genes will be an important step for the future, to take full advantage of all the information gained in model plant systems.

#### Microarray results

Comparison of the 'within-slide' and Dunnett's test analyses of pine microarray data shows that the methods agree in large part, but that the Dunnett's test is more conservative in declaring a particular gene to be differentially expressed. In the vertical wood versus compression wood comparison, 156 of the 2300 elements were designated as differentially expressed by the within-slide test, 85 up-regulated and 71 down-regulated in compression wood. Dunnett's test confirmed up-regulation of 36 of the 85 elements called up-regulated by the within-slide test, but called 6 of the 85 down-regulated. Of the 71 elements called

Table 3. Comparisons of EST abundance in *Arabidopsis* and in pine. The table shows protein sequences used as search queries, the number of ESTs from *Arabidopsis* and pine that are similar to each query at an E value of less than  $1 \times 10^{-5}$ , the fold-difference in representation of that EST in the pine collection versus the *Arabidopsis* collection, and the probability that the difference in EST abundance is due to chance alone (Audic and Claverie, 1997). These calculations do not take into consideration that 40% of the pine ESTs are from tissues or organs other than differentiating xylem.

Protein	Hits in <i>Arabidopsis</i>	Hits in pine	Fold difference	Probability ( <i>P</i> )
pirT03962 r40g3 stress-induced protein	4 ESTs (0.0036%)	47 ESTs (0.26%)	72-fold more in pine	<0.001
spQ40854 metallothionein-like protein	74 ESTs (0.067%)	204 ESTs (1.1%)	16-fold more in pine	<0.001
Gi 2765366, similar to pollen allergens	24 ESTs (0.022%)	53 ESTs (0.29%)	13-fold more in pine	<0.001
Gi 1563719 cyclophilin	119 ESTs (0.11%)	158 ESTs (0.88%)	8-fold more in pine	<0.001
pirT07139 cysteine proteinase inhibitor	34 ESTs (0.031%)	33 ESTs (0.18%)	5.8-fold more in pine	<0.001
pirT05667 (former) auxin-independent growth regulator	33 ESTs (0.030%)	30 ESTs (0.17%)	5.5-fold more in pine	<0.001
Gi 1419088 calreticulin	52 ESTs (0.046%)	28 ESTs (0.16%)	3.4-fold more in pine	<0.001
spP28014 translationally controlled tumor protein	169 ESTs (0.15%)	57 ESTs (0.31%)	2.1-fold more in pine	<0.001
pirT05950 lipid transfer protein	293 ESTs (0.26%)	73 ESTs (0.40%)	1.5-fold more in pine	<0.002
Gi 2995990 dormancy-associated protein	144 ESTs (0.13%)	26 ESTs (0.14%)	1.1-fold more in pine	<0.6

down-regulated by the within-slide test, 24 were confirmed by Dunnett's test, and 47 were not detected as down-regulated. Dunnett's test did not designate any element as differentially expressed which was not so designated by the within-slide test (Table 4).

In the juvenile wood versus mature wood comparison, 188 elements were designated as differentially expressed by the within-slide test, 113 as up-regulated in mature wood and 75 as up-regulated in juvenile wood. Of these 113 up-regulated elements, 94 were confirmed by Dunnett's test along with another 12 elements not designated as differentially expressed by the within-slide test. Of the 75 elements called up-regulated in juvenile wood by the within-slide test, 60 were confirmed by Dunnett's test, along with another 5 elements not designated as differentially expressed by the within-slide test. Table 3 shows known genes to which array elements showing differential expression by both methods are similar. The remaining elements showing differential expression either showed no similarity to any sequence in GenBank, or showed similarity only to uncharacterized genomic or EST sequences.

## Conclusions

Only one of the 20 largest pine contigs (of those without similarity to sequences in GenBank) showed significant similarity to one of the abundant poplar ESTs of unknown function. This result suggests that many of the genes highly expressed during secondary xylem formation in poplar are significantly different, either in sequence or in relative abundance, from the genes expressed during secondary xylem formation in pine. A more complete analysis of this interesting question awaits the completion of larger sets of both pine and poplar ESTs, or a separate project specifically focused on addressing the question of how many unique genes may be involved in secondary xylem formation in poplar and in pine. An equally interesting question, which must also await additional information, is whether any genes expressed during secondary xylem formation are truly specific to that process, or if secondary xylem formation uses largely the same genes as formation of primary xylem.

Comparisons of relative EST abundance in different libraries do not always agree with the microarray results about relative levels of gene expression in dif-

*Table 4.* Genes detected as differentially expressed by microarray analysis of juvenile vs. mature and compression wood vs. normal wood. EST sequences from elements of microarrays showing differences in expression in both 'within-slide' and Dunnett's test analyses were used to search GenBank (BLASTX) for similar sequences of known function, and the resulting GenBank accessions and descriptions are shown. If multiple elements were similar to a single type of sequence, then only a single entry is present in the table.

Tissue specificity	Similar to accession number	GenBank description
Juvenile wood abundant	AAC39360	low-MW heat shock protein
	P36182	heat shock protein 82
	T09248	HSP23.5
	T09253	HSP17
	A49539	XET
	P35694	BRU-1, XET-like protein
	JE0184	chitinase
	P21563	peptidyl-prolyl isomerase
	AAF17645	pectinesterase-like
	BAB01177	lipid transfer protein
	T14889	porin Mip2
	BAB09857	Phi-1-like protein
	AAC67358	acid phosphatase-like
	444790	nucleotide translocator
	CAA05979	adenine nucleotide translocator
	Q42679	<i>S</i> -adenosylmethionine decarboxylase
	Q05212	DNA damage/repair protein DRT102
	P35063	histone H2A
	AAC64128	actin
	AAG02215	Class III peroxidase
	BAA86060	senescence-related protein
	P54778	26S protease regulatory subunit 6B
	S31035	retroviral gag protein-like
S58500	auxin-induced protein IAA9	
Mature wood abundant	AAD50628	$\alpha$ -tubulin
	P41636	4-coumarate CoA ligase
	AAD02050	caffeoyl-CoA <i>O</i> -methyltransferase
	AAD23378	<i>trans</i> -cinnamate 4- hydroxylase
	P52777	phenylalanine ammonia-lyase
	AAG02215	Class III peroxidase
	S52995	arabinogalactan-like protein
	AAF75827	pine AGP5
	AAF75821	pine AGP6
	U09554	pine AGP-like 3H6
	CAB88264	callose synthase catalytic subunit-like protein
	BAB09063	cellulose synthase catalytic subunit-like protein
	AAF76468	contains a peptidase S8 domain
	BAB09397	cysteine protease-like
	AAA34123	hexameric polyubiquitin
	AAD39373	plasma membrane intrinsic protein 1
	P50300	<i>S</i> -adenosylmethionine synthase

Table 4 continued.

Tissue specificity	Similar to accession number	GenBank description
Mature wood abundant	AAC39360	low-MW heat shock protein
	P36182	heat shock protein 82
	CAA07232	putative Pi-starvation induced protein
	AAD21718	putative ribose phosphate pyrophosphokinase
	AAC33203	ATP-citrate-lyase-like
	BAA94511	ABC transporter-like
	S71769	low-MW heat shock protein
	T00801	homeobox protein-like
	T01643	DNA-J-protein-like
	U10432	lipid transfer protein
Z11487	pine globulin-2	
Compression wood abundant	U09554	AGP-like protein 3H6
	AAF75826	AGP4
	P52777	phenylalanine ammonia-lyase
	AAD21718	putative ribose phosphate pyrophosphokinase
	CAA94437	ABC transporter-like
	P93263	methionine synthase
Normal wood abundant	B71400	glycine hydroxymethyltransferase
	BAA94511	ABC transporter-like
	T16974	pectinesterase
	BAB09857	Phi-1-like protein
	AAC67358	acid phosphatase-like

ferent tissue types, and several reasons could account for the differences. Different trees were harvested at different times for the two types of studies, so true biological variation in the response is probably partially responsible. The differences could also be due to variation in the nature of the experimental methods. Microarray hybridizations and washes were carried out under relatively low-stringency conditions, so cross-hybridization between members of gene families could have occurred. The extent of cross-hybridization undoubtedly differs between elements on the array, both because the sizes of multi-gene families vary, and because some cDNA clones are full-length while others include only a small portion of the protein coding sequence and the 3'-untranslated region. Comparisons of EST abundance between libraries were carried out by using pine EST contigs as BLASTN queries of the pine EST collection, and counting the numbers of sequences from different libraries that show similarity at expect values less than  $10^{-5}$ . This could lead to inclusion of several members of a gene family in the count for a particular contig, depending on the presence of

conserved domains in some families of proteins. Some libraries used for the EST projects could contain up to six different alleles for each gene, and allelic variation within pine is sufficiently high to make it difficult to distinguish between alleles and different members of a gene family without genetic segregation data.

The two methods used for analysis of the microarray data each have strengths and weaknesses. The 'within-slide' approach is relatively easy to apply, and does not require positive control elements and construction of synthetic mRNAs. It does, however, have the disadvantage that the significance of any change in relative transcript levels for each element is dependent on all the other elements on the array. The same cDNA fragment spotted on two different arrays, and hybridized to the same two labeled cDNA pools, could yield data deemed significant from an array with other elements that show little change in expression, and yet the same ratio of differential expression might not be significant on a second array with other elements that show large changes in expression. This use of the other elements on the array as a test for significance limits

the ability of researchers to compare results between arrays that contain different but overlapping subsets of the total complement of genes in any genome. Dunnett's test requires some objective means of balancing the relative signal strength from the two channels, but does provide a measure of the experiment-wide false-positive rate.

The within-slide test as used in this study is not a stringent test for differential expression, requiring less than two standard deviations difference from the mean on a majority of arrays, but most elements on the array fail it. This test provides no means of determining the robustness of the conclusion other than permutation testing or bootstrapping, which are relatively computationally intensive methods of deriving an estimate of error rates from the data-set under study. Dunnett's test or ANOVA methods, in contrast, can be conducted at any desired level of experiment-wide error probability, and can also be adjusted to minimize the false-negative error rate or the false-positive error rate. The latter is an important consideration for investigators using microarrays to screen thousands of genes for those deserving more detailed study in a particular biological context – arguably a false-negative result could be more damaging to the outcome of such an experiment than a false-positive result, because of the failure to detect a gene of true significance to the process of interest.

#### *Genes expressed during wood formation*

Many of the genes expressed during pine secondary xylem differentiation are those genes expected to be present from previous work: cellulose synthase subunits, sucrose synthase, cell wall proteins, glucosidases and glucosyltransferases, and enzymes of lignin biosynthesis. Some pine cDNAs abundant in differentiating xylem cDNA libraries are similar to known genes which have never previously been linked with xylem differentiation or cell wall formation. One possibility is that those genes serve functions common to many cell types, but unrelated to cell wall formation, in differentiating xylem. Another possibility is that those genes have functions not yet known, and that they do play roles relevant to cell wall formation in differentiating xylem. The last class of genes are those identified by pine cDNAs which show no similarity to any known gene from any source. These may truly be pine-specific or secondary xylem-specific genes, and at least some of these genes are likely to play roles

in determining some of the aspects of secondary cell walls unique to pines.

Testing the roles in cell wall formation of candidate genes identified in pine will be difficult, due to the formidable challenges pine presents as an experimental system. Functional analysis of pine genes similar to *Arabidopsis* genes can readily be carried out in *Arabidopsis*, however, and the results confirmed in other model plants as appropriate. Pine genes not present in *Arabidopsis* clearly must be analyzed in pine, unless they can be found in maize or another model plant which provides the opportunity to carry out both forward and reverse genetics. These seem, however, to be a small fraction of all the interesting candidate genes identified in pine EST projects, so it may not be an overwhelming task to identify naturally occurring genetic variants in those genes in pine populations, then analyze the effects of those variants in controlled crosses. Natural populations of pines have relatively high frequencies of null alleles at isozyme loci, averaging about 0.3% across 22 isozyme loci (Allendorf *et al.*, 1982). This is equivalent to about one plant heterozygous for a deleterious allele per 300 plants in the population, or approximately the same frequency of mutant alleles observed in mutagenized populations of *Arabidopsis*. This approach has been fruitful in analyzing the effects of variation in cinnamyl alcohol dehydrogenase activity on lignin composition in loblolly pine (Ralph *et al.*, 1997).

The cell wall protein genes show a range of different patterns of expression. The transcripts of putative AGPs 2, 4 and 6 show similar probabilities of differential representation between vertical wood and compression wood libraries, vertical wood and shoot tip libraries, and compression wood and pollen cone libraries. Putative AGPs 1, 3, and 5 all have different patterns of differential expression among the libraries examined here, and EXP1 and GRP2 show little indication of differential expression in this analysis.

The most abundant cDNA clones identified by sequencing occur at a frequency of less than 1% of the total number of clones sequenced (50 of ca. 8500), and the frequency distribution drops rapidly from that point. There are 106 contigs that contain eight or more ESTs in a set of about 8500 normal wood ESTs analyzed, so we can estimate that about 100 transcripts occur at a frequency greater than 0.1% in differentiating normal wood. This set of 106 contigs encompasses a total of 1389 ESTs, or about 16% of the total number in the normal wood EST collection. This suggests that the hundred most abundant transcripts in differentiat-

ing pine secondary xylem account for only about 16% of the total number of mRNA molecules, and the other 84% constitute transcripts that are less abundant. This suggests that the mRNA pool in differentiating secondary xylem is quite complex, and that further EST sequencing will continue to yield novel genes, albeit with increasing numbers of cDNAs similar to those genes already identified.

Most of the contigs composed of the most abundant ESTs show sequence similarity to proteins of known or presumed function, but a significant minority (20 of 106) are either similar only to proteins of unknown function or show no similarity to public sequences. The latter group of genes may represent functions common to many or all plants, but abundant in differentiating pine xylem because of the specialized nature of the cells in that tissue. Functional genomics in differentiating pine xylem has the potential to contribute not to our understanding of wood formation, but to our understanding of cell wall formation and cellular differentiation in all plants.

### Acknowledgements

The research described in this report has been supported by grants from USDA (95-373000-1591), US Department of Energy (DE-FC07-97ID13550) and NSF (9975806). The long-term collaboration of Ernest Retzel and the Computational Biology Center at University of Minnesota is gratefully acknowledged, as are the contributions of past and present members of the Forest Biotechnology Group at North Carolina State University and the Institute of Forest Genetics of the Pacific Southwest Forest and Range Research Station of the USDA-Forest Service. The first-generation pine microarrays described here benefited greatly from the advice and assistance of Shauna Somerville and Per Villand of Carnegie Institute of Washington, Palo Alto, CA, Patrick Hurban of Paradigm Genetics, Research Triangle Park, NC, and Ernest Kawasaki of GSI Lumonics, Watertown, MA.

### References

- Allona, I., Quinn, M., Shoop, E., Swope, K., St. Cyr, S., Carlis, J., Riedl, J., Retzel, E., Campbell, M.M., Sederoff, R. and Whetten, R. 1998. Analysis of xylem formation in pine by cDNA sequencing. *Proc. Natl. Acad. Sci. USA* 95: 9693–9698.
- Altschul, S.F., Madden, 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. *Nucl. Acids Res.* 25: 3389–3402.
- Audic, S. and Claverie, J.M. 1997 The significance of digital gene expression profiles. *Genome Res.* 7: 986–995. Software available through <http://lgs-server.cnrs-mrs.fr>
- Bennett, M.D. and Leitch, I.J. 1995. Nuclear DNA amounts in angiosperms. *Ann. Bot.* 76: 113–176.
- Biermann, C.J. 1993. *Essentials of pulping and papermaking*. Academic Press, San Diego, CA.
- Chapple, C. and Carpita, N. 1998. Plant cell walls as targets for biotechnology. *Curr. Opin. Plant Biol.* 1: 179–185.
- Diatchenko, L., Lau, Y.F., Campbell, A.P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N., Sverdlov, E.D. and Siebert P.D. 1996. Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc. Natl. Acad. Sci. USA.* 93: 6025–6030.
- Dhugga, K.S., Tiwari, S.C. and Ray, P.M. 1997. A reversibly glycosylated polypeptide (RGPI) possibly involved in plant cell wall synthesis: purification, gene cloning, and trans-Golgi localization. *Proc. Natl. Acad. Sci. USA* 94: 7679–7684.
- Dunnnett, C.W. 1955. A multiple comparison procedure for comparing several treatments with a control. *J. Am. Statist. Ass.* 50: 1096–1121.
- 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, B. and Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8: 186–194.
- Friedman, N., Linial, M., Nachman, I. and Pe'er, D. 2000. Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, in press.
- Harada, H. and Côté, W.A. 1985. The structure of wood. In: T. Higuchi (Ed.) *Biosynthesis and Biodegradation of Wood Components*, Academic Press, Orlando, FL, pp. 1–42.
- Higuchi, T. 1997. *Biochemistry and Molecular Biology of Wood*. Springer-Verlag, Berlin.
- Kamm, A., Doudrick, R.L., Heslop-Harrison, J.S. and Schmidt, T. 1996. The genomic and physical organization of Ty1-copia-like sequences as a component of large genomes in *Pinus elliotii* var. *elliotii* and other gymnosperms. *Proc. Natl. Acad. Sci. USA* 93: 2708–2713.
- Kerr, M.K. and Churchill, G.A. 2000. Experimental design for gene expression microarrays. Submitted; manuscript available at <http://www.jax.org/research/churchill/pubs/index.html>.
- Kerr, M.K., Martin, M. and Churchill, G.A. 2000. Analysis of variance for gene expression microarray data. Submitted; manuscript available at <http://www.jax.org/research/churchill/pubs/index.html>.
- Kinlaw, C.S., Ho, T., Gerttula, S.M., Gladstone, E. and Harry, D.E. 1996. Gene discovery in loblolly pine through cDNA sequencing. In: *Somatic Cell Genetics and Molecular Genetics of Trees* (Forestry Sciences vol. 49), Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 175–182.
- Kinlaw, C. and Neale, D. 1997. Complex gene families in pine genomes. *Trends Plant Sci.* 2: 356–359.
- Kossack, D. 1989. The IFG copia-like element: characterization of a transposable element present in high copy number in *Pinus* and a history of the pines using IFG as a marker. Ph.D. dissertation, University of California at Davis, CA.
- Kossack, D.S. and Kinlaw, C.S. 1999 IFG, a gypsy-like retrotransposon in *Pinus* (Pinaceae), has an extensive history in pines. *Plant Mol. Biol.* 39: 417–426.

- Kriebel, H.B. 1985. DNA Sequence components of *Pinus strobus* nuclear genome. *Can. J. For. Res.* 15: 1–4.
- Lewin, M. and Goldstein, I.S. 1991. *Wood Structure and Composition*. Marcel Dekker, New York.
- Loopstra, C.A. and Sederoff, R.R. 1995. Xylem-specific gene expression in loblolly pine. *Plant Mol. Biol.* 27: 277–291.
- Loopstra, C.A., Puryear, J.D. and No, E.G. 2000. Purification and cloning of an arabinogalactan-protein from xylem of loblolly pine. *Planta* 210: 686–689.
- Megraw, R.A. 1985. *Wood Quality Factors in Loblolly Pine: the influence of tree age, position in tree, and cultural practice on wood specific gravity, fiber length, and fibril angle*. TAPPI Press, Atlanta, GA.
- Mellerowicz, E.J., Baucher, M., Sundberg, B. and Boerjan, W. 2001. Unravelling cell wall formation in the woody dicot stem. *Plant Mol. Biol.*, this issue.
- Meyer-Berthaud, B., Scheckler, S.E. and Wendt, J. 1999. *Archaeopteris* is the earliest known modern tree. *Nature* 398: 700–701.
- Murray, B.G. 1998. Nuclear DNA amounts in gymnosperms. *Ann. Bot.* 82: 3–15.
- Newton, M.A., Kendzioriski, C.M., Richmond, C.S., Blattner, F.R. and Tsui, K.W. 2000. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.*, in press.
- O'Malley, D., Whetten, R., Bao, W., Chen, C.-L. and Sederoff, R.R. 1993. The role of laccase in lignification. *Plant J.* 4: 751–757.
- O'Malley, D.M., Grattapaglia, D., Chaparro, J.X., Wilcox, P.L., Amerson, H.V., Liu, B.-H., Whetten, R., McKeand, S.E., Kuhlman, E.G., McCord, S., Crane, B. and Sederoff, R.R. 1996. Molecular markers, forest genetics and tree breeding. In: J.P. Gustafson and R.B. Flavell (Eds.) *Genomes of Plants and Animals: Proceedings of the 21st Stadler Symposium* (Columbia, MO), Plenum, New York, pp. 87–102.
- Ralph, J., MacKay, J.J., Hatfield, R.D., O'Malley, D.M., Whetten, R.W. and Sederoff, R.R. 1997. Abnormal lignin in a loblolly pine mutant. *Science* 277: 235–239.
- Reiter, W.D. 1998. The molecular analysis of cell wall components. *Trends Plant Sci.* 3: 27–32.
- Saltman, D., Thompson, L. and Bennett, K.M. 1998. *Pulp and Paper Primer*. TAPPI Press, Atlanta, GA.
- Schouten, J., de Kam, R.J., Fetter, K. and Hoge, J.H. 2000. Over-expression of *Arabidopsis thaliana* SKP1 homologues in yeast inactivates the Mig1 repressor by destabilising the F-box protein Grr1. *Mol. Gen. Genet.* 263: 309–319.
- Sederoff, R., Campbell, M., O'Malley, D. and Whetten, R. 1994. Genetic regulation of lignin biosynthesis and the potential modification of wood by genetic engineering in loblolly pine. *Rec. Adv. Phytochem.* 28: 313–355.
- Somerville, C. and Somerville, S. 1999. Plant functional genomics. *Science* 285: 380–383.
- Somssich, I.E., Wernert, P., Kiedrowski, S. and Hahlbrock, K. 1996. *Arabidopsis thaliana* defense-related protein ELI3 is an aromatic alcohol:NADP(+) oxidoreductase. *Proc. Natl. Acad. Sci. USA* 93: 14199–14203.
- Sterky, F., Regan, S., Karlsson, J., Hertzberg, M., Rohde, A., Holmberg, A., Amini, B., Bhalerao, R., Larsson, M., Villarroel, R., Van Montagu, M., Sandberg, G., Olsson, O., Teeri, T.T., Boerjan, W., Gustafsson, P., Uhlen, M., Sundberg, B. and Lundeberg, J. 1998. Gene discovery in the wood-forming tissues of poplar: analysis of 5692 expressed sequence tags. *Proc. Natl. Acad. Sci. USA* 95: 13330–13335.
- Timell, T.E. 1986 *Compression Wood in Gymnosperms* (3 vols.). Springer-Verlag, Berlin.
- Wakamiya, I., Newton, R.J., Johnston, J.S. and Price, H.J. 1993. Genome size and environmental factors in the genus *Pinus*. *Am. J. Bot.* 80: 1235–1241.
- Winzeler, E.A., Schena, M. and Davis, R.W. 1999. Fluorescence-based expression monitoring using microarrays. *Meth. Enzymol.* 306: 3–18.
- Wojtaszek, P. 2000. Genes and plant cell walls: a difficult relationship. *Biol. Rev. Camb. Phil. Soc.* 75: 437–475.
- Wolfinger, R.D., Gibson, G., Wolfinger, E.D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. and Paules, R.S. 2000. Assessing gene significance from cDNA microarray expression data via mixed models. Manuscript available from <http://statgen.ncsu.edu/ggibson/Publications/WGetc.pdf>
- Zhang, Y., Sederoff, R.R. and Allona, I. 2000. Differential expression of genes encoding cell wall proteins in vascular tissues from vertical and bent pine trees. *Tree Physiol* 20: 457–466.
- Zobel, B.J. and Sprague, J.R. 1998. *Juvenile Wood in Forest Trees*. Springer-Verlag, Berlin.
- Zobel, B.J. and van Buitenen, J. P. 1989. *Wood Variation: Its Causes and Control*. Springer-Verlag, Berlin.