Available online at www.sciencedirect.com
SCIENCE DIRECT®
PERGAMON
Phytochemistry 64 (2003) 1097–1112
PHYTOCHEMISTRY
www.elsevier.com/locate/phytochem

An in silico assessment of gene function and organization of the phenylpropanoid pathway metabolic networks in *Arabidopsis thaliana* and limitations thereof

Michael A. Costa, R. Eric Collins, Aldwin M. Anterola, Fiona C. Cochrane, Laurence B. Davin, Norman G. Lewis*

Institute of Biological Chemistry, Washington State University, Pullman, WA 99164-6340, USA

Received 7 January 2003; received in revised form 5 August 2003

Abstract

The *Arabidopsis* genome sequencing in 2000 gave to science the first blueprint of a vascular plant. Its successful completion also prompted the US National Science Foundation to launch the Arabidopsis 2010 initiative, the goal of which is to identify the function of each gene by 2010. In this study, an exhaustive analysis of The Institute for Genomic Research (TIGR) and The Arabidopsis Information Resource (TAIR) databases, together with all currently compiled EST sequence data, was carried out in order to determine to what extent the various metabolic networks from phenylalanine ammonia lyase (PAL) to the monolignols were organized and/or could be predicted. In these databases, there are some 65 genes which have been annotated as encoding putative enzymatic steps in monolignol biosynthesis, although many of them have only very low homology to monolignol pathway genes of known function in other plant systems. Our detailed analysis revealed that presently only 13 genes (two PALs, a cinnamate-4-hydroxylase, a *p*-coumarate-3-hydroxylase, a ferulate-5-hydroxylase, three 4-coumarate-CoA ligases, a cinnamic acid *O*-methyl transferase, two cinnamoyl-CoA reductases) and two cinnamyl alcohol dehydrogenases can be classified as having a bona fide (definitive) function; the remaining 52 genes currently have undetermined physiological roles. The EST database entries for this particular set of genes also provided little new insight into how the monolignol pathway was organized in the different tissues and organs, this being perhaps a consequence of both limitations in how tissue samples were collected and in the incomplete nature of the EST collections. This analysis thus underscores the fact that even with genomic sequencing, presumed to provide the entire suite of putative genes in the monolignol-forming pathway, a very large effort needs to be conducted to establish actual catalytic roles (including enzyme versatility), as well as the physiological function(s) for each member of the (multi)gene families present and the metabolic networks that are operative. Additionally, one key to identifying physiological functions for many of these (and other) unknown genes, and their corresponding metabolic networks, awaits the development of technologies to comprehensively study molecular processes at the single cell level in particular tissues and organs, in order to establish the actual metabolic context.

Keywords: *Arabidopsis thaliana*; Cruciferae; Phenylpropanoid-acetate pathway; Lignin/lignan biosynthesis; Suberization; Sporopollenin; Functional analysis; In silico; Expressed sequence tags (ESTs); Metabolic networks; Digital northern

1. Introduction

The National Science Foundation (NSF) Arabidopsis 2010 Initiative in the USA has as its mission the identification of the physiological functions of all of the *Arabidopsis thaliana* genes by 2010, including their cellular, organismal and evolutionary relationships. This approach will thus enable us to further understand, in a whole plant context, many of the most fundamental aspects of plant metabolism and the gene functions associated with same. The overall strategy being undertaken at present essentially uses a systems approach, this being a marked departure from the more conventional practice of, for example, defining a particular scientific goal based on an observed biological phenomenon. *Arabidopsis* was chosen since this angiosperm, a representative of the mustard family, has a relatively small but fully sequenced 120-megabase genome

* Corresponding author. Tel.: +1-509-335-8382; fax: +1-509-335-8206.
E-mail address: lewisn@wsu.edu (N.G. Lewis).

doi:10.1016/S0031-9422(03)00517-X

containing an estimated 25,000+ genes; it can therefore serve as a convenient model for addressing fundamental questions of biological structure and function common to all eukaryotes (Meinke et al., 1998; The Arabidopsis Genome Initiative, 2000).

As part of the NSF Arabidopsis 2010 Initiative, this laboratory has recently undertaken the task of studying some 248 genes found in the *A. thaliana* databases that display sequence homology to genes primarily involved in selected and/or suspected aspects of phenylpropanoid/phenylpropanoid-acetate pathway metabolism and/or cell wall formation, with a particular emphasis being placed on the lignin/lignan biosynthetic pathways. The overall goal is to isolate and characterize all of these homologues from *A. thaliana* and to determine their functional genomic roles by studying not only the patterns of gene expression and the effects of modulating same, but also in comprehensively characterizing the proteins and enzymes at the chemical, biochemical and structural levels.

The presumed phenylpropanoid pathway networks of interest can be contemplated as complex biological regulatory systems that have evolved in vascular plants during their successful transition to land, and which are ultimately essential for their growth, development and survival. However, how these metabolic networks are organized and differentially controlled is not yet understood. Nevertheless, it is these coordinated phenylpropanoid pathway networks from phenylalanine ammonia lyase (PAL) onwards [and preceding transcriptionally coordinated metabolic processes (Anterola and Lewis, 2002)] that eventually differentially afford the phenylpropanoid components of the lignins, lignans, hydroxycinnamic acids, flavonoids, suberins, sporopollenins, cutins and other related constituents in various cell types, tissues and organs (Lewis and Davin, 1994, 1999; Lewis et al., 1999). Furthermore, when considered from both evolutionary and functional perspectives, such networks have afforded vascular plants with competitive advantages for successful land plant adaptation (Lewis and Davin, 1994, 1999; Lewis et al., 1999), which include: mechanisms for obtaining and transporting water and nutrients (Lewis and Yamamoto, 1990); maintenance of a high water potential which facilitates active metabolism in desiccating environments (Bernards et al., 1995; Bernards and Lewis, 1998); minimizing the effects of temperature, humidity and (UV) light variations; withstanding and modulating forces of compression which act upon plant structures during growth and development (Lewis et al., 1999), and formation of specialized structures which permit lengthy maintenance of pollen grain viability (Wiermann and Gubatz, 1992).

Interestingly, the apparent partial overlap of these branchpoint phenylpropanoid pathways (e.g. to flavonoids, lignans, sporopollenins, etc.) has been viewed by some researchers as evidence for both metabolic redundancy and metabolic channeling (Hrazdina and Jensen, 1992). In contrast, our own view is that of a series of distinct, sophisticated, biochemical networks which are present in different tissues and cell types, and which enable the various metabolic branch pathways to be differentially controlled during each and every aspect of plant growth and development (Anterola and Lewis, 2002). However, no single plant species has yet been systematically examined to determine either how such proposed networks in phenylpropanoid pathway associated metabolism are organized and controlled during each stage of plant growth and development. In a somewhat related manner, little is even understood about the precise nature of the oxidative enzyme networks responsible for generating the free radical intermediates involved (at least in part) in the assembly of lignins, lignans, suberins and other phenol radical-radical linked products. Hence, the sequencing of the *Arabidopsis* genome now provides the opportunity for the systematic dissection of the proposed complex metabolic networks in order to identify how they are actually differentially organized during growth and development. Moreover, given that apparently there are various multigene families for most of the enzymatic steps, such proposed networks would be expected to be differentially controlled. This could occur as distinct gene functions within particular cell types, and/or it could be a result of differential expression within the same cell type with distinct metabolic pathways being activated concurrently and/or differentially. It is one of our goals to identify how such overall organization is achieved.

In this regard, it was anticipated that an in silico analysis could provide some useful initial insights in order to bridge the gap between information stored within the *Arabidopsis* genome, and that already known about [quantifiable] spatial and temporal gene expression patterns (Bouchez and Höfte, 1998). This is because the relative abundance of expressed sequence tags (ESTs) from available databases, which are derived from libraries from different organs and plants, might be anticipated to provide important clues as to possible physiological and/or developmental functions and hence of the tissue-specific organization of the proposed networks involved. Such EST analyses have been applied in studies as diverse as toxicology (Fielden et al., 2002), inner-ear function (Klockars et al., 2003), and plant metabolic pathways (Allona et al., 1998; Sterky et al., 1998; White et al., 2000; Hertzberg et al., 2001).

For example, to learn more about possible gene expression patterns involved in wood formation, Allona et al. (1998) performed a preliminary analysis of xylem formation in loblolly pine (*Pinus taeda*) albeit based only on 1097 ESTs. Using this approach, genes putatively encoding regulatory proteins were identified, as were others presumed to be involved in cell wall formation, in addition to lignin and carbohydrate biosynthetic

enzymes. As expected, other than their grouping into broad enzyme classes (e.g. cytochrome P450s), the actual physiological roles of unknown partially sequenced genes could not be identified at the time of the investigation. For example, it was not until 2002 with functional genomic studies, that a gene of unknown function and partially annotated as a cytochrome P450, was established to be a *p*-coumarate 3-hydroxylase in loblolly pine (*P. taeda*) (Anterola et al., 2002), based partly on previous findings by Schoch et al. (2001). In an even more comprehensive manner, Hertzberg et al. (2001) identified an unique tissue-specific transcript profile for a well-defined developmental gradient occurring during xylogenesis in the secondary xylem of poplar trees. These data suggested, as would be predicted, that genes encoding lignin biosynthetic enzymes, transcription factors and other potential xylogenesis regulators were under strict developmental stage-specific transcriptional regulation; however, again no new physiological functions for any of the genes were identified. In other approaches, Klockars et al. (2003) used an in silico analysis of mouse inner-ear transcripts to identify those putatively associated with specific roles in auditory or vestibular functions, whereas White et al. (2000) employed an "electronic or digital northern" technique to estimate gene expression involved in the conversion of photosynthate into oil in developing seeds of *Arabidopsis*; in the latter case, the ESTs revealed patterns of gene expression associated with specific plant tissues and/or growth conditions.

In spite of the obvious limitations in identifying either precise physiological functions and temporal/spatial relationships of both genes and gene products on a cell-by-cell basis, this "electronic or digital northern" might still provide possible clues into the organization of metabolic routes of interest (White et al., 2000). Accordingly, we thus considered it instructive to initially employ a computerized "database mining" approach to generate three-dimensional graphical representations of reprocessed EST database collections of specific tentative consensus (TC) cDNAs of interest, i.e. as preliminary evaluations of gene expression patterns via "digital northern" analyses. Initial source information was retrieved from The Institute for Genomic Research (TIGR) and The Arabidopsis Information Resource (TAIR) databases in order to facilitate further understanding of the presumed metabolic networks involved in phenylpropanoid metabolism including their regulatory systems. As described below, however, the EST database entries only provided very limited insights into both the identification and organization of the networks involved post-phenylalanine (**1**) to the monolignols **2**–**4**. These data also underscored the need for both a full biochemical clarification of the roles of each protein/enzyme of interest, together with the determination of the actual metabolic context of each in vivo.

2. Results and discussion

The three possible and distinct networks integrally associated with the phenylpropanoid pathway, and of primary interest to us, include temporal and spatial organization in planta of the various: (1) radical–radical coupling/polymerization networks for phenylpropanoid coupling/polymerization leading to formation of lignins, lignans, suberins, sporopollenins, etc. and how such proteins and enzymes mechanistically function; (2) downstream metabolic networks, following phenylpropanoid coupling of monolignols **2**–**4**, as well as the nature and function of the proteins and enzymes involved; and (3) networks that exist in *Arabidopsis* and which are involved in the conversion of phenylalanine (**1**) through to the monolignols **2**–**4** and associated metabolites (Fig. 1). This particular contribution, however, addresses (a) only the latter network, (b) the information gained thus far and (c) the caveats involved in data interpretation.

In order to do this, both TIGR and TAIR databases were initially analyzed, with screening of the ESTs providing a summary of non-redundant tentative consensus (TC) DNA sequence homologues for proteins putatively associated with each step of the phenylpropanoid pathway, i.e. those related to PAL and others possibly en route to monolignols **2**–**4**. Moreover, in order to obtain statistically meaningful data, only data from libraries having over 5000 ESTs were used, as well as those from libraries obtained from well-defined organs. The results of this analysis are summarized in Fig. 2. Subsequent data analysis also permitted deduction of the presumed signal peptide cutoff sites and the subcellular target locations for each gene of interest using the SignalP (Nielsen et al., 1997), TargetP (Emanuelsson et al., 2000) and iPSORT (Bannai et al., 2002) programs, as well as identifying the specific tissues in which these proteins have a high probability of being found.

It is important to note, however, that although the "entry point" into the pathway is through phenylalanine ammonia lyase (PAL) and/or tyrosine ammonia lyase (TAL) (see Lewis et al., 1999), the reader should be cognizant that there also needs to be concurrent activation of the chorismate/shikimate pathway (and other preceding pathways) in order to supply the rate-limiting amounts of Phe (**1**)/Tyr (**5**) precursors needed for phenylpropanoid metabolism (Anterola et al., 1999, 2002). For simplicity, Fig. 1 summarizes only the currently understood pathway steps from PAL to cinnamyl alcohol dehydrogenases (CAD). Interestingly, even with this relatively small subset of enzymatic steps (and genes), there are within the *Arabidopsis* genome database *putatively* four PAL homologues, one cinnamate 4-hydroxylase (C4H), one *p*-coumarate 3-hydroxylase (C3H), two ferulate 5-hydroxylases (F5H) (and each of their associated NADPH dependent reductases), 14

coumarate CoA ligases (4CL), 4 cinnamoyl CoA *O*-methyltransferases (CCOMT), 17 caffeic acid *O*-methyltransferases (COMT), 11 cinnamyl CoA reductases (CCR) and 9 CAD (like) proteins in *Arabidopsis*. Of these enzymes, both C4H and C3H have established rate-limiting roles in carbon allocation to the pathways(s) and in determining monolignol composition (Anterola et al., 1999, 2002) as to be expected from metabolic control analysis (Fell, 1997). However, curiously other researchers have indicated that each and every step of the phenylpropanoid pathway has a rate-limiting role [e.g., CAD (Baucher et al., 1996)], this assertion now being fully disproven (Anterola and Lewis, 2002).

Each of the enzymatic steps in monolignol formation is described below.

2.1. Phenylalanine ammonia lyase (PAL)

PAL catalyzes the non-oxidative deamination of L-Phe (**1**) to afford *trans*-cinnamic acid (**6**) and ammonium ion (Koukol and Conn, 1961), with the cinnamate (**6**) so formed being, depending upon the tissue, cell type and organ, *differentially* metabolized into lignins, lignans, hydroxycinnamic acids, suberins, flavonoids, etc. (Lewis and Davin, 1999; Lewis et al., 1999; Croteau et al., 2000). Interestingly, the enzyme has no co-factor requirement, and the ammonium ion released is used to regenerate more L-Phe (**1**) as needed in a tightly regulated nitrogen recycling process (van Heerden et al., 1996). Furthermore, transcription of PAL-encoding genes is controlled during development and in response to environmental cues, [e.g. in French bean (*Phaseolus*

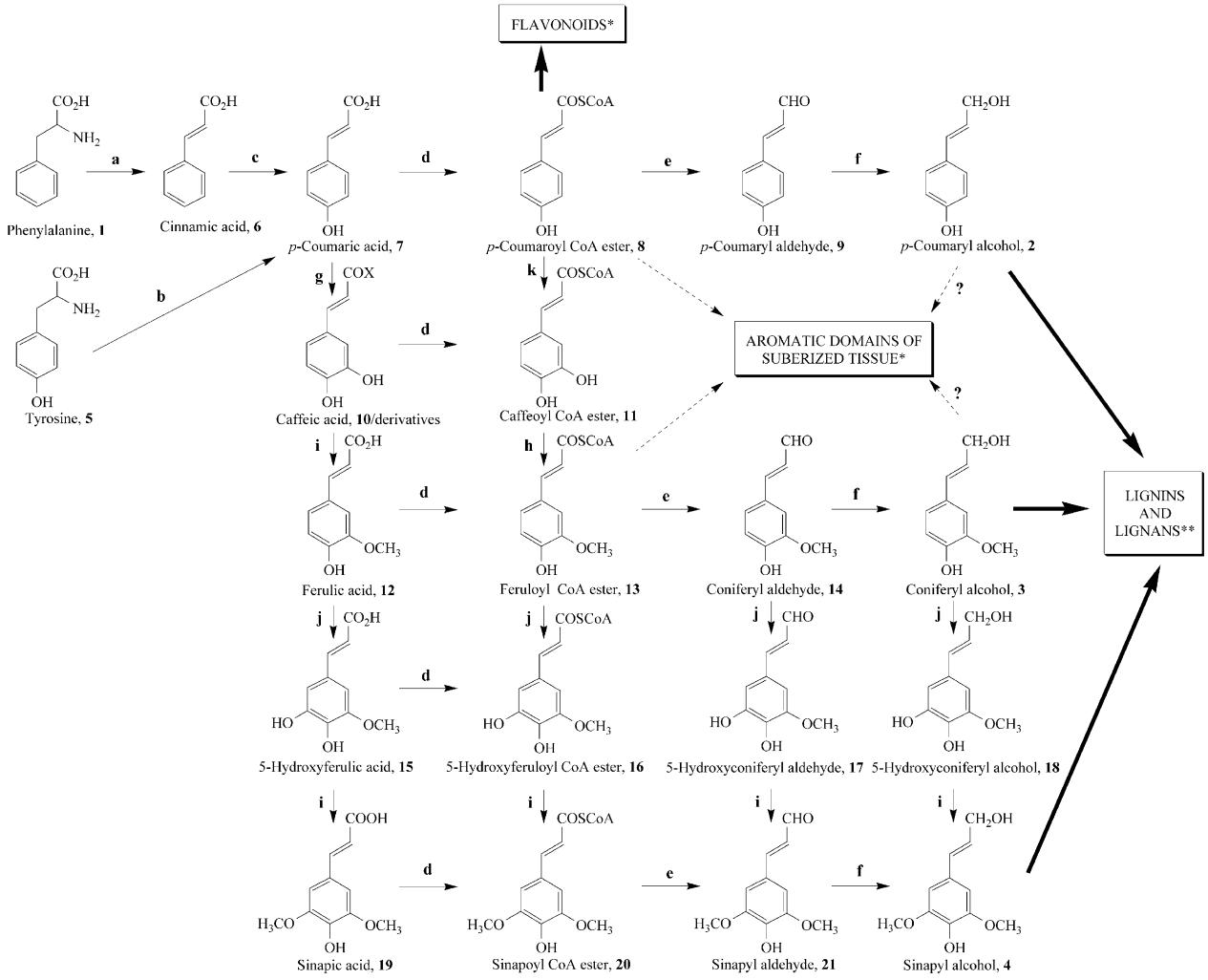

Fig. 1. Major segments of phenylpropanoid metabolism in vascular plants as currently understood: (a) phenylalanine ammonia-lyase (PAL), (b) tyrosine ammonia-lyase (TAL), (c) 4-cinnamate hydroxylase (C4H) and its associated reductase, (d) 4-coumarate:CoA ligase (4CL), (e) cinnamoyl-CoA reductase (CCR), (f) cinnamyl alcohol dehydrogenase (CAD), (g) *p*-coumarate 3-hydroxylase (C3H), (h) caffeoyl-CoA *O*-methyltransferase (CCOMT), (i) caffeate *O*-methyltransferase (COMT), (j) ferulate 5-hydroxylase (F5H) and associated reductase and (k) *p*-coumaroyl CoA 3-hydroxylase and associated reductase (note: X = OH, quinate, shikimate, or an unknown functional group).

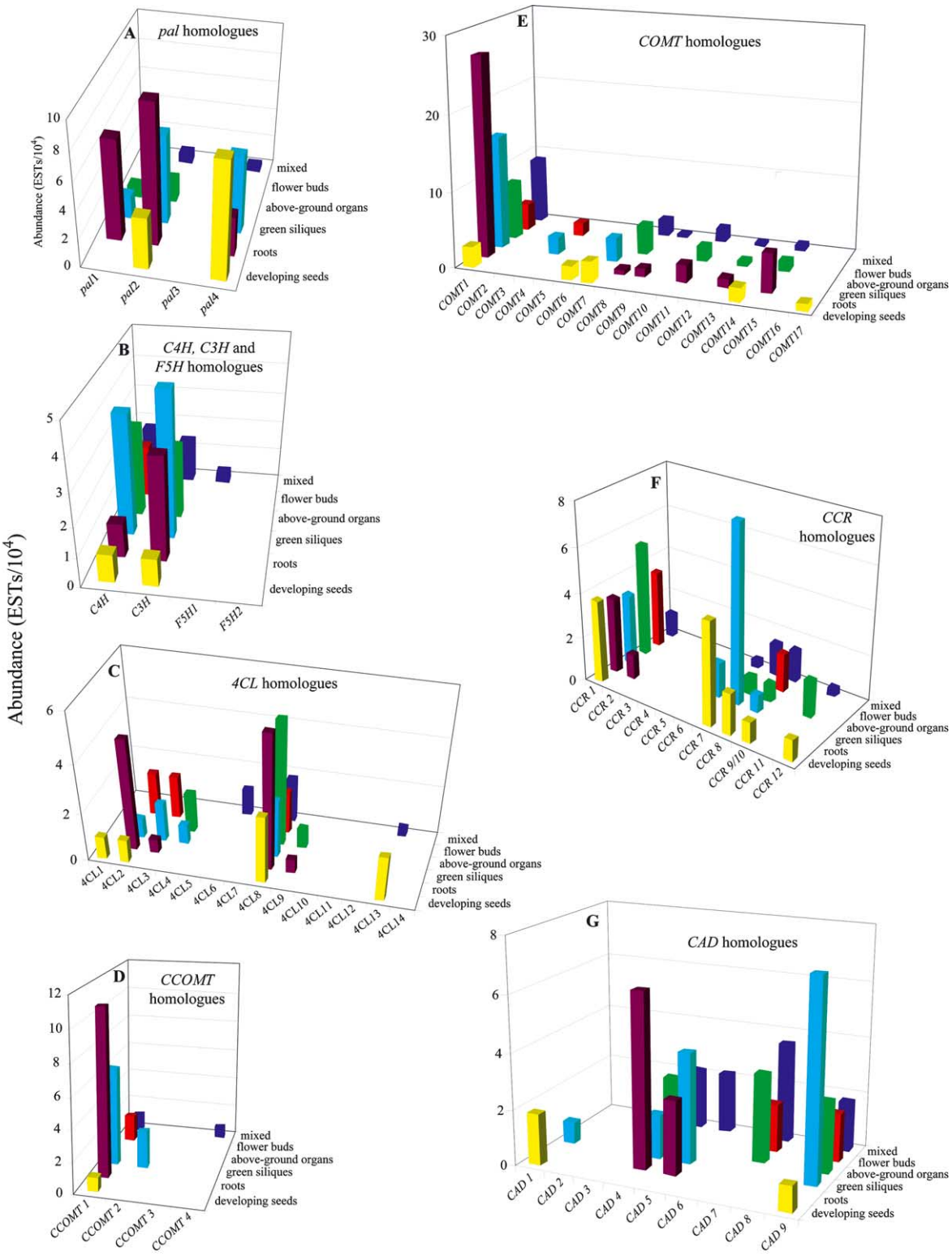

Fig. 2. Reprocessed EST data for *Arabidopsis* in terms of relative abundances of tissue expression patterns: (A) *PAL* homologues (*PAL1–PAL4*). *PAL1* (At2g37040), *PAL2* (At3g53260); *PAL3* (At5g04230) and *PAL4* (At3g10340) (code: for *PAL1*, At: *Arabidopsis thaliana*; 2 = chromosome number; 37040, position on chromosome); (B) *C4H*, *C3H* and *F5H* homologues (*F5H1* and *F5H2*); *C4H* (At2g30490), *C3H* (At2g40890); *F5H1* (At4g36220) and *F5H2* (At5g04330); (C) *4CL* homologues (*4CL1–4CL14*, see Table 1); (D) *CCOMT* homologues (*CCOMT1–CCOMT4*); *CCOMT1* (At4g34050), *CCOMT2* (At4g26220), *CCOMT3* (At1g67990) and *CCOMT4* (At1g67980); (E) *COMT* homologues (*COMT1–COMT17*, see Table 2); (F) *CCR* homologues (*CCR1–CCR12*, see Table 3); and (G) *CAD* homologues (*CAD1–CAD9*, see Table 4).

vulgaris) (Liang et al., 1989)]. An AC-rich Pal-box binding factor (Ntlim1) has also been identified in tobacco and it is highly similar to the LIM protein family which affects levels of phenylpropanoid pathway gene transcripts (e.g. PAL, 4CL and CAD) and lignin contents (Kawaoka et al., 2000; Kawaoka and Ebinuma, 2001). Overexpression of MYB proteins (AmMYB308 and AmMYB330, from *Antirrhinum*) in tobacco has additionally been shown to downregulate lignin biosynthesis (Tamagnone et al., 1998). However, how regulation of different members of PAL multigene families is effectuated, and which may result in different metabolic fates as regards to the products formed is not understood. Yet, by contrast, previous studies in *Arabidopsis* have essentially focused only on *PAL1* expression profiles (Ohl et al., 1990; Kubasek et al., 1992; Leyva et al., 1995; Giacomin and Szalay, 1996), and have thus overlooked almost entirely the various members of this multigene family in *Arabidopsis*, as well as the actual (demonstrable) physiological roles of each.

In this context, EST and genomic sequence analyses of each of the four cytosolic *Arabidopsis* PAL homologues yielded interesting, albeit preliminary, observations (Fig. 2A). For simplicity in presentation, each homologue is designated 1, 2, 3, etc. with the figure legend indicating the actual chromosome locus. Relative to *PAL1*, the three other *PAL* homologues display 92.7, 79.9 and 85.3% similarity, and 90.3, 73.9 and 80.5% identity, respectively. All are cytosolic and lack any targeting peptide sequence. Of these, only three of the four *PAL* homologues were detected in the EST database (Fig. 2A), with *PAL3* escaping detection perhaps due to low copy number, transient expression and/or being inducible only under certain conditions. Additionally, *PAL1*, *2* and *4* are differentially expressed in the various tissues examined, although with no indication of either distinct spatial and/or temporal expression patterns in, for example, particular cell types. *PAL2* and *PAL4* also apparently differ from *PAL1* in that both are mainly expressed in above-ground (presumably lignified) organs, green siliques and roots, whereas *PAL1* seems to be mainly present in just the roots and green siliques. Interestingly, the relative abundances of *PAL1* and *PAL2* ESTs are highest overall in root tissues, this perhaps being an indication that the ESTs for the above-ground organs were from *Arabidopsis* tissues which were not (very) actively lignifying at the time(s) of collection. Furthermore, the presence of *PAL2* in essentially every tissue known to undergo lignification (including developing seeds) might suggest it to be the *most likely* candidate for an exclusive and/or dominant role in targeting carbon to the lignin pathway. *PAL4* was also highly abundant in developing seed tissues whereas *PAL1* was not, suggesting a different metabolic role for *PAL4* within this tissue. Finally, no ESTs for any of the PAL isoforms were detected in flower buds.

Provisionally, these data can be placed in context with results obtained during an earlier study of *PAL1* expression (using an ~1.8 kb *Arabidopsis PAL1* promoter fused to the *GUS* reporter gene) at different developmental stages, with visualization of expression patterns using X-gluc (5-bromo-4-chloro-3-indoyl-β-glucuronide) (Ohl et al., 1990). In the data reported (together with our own re-analysis), staining was most intense in primary root vascular tissue (but not in either root tips or hairs, and perhaps not in the epidermal regions), and to a lesser extent (decreased overall intensity) in leaf vascular tissue for plants from 2 to 25 days old. In the more mature plants (~25 days old and older), staining was most evident (strongest) in root and leaf vascular tissues, but no data were reported showing either its expression or lack thereof in lignifying stem tissue. Furthermore, although not detected in developing flower buds, *PAL1* expression was noted in mature floral organs, including the sepals, stamens, and the tip and base of the carpel, but not in the petals. Expression was also detected in the developing siliques. In general, these data are in agreement with that noted for *PAL1* EST database tissue-specific expression, and it again *appears* to provisionally suggest that *PAL1* may have no direct involvement in lignification. However, this will only be unambiguously determined when the expression patterns of all four *PAL* homologues are known, together with gaining the knowledge of the actual metabolic context for each.

Thus, when considered together, the preliminary EST data for each *PAL* homologue helps provisionally explain some of the difficulties and uncertainties previously encountered in assigning/identifying specific physiological functions to each member of the *PAL* multigene family, regardless of plant species under investigation. It also emphasizes the absolute necessity to systematically establish the individual cellular and tissue patterns of expression levels at each stage of plant growth and development, as well as identifying their individual physiological roles (i.e. metabolic networks involved) and even their individual substrate specificities, i.e. for Phe (**1**) and Tyr (**5**). This also includes determination of their patterns of induction of gene expression (as well as metabolites actually formed) in response to wounding, UV-light exposure, pathogen attack and/or other stresses.

Additionally, the EST data calls into question the overall utility of employing heterologous promoters (e.g. CaMV 35S) as a means to ostensibly dissect physiological roles of specific genes, since such strategies can presumably at best only give gross (averaging) of the various PAL-mediated pathways, rather than those due to specific isoform effects. Put in another way, the temporal and spatial patterns of expression of native *Arabidopsis PAL* promoters, coupled with actual substrate specificities and metabolic pathway (end product)

identification, are needed to unambiguously clarify both the physiological roles of each isoform, and the corresponding metabolic networks that each is associated with.

2.2. Cinnamate-4-hydroxylase (C4H) and p-coumarate-3-hydroxylase (C3H)

Both enzymes, in conjunction with their NADPH-dependent reductases, catalyze sequential regiospecific hydroxylations at C-4 of cinnamate (**6**) and C-3 of *p*-coumarate (**7**) (or an analog thereof), respectively (Benveniste et al., 1986; Schoch et al., 2001). Both enzymes in *Arabidopsis* are encoded by single genes, and both have signal peptides (iPSORT program) that indicate targeting to the secretory system; in addition, the associated reductases have hydrophobic domains at the N-terminus for anchoring into the plasma membrane (Urban et al., 1997; Kida et al., 1998). C4H and C3H are additionally both established as having rate-limiting roles (Anterola et al., 1999, 2002) in carbon allocation/determination of monolignol composition in *P. taeda* cell cultures, an observation in harmony with the re-analysis of data obtained from various transgenic plant lines (Anterola and Lewis, 2002). Furthermore, there is evidence linking C4H regulation with a MYB transcription factor (AtMYB4) (Jin et al., 2000).

Provisionally, C4H can be considered as intimately involved in all aspects of phenylpropanoid metabolism in *A. thaliana*, and should therefore be colocalized in the same cell types as each of the corresponding four *PAL* homologues; this, however, assumes that none of the four *PAL* homologues actually encode a TAL isoform. If the latter occurred, this would potentially eliminate the need for C4H in, for example, that particular metabolic network.

Nevertheless, with only one gene encoding C4H and C3H, respectively, this provides an excellent opportunity to compare and contrast differential patterns of expression of the four *PAL* homologues with that of *C4H* and *C3H*, and to place these in a sound metabolic context. In this regard, it is anticipated that C3H would only be found in cells undergoing caffeate (**10**) (or caffeate derivative) metabolism, and hence this offered the opportunity to systematically distinguish its expression pattern from that of PAL–C4H metabolism directed to, for example, flavonoid and *p*-coumaroyl (**2**) formation, i.e. on either a spatial and/or temporal basis.

Analysis of the EST database entries (Fig. 2B) provided only limited insight into tissue-specific expression, with both *C4H* and *C3H* transcripts being present in all tissues examined, i.e. developing seeds, roots, green siliques, above-ground organs, flower buds and in mixed tissues. As such, these data gave no basis for distinction of any metabolic and/or tissue-specific functions. (In this regard, a further complication is the possible intermediary role of a *p*-coumarate (**7**) ester (such as quinate/shikimate) for the 3-hydroxylation step.) Clearly, this and the relative subcellular locations of both C3H and C4H need to be determined in order to fully understand the organizational basis in vivo for both hydroxylation steps.

Arabidopsis C4H gene expression has, however, been reported previously (Bell-Lelong et al., 1997; Nair et al., 2002), using ~2.9 kb C4H promoters fused to the GUS-reporter gene, albeit with no definitive metabolic context being described. Nevertheless, both studies demonstrated that the *C4H* gene was expressed throughout the stems (xylem and sclerified parenchyma), leaves, roots (but not root tips/root hairs), petioles, and green siliques (i.e. in agreement with a dominant role in lignin biosynthesis), as well as being present in floral organs and developing seeds (Bell-Lelong et al., 1997). However, as for *PAL1*, the actual cell types involved were only partially identified and the precise physiological (metabolic pathway) roles in the various tissues and organs—including those in lignifying tissues—were not determined. Additionally, *C4H* was found to be strongly expressed throughout the entire primary root (stele and cortex cells, with the latter being particularly evident adjacent to the root tip), but no precise metabolic context was established.

These observations can presumably also be placed in an additional metabolic context, since in a study by Saslowsky and Winkel-Shirley (2001) both flavonoid pathway enzymes, chalcone synthase (CHS) and chalcone isomerase (CHI), were detected in root cortex and epidermal cells (where the flavonoids accumulate), i.e. indicating that at least in *those* cells, C4H is primarily (if not fully) associated with the flavonoid pathway. It will thus be instructive in future to systematically identify which *PAL* gene (or genes) is (are) associated with these specific cells and the other cell types during *Arabidopsis* growth and development. Furthermore, it will be of considerable interest to understand how C4H (in the different cells, tissues and organs where it is expressed) is differentially regulated vis-à-vis the distinct metabolic branches.

In a somewhat related manner, a *C3H*:promoter:*GUS* fusion study indicated that in *Arabidopsis C3H* expression was often colocalized with *C4H*, although some significant differences in expression were noted (Nair et al., 2002). For example, in primary root tissue, *C3H* expression occurred in the stele, but not within either cortex and/or epidermal cells, where *C4H*, *CHS* and *CHI* were detected; this is in agreement with the latter being involved in flavonoid biosynthesis in such cell types.

2.3. 4-Coumarate CoA ligase (4CL)

Generically, this ATP-requiring class of enzymes converts *p*-coumaric acid (**7**) and potentially other substituted

cinnamic acids, e.g. **10**, **12**, **15** and **19** into the corresponding CoA esters **8**, **11**, **13**, **16** and **20**, respectively. Initially, in *Arabidopsis* (Lee et al., 1995), 4CL was reported to exist as a single gene (*4CL1*), this subsequently being extended to a three membered gene family (*4CL1–4CL3*) by 1999 (Ehlting et al., 1999). However, following completion of the *Arabidopsis* genome, some 14 genes were annotated as *putative* 4CLs, with identities of the corresponding proteins ranging widely from ~83 to ~40% relative to the originally isolated 4CL1 (see Table 1). It should be noted, though, that the precise biochemical and physiological functions of each are largely unknown, although all of them contain an AMP binding domain.

The available EST database entries (Fig. 2C) provide some preliminary insight into possible patterns of *putative 4CL* expression in developing *Arabidopsis*. In the various tissues examined, only eight of the *putative 4CL*s were detected, namely *4CL1–4CL4*, *4CL6*, *4CL8*, *4CL9* and *4CL13*, with *4CL8* being the most abundant, i.e. *4CL1*, *2*, *8* and *13* were noted in developing seeds, *4CL2*, *3*, *8* and *9* were present in roots, *4CL2*, *3*, *4* and *8* were in green siliques, whereas *4CL4*, 8 and *9* were in above-ground organs. Additionally, *4CL2*, *3* and *8* were found in flower buds and *4CL6*, *8* and *13* were present in mixed tissues. As for *PAL3*, the remaining six isoforms were not detected, this perhaps being due to low copy number, transient expression and/or only being inducible under certain conditions. As to be expected, these data provisionally suggest that different 4CL isoforms are specifically targeted to distinct metabolic pathways, e.g. to lignins, lignans, various hydroxycinnamate derivatives, as well for formation of, for example, the phenolic components of suberin, sporopollenin and cutin.

However, as summarized in Table 1, a closer analysis of the 14 putative *4CL* genes reveals a PTS1 motif (Jay Shockey, personal communication and Hayashi et al., 1997; Kragler et al., 1998) at the C-terminus for *4CL4*, *6*, *7*, *9* and *12–14*, this in turn representing a motif generally interpreted for peroxisome targeting. This may imply that none of the presumed peroxisome targeted enzymes are actually 4CLs, since the latter are generally believed to be cytosolic and lack a signal peptide and/or targeting sequence. Obviously, there is an urgent need to systematically establish the precise biological functions of each of the 14 *putative* 4-coumarate CoA ligase isoforms.

Table 1
Putative *Arabidopsis* 4-coumarate CoA ligases as annotated by TIGR

		Similarity (%)	Identity (%)	Peroxisome targeting signal, PTS1
4CL1	At1g51680	100.0	100.0	No
4CL2	At3g21240	87.2	83.4	No
4CL3	At1g65060	71.3	61.3	No
4CL4	At5g63380	50.6	37.4	Yes
4CL5	At3g21230	74.5	66.4	No
4CL6	At4g05160	55.3	43.9	Yes
4CL7	At4g19010	49.6	39.8	Yes
4CL8	At3g48990	42.8	31.1	No
4CL9	At1g20510	50.6	40.9	Yes
4CL10	At1g62940	52.8	39.6	No
4CL11	At5g38120	47.2	36.2	No
4CL12	At1g20500	51.1	39.3	Yes
4CL13	At1g20490	49.7	38.1	Yes
4CL14	At1g20480	49.8	39.1	Yes

As there are currently no reports on detailed (putative) *4CL* expression patterns (e.g. using the *GUS*-reporter fusion strategy), we have no knowledge of the precise cellular regions of expression in the various tissues and organs of *Arabidopsis* at different stages of growth and development. There are, however, some preliminary insights into possible biochemical functions for *4CL1-3*, as well as of their potential tissue-specific patterns of expression using northern blot analyses.

In this regard, phylogenetic comparison of *Arabidopsis* 4CL1-3 with other known 4CL amino acid sequences has identified two major clusters (Class I and Class II) that have evolved within the angiosperms (Ehlting et al., 1999): 4CL1 and 2 are categorized in Class I, together with, for instance, a presumed lignin-specific aspen 4CL1 (Hu et al., 1998), and parsley 4CL1 and 2, *Lithospermum* 4CL1 (Ehlting et al., 1999) and *Populus trichocarpa* 4CL3 (Cukovic et al., 2001). On the other hand, 4CL3 is a Class II 4CL, together with aspen 4CL2 (Hu et al., 1998), rice 4CL1 and 2, *Lithospermum* 4CL2 (Ehlting et al., 1999) and *P. trichocarpa* 4CL4 (Cukovic et al., 2001).

In terms of their potential substrate specificities, crude extracts of recombinant *Arabidopsis* 4CL1 and 4CL3 apparently showed a preference for *p*-coumarate (**7**) (relative to the other hydroxycinnamate substrates evaluated), whereas 4CL2 was more specific for caffeate (**10**). None converted sinapate (**19**) into its corresponding CoA ester (**20**), and ferulate (**12**) was a relatively poor substrate compared to *p*-coumarate (**7**), with only 4CL1 and 4CL3 apparently catalyzing this conversion. Ehlting et al. (2001) have further investigated the substrate recognition domains of 4CL1 and 4CL2 using a domain swapping approach generating chimeric proteins. They have shown that there are two small contiguous domains, which clearly have profound effects on substrate utilization by both 4CLs.

Of these isoforms, however, 4CL1 is considered to possibly have a specific role in lignin biosynthesis in *Arabidopsis*, even though the EST data (Fig. 2C) only indicated expression in developing seed, and thus no obvious correlation with lignification. Possible involvement in lignification is based on the following two observations: (1) Northern blot analyses suggest it to be strongly expressed in seedling roots and bolting stems,

with more moderate levels of expression in the siliques and lower levels in mature flowers. However, *4CL1* expression was also induced upon fungal attack, wounding and UV illumination; therefore, a role in plant defense cannot be excluded as well; (2) down-regulation of 4CL1 (using either the CaMV 35S and/or a parsley *4CL* promoter) resulted in significant reductions in lignin levels, albeit with *unusual* changes in monomeric composition apparently being observed (Lee et al., 1997; Anterola and Lewis, 2002). As noted earlier with the *PAL* multigene family, however, it would have been much more instructive and definitive to utilize the specific native promoter of *4CL* rather than employing heterologous promoters.

As regards the other two partially characterized *4CL* genes, *4CL2* was most strongly expressed in seedling roots, but at lower levels in all other organs, and was also inducible by wounding, fungal treatment and UV irradiation. Moreover, since it shows a higher affinity for caffeic acid (**10**) [but not ferulic acid (**12**)], a role in lignin biosynthesis has also been contemplated. *4CL3*, by contrast, had highest transcript levels in floral organs, with lower levels in siliques, leaves, bolting stems and seedling roots. It is also induced by UV illumination, but not by either wounding and/or fungal treatment. Together, these data gave rise to the possibility that 4CL3 may be involved in flavonoid metabolism, i.e. to afford the flower pigments and UV-protecting flavonols/anthocyanins.

In summary, even with there being fourteen putative *4CL* genes, much needs to be done to delineate their individual biochemical functions, as well as placing their particular patterns of gene expression into a sound metabolic context. In this way, it should be possible to identify the true physiological roles of each gene in vivo, particularly given the quite broad substrate versatility that many 4-coumarate CoA ligases have with various hydroxycinnamic acid substrates in vitro.

2.4. Caffeoyl-CoA O-methyltransferase (CCOMT)

Since the discovery of a distinct class of *O*-methyltransferases catalyzing the conversion of caffeoyl CoA (**11**) into feruloyl CoA (**13**) (Pakusch et al., 1989; Pakusch and Matern, 1991; Schmitt et al., 1991), it is now well established that vascular plants contain at least two classes of *O*-methyltransferases involved in phenylpropanoid/monolignol pathway metabolism. Depending upon the class, these can either act on the corresponding CoA esters (CCOMTs), or are involved in the final methylation step leading to syringyl moieties (so called caffeic acid *O*-methyltransferases, COMTs). However, the rather broad enzyme versatility displayed by each class of the *O*-methyltransferases in vitro has led to much uncertainty as regards precise physiological functions (particularly for CCOMT). For example, alfalfa CCOMT is capable of converting caffeoyl (**11**) and 5-hydroxyferuloyl CoA (**16**) into feruloyl (**13**) and sinapoyl CoA (**20**), but does not efficiently use hydroxycinnamic acids **7**, **12** and **19** as substrates (Inoue et al., 1998). Thus, there is currently some uncertainty as to whether CCOMT is involved in guaiacyl, syringyl and/or guaiacyl–syringyl lignin biosynthesis (Anterola and Lewis, 2002).

Arabidopsis contains some four *CCOMT* homologues (*CCOMT1–CCOMT4*) which, when compared to the alfalfa *CCOMT* (Inoue et al., 1998), display between ~94–65% identity and ~89–53% similarity at the amino acid level, respectively. Relative to *Arabidopsis* CCOMT 1 (the closest homologue to the alfalfa CCOMT), the three other *Arabidopsis* homologues have 70.3, 63.6 and 67.1% similarity and 58.6, 50.2 and 54.8% identity, respectively; all are cytosolic and lack any targeting peptide sequence(s). Currently, however, no *CCOMT* has been functionally characterized in *Arabidopsis*, nor are their detailed patterns of gene expression known.

Analysis of the EST database entries (Fig. 2D) indicates that of the four putative *CCOMT*s, only *CCOMT1* is expressed in all of the tissues examined, except for the above-ground organs. On the other hand, *CCOMT2* was only found in green siliques, whereas *CCOMT4* was present in mixed tissues and *CCOMT3* was not detected. It is quite unusual that no CCOMT homologue was detected in the above-ground organs, and again calls into question the stage(s) of *Arabidopsis* development that were selected for the EST databases, particularly if *CCOMT1* has a role in lignification. Obviously, there is an urgent need for the systematic delineation of the function of each *CCOMT*, with both a biochemical and metabolic context being established.

2.5. Ferulate 5-hydroxylase (F5H)

The enzyme, trivially known as ferulate 5-hydroxylase (F5H), catalyzes the cytochrome P450 dependent (with its associated reductase) introduction of a hydroxyl functionality at the 5-position. Originally, the substrate was considered to be ferulic acid (**12**) (Grand, 1984), thereby affording 5-hydroxyferulic acid (**15**), this representing a natural product first isolated from maize (*Zea mays*) (Ohashi et al., 1987). Later studies suggested that the substrate undergoing hydroxylation was either 5-hydroxyconiferyl aldehyde (**17**) or 5-hydroxyconiferyl alcohol (**18**) (Chen et al., 1999), and subsequent investigations with the corresponding recombinant F5H cytochrome P-450 from *Arabidopsis* (Humphreys et al., 1999) and sweetgum (Osakabe et al., 1999) supported this contention.

In *Arabidopsis*, however, there are two *F5H* genes (so called *F5H1* and *F5H2*, with F5H2 being 76.6% similar and 67.6% identical to F5H1. Of these, the null mutant *fah1* (mutating *F5H1*) gave an interesting phenotype, where essentially all of the detectable lignin component

was now derived from coniferyl alcohol (**3**). This was due to blocking of the corresponding 5-hydroxylation step (Meyer et al., 1998) and hence the inability to form the syringyl component of the lignin.

When analyzing the EST database entries for *Arabidopsis F5H1* and *F5H2* (Fig. 2B), it can be seen that neither was detectable in any of the organs examined, except for *F5H1* which was in the mixed tissue fraction. Again this could be due to low copy number, transient expression and/or not being inducible under the conditions employed. Clearly, as in other cases, the biochemical function of each form of *F5H* needs to be established, as do the individual patterns of gene expression and the actual metabolic context of each determined.

2.6. Caffeic acid O-methyltransferase (COMT)

Originally, this particular enzyme was considered to be exclusively involved in all *O*-methylation reactions during monolignol **2**–**4** biosynthesis. It was thought to be bifunctional in angiosperms, in terms of phenylpropanoid substrate specificity, since various crude plant extracts (Shimada, 1972; Shimada et al., 1973) were reportedly capable of converting both caffeic (**10**) and 5-hydroxyferulic (**15**) acids into ferulic (**12**) and sinapic acids (**19**), respectively; on the other hand, with the corresponding gymnosperm plant crude extracts the preferred substrate was caffeic acid (**10**), and it was thus initially concluded that gymnosperm COMTs were monofunctional.

To date, there are no known COMT homologues of very high similarity/identity present in gymnosperms at the point of writing; thus, the actual physiological significance of these crude COMT assays in gymnosperms is now questionable. However, a COMT homologue has been described in *Pinus radiata* (Moyle et al., 1999), which shows low similarity and identity (~50%) to that of the COMTs reportedly involved in sinapyl alcohol (**4**) biosynthesis in tobacco (GenBank Accession number: X74452, Jaeck et al., 1996) and aspen (GenBank Accession number: X62096, Bugos et al., 1991). It also has ~98.5% similarity/identity with a putative hydroxycinnamic acid/hydroxycinnamoyl CoA ester *O*-methyltransferase (AEOMT) from *P. taeda* (Li et al., 1997), whose physiological function has been questioned (Anterola et al., 2002). Furthermore, no heterologous expression/kinetic studies of the *P. radiata* "COMT" have been carried out in order to investigate its possible function. Its promoter has, however, been isolated, which, when fused to the *GUS* reporter gene, is expressed in both ray parenchyma and developing tracheary elements of the xylem (using transient expression) (Moyle et al., 2002).

A multifunctional *O*-methyltransferase, termed as a pinosylvin 3-*O*-methyltransferase, has also been purified from Scots pine (*P. sylvestris*), its encoding gene cloned and the corresponding protein heterologously expressed in *Escherichia coli*. This heterologous protein methylates pinosylvin, caffeic acid (**10**), caffeoyl CoA (**11**) and quercitin with K_m values of 14, 28, 26 and 26 μM, respectively, and with quite similar K_m/V_{max} values (Chiron et al., 2000). Taken together, all of these data underscore the difficulties in assigning true "physiological" roles to particular OMT isoforms.

A more incisive study, directed to identifying the actual physiological roles of angiosperm *COMT*s, was next carried out (Atanassova et al., 1995). This established that this gene (or a homologue thereof) in tobacco actually encoded an enzyme required for the second methylation step of a 5-hydroxylated substrate to ultimately give sinapyl alcohol (**4**); apparently, COMT downregulation, by contrast, had no effect on guaiacyl contents (i.e. to coniferyl alcohol (**3**) in the tobacco lignin) (Anterola and Lewis, 2002). It should be noted though that in the tobacco study, there are several *COMT* homologues and, since the CaMV 35S promoter was also used rather than its native promoter, there remains the question as to which of the COMT homologues was actually responsible for the effects on downregulating sinapyl alcohol (**4**) formation.

Analysis of the *Arabidopsis* genomic sequence indicates that some seventeen genes are annotated as putative *COMT*s, and these are tabulated in Table 2. Of these *Arabidopsis* homologues, COMT1 has the highest similarity (78.8 and 84.0%) and identity (75.5 and 79.6%) to that of the COMTs reportedly involved in sinapyl alcohol (**4**) biosynthesis in tobacco (Jaeck et al., 1996). This homologue would, therefore, represent the most likely candidate for this physiological role, particularly since relative to *Arabidopsis* COMT1, the other homologues showed significantly less similarity (~62–42%)

Table 2
Putative *Arabidopsis* COMTs as annotated by TIGR

	Locus name	Similarity (%)	Identity (%)
COMT1	At5g54160	100.0	100.0
COMT2	At1g51990	55.9	47.5
COMT3	At5g53810	56.2	48.5
COMT4	At3g53140	47.4	39.4
COMT5	At5g37170	54.6	42.9
COMT6	At1g77520	57.5	48.6
COMT7	At1g63140	58.4	47.9
COMT8	At1g21130	60.8	51.0
COMT9	At1g77530	59.9	49.9
COMT10	At1g33030	57.5	47.1
COMT11	At1g21100	60.2	50.4
COMT12	At1g21120	59.9	50.4
COMT13	At1g21110	59.4	49.3
COMT14	At1g62900	62.3	54.4
COMT15	At1g76790	58.3	47.7
COMT16	At4g35150	42.1	32.5
COMT17	At4g35160	43.1	32.8

and identity (54–33%) (Table 2), respectively. Thus, on this homology basis, *COMT1* was originally viewed as being lignin specific (Zhang et al., 1997). Indeed, this same COMT was subsequently demonstrated to methylate 5-hydroxyferulic acid (**15**), 5-hydroxyconiferyl aldehyde (**17**) and 5-hydroxyconiferyl alcohol (**18**) in coupled assays with recombinant F5H (Humphreys et al., 1999; Humphreys and Chapple, 2002), although no comparative kinetic data were reported. However, further biochemical characterization of the recombinant *Arabidopsis* COMT1 protein then indicated that it was *not* involved in lignin biosynthesis. In contrast, it was reported to be involved in flavonoid biosynthesis, being able to convert quercetin to its 3′-*O*-methyl derivative, isorhamnetin whereas caffeic (**10**) and 5-hydroxyferulic (**15**) acids were poor substrates (Muzac et al., 2000).

More recently, Goujon et al. (2003) reported an *Arabidopsis* knockout mutant, *Atomt1*, which contains next to the T-DNA border the promoter-less β-glucuronidase gene inserted as a translational fusion in the *COMT1* gene, i.e., perturbing expression of the COMT1 protein, but with the GUS gene under control of the *COMT1* promoter. The *Atomt1* mutant was very much reduced in overall OMT activity in stem crude extracts, e.g., with 6.6, 16.1, 14.9 and 17.0% activity remaining (compared to wild type) when caffeic acid (**10**), 5-hydroxyferulic acid (**15**), 5-hydroxyconiferyl aldehyde (**17**) and 5-hydroxyconiferyl alcohol (**18**) were used as substrates, respectively. However, when quercetin was used as substrate the *Atomt1* mutant showed only 1.1% activity when compared to the wild type, this apparently being consistent with the work by Muzac et al. (2000). On the other hand, the levels of flavonoids between mutant and wild type were very similar and in both cases the amounts of isorhamnetin were low suggesting no significant differences in the ability to synthesize flavonoids between the mutant and the wild type. Estimated lignin contents and monomer compositions also indicated this COMT to be involved in sinapyl alcohol (**4**) formation. That is, Klason lignin contents between the wild type plants and the *Atomt1* mutant were very similar, e.g., 14.23 and 15.51% dry weight of extract-free stem, respectively, whereas thioacidolysis showed that the levels of sinapyl alcohol (**4**) derived units were reduced from 330 to 11 μmol/g Klason lignin in the *Atomt1* mutant, with 5-hydroxylated units increasing from 8 to 20 μmol/g Klason lignin. Analysis of sinapoyl esters also showed that their levels were reduced by half in the *Atomt1* mutant when compared to wild type, with accumulation of 5-OH ferulic acid derivatives (5-OH feruloyl malate and 5-OH feruloyl glucose). Thus, taken together, these results indicate that COMT1 is indeed involved in both lignification and in sinapoyl ester biosynthesis, and not in flavonoid formation.

Furthermore, since the *Atomt1* mutant contained the GUS gene under control of the *COMT1* promoter, histochemical staining showed that GUS expression was restricted to root, leaf, stem and cotyledon vascular tissues. In the stem GUS activity was associated with the xylem, mature phloem and differentiating fibers, whereas in the flowers it was restricted to the sepal veins and in the silique to the lignified regions.

Other than the *COMT1* gene, however, none of the other *Arabidopsis COMT* homologues have been characterized in terms of their biochemical conversions in vitro, and none (excluding *COMT1*) have been studied to delineate either patterns of gene expression and/or their true metabolic contexts (roles) in vivo.

Interestingly, the analysis of the EST database entries (Fig. 2E) indicated that *COMT1* is expressed in all organs (particularly root tissues), whereas all others were either undetected (i.e. *COMT2*, *3*, *5*, *10*, *12* and *16*) and/or showed quite variable patterns of expression. As before for *4CL* and *PAL* homologues, there is thus an urgent need to now delineate the specific roles of each *COMT* homologue in vivo, paying particular attention to the patterns of gene expression as dictated by the corresponding native promoters, and the corresponding metabolic context, for each. The data obtained, however, again demonstrate thus far the very limited utility of the EST databases in identifying possible physiological functions, as well as highlighting the dangers in drawing conclusions on function on the basis of homology in the absence of demonstrable biochemical conversion data.

2.7. Cinnamoyl CoA reductase (CCR)

Originally, the cinnamoyl CoA reductases (CCRs) in various plant species have been considered to catalyze the NADPH-dependent conversions of *p*-coumaroyl CoA (**8**), feruloyl CoA (**13**) and sinapoyl CoA (**20**) into the corresponding aldehydes **9**, **14** and **21** (Fig. 1). Analysis of the *Arabidopsis* database indicates that there are 11 putative *CCR* homologues annotated with greatly differing similarities and identities from 85.8 to 40.9% and 82.8 to 32.8%, respectively, relative to CCR1 (see

Table 3
Putative *Arabidopsis* CCRs as annotated by TIGR

	Locus name	Similarity (%)	Identity (%)
CCR1	At1g15950	100	100
CCR2	At1g80820	85.8	82.8
CCR3	At2g02400	54.5	44.6
CCR4	At1g68540	50.6	40.7
CCR5	At5g14700	40.9	34.1
CCR6	At2g23910	44.4	33.6
CCR7	At4g30470	43.5	32.8
CCR8	At2g33590	52.7	45.4
CCR9/10	At1g76470	50.8	41.6
CCR11	At2g33600	51.1	44.4
CCR12	At5g58490	52.9	45.1

Table 3). All of these are presumed to be cytosolic, since they have no signal peptide targeting sequences. Of these, only *CCR1* and *CCR2* have been partially characterized in *Arabidopsis* (Lauvergeat et al., 2001), with the former being implicated with a role in lignin biosynthesis. This is based on the following: (a) its homology (75% identity) to a lignin-specific *Eucalyptus gunnii* CCR (Lacombe et al., 1997); (b) its high affinity for feruloyl (**13**) over sinapoyl (**20**) CoA, as compared to CCR2; (c) Northern blot analyses of *CCR1* transcripts, which revealed its presence in stem and floral tissues, and to a much lesser extent, in leaf tissue; and (d) a mutant defective in the *CCR1* gene, which gave phenotypes with severely (~50%) reduced lignin contents (Jones et al., 2001). In contrast, *CCR2* was strongly (albeit transiently) expressed in leaf tissue upon infection with *Xanthomonas campestris* (Lauvergeat et al., 2001) suggesting a role in plant defense.

The EST database entries (Fig. 2F) indicated that only eight of the 11 putative *CCR* homologues could be detected in various tissues and organs, but not *CCR3*, *4* and *5*. Of these, *CCR1* was apparently expressed at very similar levels in all tissues examined (including above-ground organs), this perhaps being consistent with its proposed role in lignification. On the other hand, the *CCR2* homologue was present in root tissue, in contrast to its induction in leaf tissue upon infection as noted above. Additionally, *CCR6* was only found in developing green siliques, and to a much lesser extent in mixed tissues, whereas *CCR7* was mainly expressed in green siliques and developing seeds, with lower transcript levels in the above-ground organs and mixed plant tissue. *CCR8* was also present in most organs (except root tissue), whereas *CCR9*/*10* and *12* were found only in developing seeds and *CCR11* was in the above-ground organs and mixed tissues.

Thus, as before, in spite of their annotations as cinnamoyl CoA reductases, it should be evident that much remains to be done to define (a) the patterns of gene expression for each of these homologues and (b) their precise metabolic roles in vivo. To this point, however, the vast majority of these genes have no known physiological functions or one(s) that can readily be deduced based on the information currently available.

2.8. Cinnamyl alcohol dehydrogenase (CAD)

Generically, cinnamyl alcohol dehydrogenases (CADs) catalyze the NADPH-dependent reduction of cinnamyl aldehydes to cinnamyl alcohols, and are considered as being able to reduce aldehydes **9**, **14**, **21**, and more recently **17** (Li et al., 2001). However, while differences in overall substrate versatilities have been noted between gymnosperm and angiosperm CADs (Lüderitz and Grisebach, 1981), there is a recent claim of a CAD isoform in aspen able to *specifically* reduce sinapyl aldehyde (**21**) (Li et al., 2001). This was coined sinapyl aldehyde dehydrogenase (SAD) by the authors, but, this claim was challenged since the data provided gave no proof for a specific SAD; instead a variety of cinnamyl aldehydes could be very effectively reduced by this dehydrogenase (Anterola and Lewis, 2002).

Prior to our report, there were no reports of the biochemical characterization of any *Arabidopsis* CAD isoform, even though the *Arabidopsis* genomic sequence has nine CAD homologues annotated. All of these, as expected, lack signal peptides and are presumed cytosolic. When compared with each other (relative to the arbitrarily chosen CAD1), these have 60.5–55.5% and 51.3–44.1% sequence similarities and identities to each other (Table 4). That is, there is a very great variability in this gene family, as noted for most of the other biochemical steps described thus far that have been annotated with putative functions. Furthermore, when compared with a lignin-specific tobacco CAD (Knight et al., 1992; Halpin et al., 1994), only isoforms CAD4 and CAD5 displayed significant similarity (81.5 and 82.9%) and identity (75.1 and 76.5%) to the tobacco CAD, suggesting that *Arabidopsis* CAD4 and CAD5 were likely to be lignin-specific (Table 4). Isoforms CAD6, 7 and 8 also displayed 77.3, 77.6 and 78.0%

Table 4
Putative *Arabidopsis* CADs as annotated by TIGR

	Locus name	CAD1		Tobacco		Aspen	
		Similarity (%)	Identity (%)	Similarity (%)	Identity (%)	Similarity (%)	Identity (%)
CAD1	At1g72680	100.0	100.0	57.1	46.0	58.6	49.3
CAD2	At2g21730	59.5	50.4	62.4	51.7	71.0	62.6
CAD3	At2g21890	60.5	51.3	64.0	52.1	70.2	61.5
CAD4	At3g19450	55.5	42.3	81.5	75.1	62.5	53.3
CAD5	At4g34230	56.5	44.1	82.9	76.5	62.6	53.1
CAD6	At4g37970	57.7	47.9	61.3	51.0	77.3	68.5
CAD7	At4g37980	56.7	46.2	61.4	50.1	77.6	71.4
CAD8	At4g37990	57.5	45.9	63.4	52.7	78.0	72.1
CAD9	At4g39330	58.3	50.1	60.2	50.7	74.4	67.5

Homology to *Arabidopsis* CAD1, tobacco CAD (Knight et al., 1992; Halpin et al., 1994) and aspen CAD (Li et al., 2001).

similarity and 68.5, 71.4 and 72.1% identity, respectively, to the claimed sinapyl aldehyde (**21**)-specific isoform present in aspen (Li et al., 2001). It should be noted that in concurrent work studying the biochemical role of each putative CAD isoform in *Arabidopsis*, both CAD4 and CAD5 have been shown to be lignin-specific, whereas no evidence for a specific SAD was obtained (Kim et al., in press).

In terms of EST database entries, only five of the *CAD* homologues (*CAD1*, *4*, *5*, *7* and *9*) were detected, whereas *CAD2*, *3*, *6* and *8* were not (Fig. 2G); in the latter cases, this is again perhaps due to low copy numbers, transient expression and/or being inducible only under certain conditions. This notwithstanding, the *CAD4* and *CAD5* isoforms which are involved in lignin biosynthesis were detected in roots, green siliques and mixed tissues with *CAD4* also being present in above-ground organs. Interestingly, *CAD7* and *CAD9* were also present in above-ground organs, mixed tissues and floral organ tissues, with the latter being detected in developing seeds and green siliques as well. *CAD1*, by contrast, was only found in developing seeds and green siliques.

However, given that only a few of these putative CAD isoforms have now known physiological roles (Kim et al., in press), it is again evident that a much more detailed knowledge of their individual patterns of gene expression is required as is the identification of their precise physiological roles. The EST sequence data thus are at best of limited utility at present.

3. Concluding remarks

Completion of the sequencing of the *Arabidopsis* genome in 2000 brought to science the entire blueprint of this highly studied organism. This, in turn, now offers the opportunity to systematically dissect, delineate and identify the true physiological functions of its various genes, including their responses to a variety of environmental conditions. However, the annotation of gene sequences in the public databases as having a particular and/or putative function (e.g. in the pathway to the monolignols **2**–**4**) must be viewed presently with considerable caution. This is because annotations are being made even when the homologies (similarities and identities) are very low to genes of known function; indeed, in contrast, it is well known that minute changes in homology can result in vastly different physiological functions [e.g. flavonoid versus stilbene formation (Schröder, 1999)]. As a further observation, the EST databases are also greatly lacking in terms of interpretable informational context. For example, notwithstanding the apparent enormous gaps in expression patterns, the data so collected give no insights into metabolic activities within distinct cell types in a particular tissue. Indeed, even for the known pathway steps in monolignol biosynthesis, several steps are not detectable based on the EST database collections garnered thus far. Thus, in many cases at present these databases are of most limited utility. On the other hand, the *Arabidopsis* genome sequence provides a blueprint from which true gene function can be systematically identified. This will, however, demand development of new technologies targeted to the study of individual cell types within particular tissues, whether at the RNA, DNA, protein/enzyme or metabolite level.

4. Experimental

4.1. Database analysis

In this investigation, the Institute for Genomic Research (TIGR) database, which attempts to elucidate and/or predict gene function of the 25,000+ genes of *A. thaliana*, together with The Arabidopsis Information Resource (TAIR) database, were analyzed. The purpose was to identify all potential tentative consensus sequence homologues of proteins with established and/or suspected (i.e. annotated) roles in the phenylpropanoid pathway, particularly those involved in lignin/lignan biosynthesis. Searches were performed using both protein identifier keywords and known sequences.

4.2. Signal peptide and targeting analysis

Amino acid sequences of each protein of interest, obtained from database searches, were subjected to the Center for Biological Sequence Analysis (CBS) website's SignalP (http://www.cbs.dtu.dk/services/SignalP/) and TargetP (http://www.cbs.dtu.dk/services/TargetP/) programs as well as the Human Genome Center, Institute of Medical Science, University of Tokyo, iPSORT (http://www.hypothesiscreator.net/iPSORT/) program (Bannai et al., 2002). This was in order to attempt to deduce each signal peptide sequence and its signal peptide cutoff as well as the probable subcellular location of each mature processed protein, e.g. in the chloroplast, mitochondria, secretory system, and/or nucleus/cytoplasm. The SignalP V2.0 program server predicts the presence and location of the most likely signal peptide cleavage sites in amino acid sequences, by incorporating predictions of cleavage sites and signal peptides/non-signal peptides based on a combination of several artificial neural networks and hidden Markov models (Nielsen et al., 1997). The program TargetP v1.01 predicts the subcellular locations of eukaryotic protein sequences based on analyses of N-terminal presequences, i.e. including chloroplast transit peptides, mitochondrial targeting peptides or secretory pathway signal peptides (Emanuelsson et al., 2000).

4.3. Tissue distribution expression profile analysis

Information obtained from the TIGR *A. thaliana* Gene Index (AtGI) EST (Expressed Sequence Tag) database are summarized in graphical form (Fig. 2), thereby providing a preliminary indication as regards in which tissue each gene may be dominantly expressed. Results are based on the Expression Reports for the number of ESTs for each tentative consensus (TC) DNA sequence derived from overlapping ESTs. Of these, only those with >5000 ESTs were considered, this totaling six libraries as follows: (1) *A. thaliana* roots Columbia with 17,573 ESTs (AtGI Library 5336) (2) *A. thaliana* developing seeds (collected 5–13 days after flowering) comprising 10,800 ESTs (AtGI Library 5564), (3) *A. thaliana* green siliques Columbia having 12,589 ESTs (AtGI Library 5339), (4) *A. thaliana* flower buds Columbia with 5719 ESTs (AtGI Library 5337), (5) *A. thaliana* above-ground organs 2–6-week-old having 12,264 ESTs (AtGI Library 5335), and (6) Lambda-PRL2 from mixed tissues comprising 27,631 ESTs (AtGI Library NH11). Data from each of the genes of interest obtained from the database searches were calculated as a percentage of the total number of ESTs from a particular tissue.

Acknowledgements

This research project is supported by the National Science Foundation Arabidopsis 2010 (MCB-0117260), the National Aeronautics and Space Administration (NAG 2-1513) and the G. Thomas and Anita Hargrove Center for Plant Genomic Research.

References

Allona, I., Quinn, M., Shoop, E., Swope, K., St. Cyr, S., Carlis, J., Riedl, J., Retzel, E., Campbell, M.M., Sederoff, R., Whetten, R.W., 1998. Analysis of xylem formation in pine by cDNA sequencing. Proceedings of the National Academy of Sciences of the United States of America 95, 9693–9698.

Anterola, A.M., Jeon, J.-H., Davin, L.B., Lewis, N.G., 2002. Transcriptional control of monolignol biosynthesis in *Pinus taeda*: factors affecting monolignol ratios and carbon allocation in phenylpropanoid metabolism. The Journal of Biological Chemistry 277, 18272–18280.

Anterola, A.M., Lewis, N.G., 2002. Trends in lignin modification: a comprehensive analysis of the effects of genetic manipulations/mutations on lignification and vascular integrity. Phytochemistry 61, 221–294.

Anterola, A.M., van Rensburg, H., van Heerden, P.S., Davin, L.B., Lewis, N.G., 1999. Multi-site modulation of flux during monolignol formation in loblolly pine (*Pinus taeda*). Biochemical and Biophysical Research Communications 261, 652–657.

Atanassova, R., Favet, N., Martz, F., Chabbert, B., Tollier, M.-T., Monties, B., Fritig, B., Legrand, M., 1995. Altered lignin composition in transgenic tobacco expressing *O*-methyltransferase sequences in sense and antisense orientation. The Plant Journal 8, 465–477.

Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., Miyano, S., 2002. Extensive feature detection of N-terminal protein sorting signals. Bioinformatics 18, 298–305.

Baucher, M., Chabbert, B., Pilate, G., Van Doorsselaere, J., Tollier, M.-T., Petit-Conil, M., Cornu, D., Monties, B., Van Montagu, M., Inzé, D., Jouanin, L., Boerjan, W., 1996. Red xylem and higher lignin extractability by down-regulating a cinnamyl alcohol dehydrogenase in poplar. Plant Physiology 112, 1479–1490.

Bell-Lelong, D.A., Cusumano, J.C., Meyer, K., Chapple, C., 1997. Cinnamate-4-hydroxylase expression in *Arabidopsis*. Plant Physiology 113, 729–738.

Benveniste, B., Gabriac, B., Durst, F., 1986. Purification and characterization of the NADPH-cytochrome P-450 (cytochrome *c*) reductase from higher plant microsomal fractions. The Biochemical Journal 235, 365–373.

Bernards, M.A., Lewis, N.G., 1998. The macromolecular aromatic domain in suberized tissue: a changing paradigm. Phytochemistry 47, 915–933.

Bernards, M.A., Lopez, M.L., Zajicek, J., Lewis, N.G., 1995. Hydroxycinnamic acid-derived polymers constitute the polyaromatic domain of suberin. The Journal of Biological Chemistry 270, 7382–7386.

Bouchez, D., Höfte, H., 1998. Functional genomics in plants. Plant Physiology 118, 725–732.

Bugos, R.C., Chiang, V.L., Campbell, W.H., 1991. cDNA cloning, sequence analysis and seasonal expression of lignin-bispecific caffeic acid/5-hydroxyferulic acid *O*-methyltransferance of aspen. Plant Molecular Biology 17, 1203–1215.

Chen, F., Yasuda, S., Fukushima, K., 1999. Evidence for a novel biosynthetic pathway that regulates the ratio of syringyl to guaiacyl residues in lignin in the differentiating xylem of *Magnolia kobus* DC. Planta 207, 597–603.

Chiron, H., Drouet, A., Claudot, A.-C., Eckerskorn, C., Trost, M., Heller, W., Ernst, D., Sandermann Jr, H., 2000. Molecular cloning and functional expression of a stress-induced multifunctional *O*-methyltransferase with pinosylvin methyltransferase activity from Scots pine (*Pinus sylvestris* L.). Plant Molecular Biology 44, 733–745.

Croteau, R., Kutchan, T.M., Lewis, N.G., 2000. Natural products (secondary metabolites). In: Buchanan, B.B., Gruissem, W., Jones, R.L. (Eds.), Biochemistry and Molecular Biology of Plants. American Society of Plant Physiologists, pp. 1250–1317.

Cukovic, D., Ehlting, J., VanZiffle, J.A., Douglas, C.J., 2001. Structure and evolution of 4-coumarate:coenzyme A ligase (*4CL*) gene families. Biological Chemistry 382, 645–654.

Ehlting, J., Büttner, D., Wang, Q., Douglas, C.J., Somssich, I.E., Kombrink, E., 1999. Three 4-coumarate:coenzyme A ligases in *Arabidopsis thaliana* represent two evolutionary divergent classes in angiosperms. The Plant Journal 19, 9–20.

Ehlting, J., Shin, J.J.K., Douglas, C.J., 2001. Identification of 4-coumarate:coenzyme A ligase (4CL) substrate recognition domains. The Plant Journal 27, 455–465.

Emanuelsson, O., Nielsen, H., Brunak, S., von Heijne, G., 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. The Journal of Molecular Biology 300, 1005–1016.

Fell, D., 1997. Understanding the Control of Metabolism. Portland Press, London.

Fielden, M.R., Matthews, J.B., Fertuck, K.C., Halgren, R.G., Zacharewski, T.R., 2002. In silico approaches to mechanistic and predictive toxicology: an introduction to bioinformatics for toxicologists. Critical Reviews in Toxicology 32, 67–112.

Giacomin, L.T., Szalay, A.A., 1996. Expression of *PAL1* promoter luciferase gene fusion in *Arabidopsis thaliana* in response to infection by phytopathogenic bacteria. Plant Science 116, 59–72.

Goujon, T., Sibout, R., Pollet, B., Maba, B., Nussaume, L., Bechtold, N., Lu, L., Ralph, J., Mila, I., Barrière, Y., Lapierre, C., Jouanin,

L., 2003. A new *Arabidopsis thaliana* mutant deficient in the expression of *O*-methyltransferase impacts lignins and sinapoyl esters. Plant Molecular Biology 51, 973–989.

Grand, C., 1984. Ferulic acid 5-hydroxylase: a new cytochrome P450 dependent enzyme from higher plant microsomes involved in lignin synthesis. FEBS Letters 169, 7–11.

Halpin, C., Knight, M.E., Foxon, G.A., Campbell, M.M., Boudet, A.M., Boon, J.J., Chabbert, B., Tollier, M.-T., Schuch, W., 1994. Manipulation of lignin quality by downregulation of cinnamyl alcohol dehydrogenase. The Plant Journal 6, 339–350.

Hayashi, M., Aoki, M., Kondo, M., Nishimura, M., 1997. Changes in targeting efficiencies of proteins to plant microbodies caused by amino acid substitutions in the carboxy-terminal tripeptide. Plant and Cell Physiology 38, 759–768.

Hertzberg, M., Aspeborg, H., Schrader, J., Andersson, A., Erlandsson, R., Blomqvist, K., Bhalerao, R., Uhlén, M., Teeri, T.T., Lundeberg, J., Sundberg, B., Nilsson, P., Sandberg, G., 2001. A transcriptional roadmap to wood formation. Proceedings of the National Academy of Sciences of the United States of America 98, 14732–14737.

Hrazdina, G., Jensen, R.A., 1992. Spatial organization of enzymes in plant metabolic pathways. Annual Review of Plant Physiology and Plant Molecular Biology 43, 241–267.

Hu, W.-J., Kawaoka, A., Tsai, C.-J., Lung, J., Osakabe, K., Ebinuma, H., Chiang, V.L., 1998. Compartmentalized expression of two structurally and functionally distinct 4-coumarate:CoA ligase genes in aspen (*Populus tremuloides*). Proceedings of the National Academy of Sciences of the United States of America 95, 5407–5412.

Humphreys, J.M., Chapple, C., 2002. Rewriting the lignin roadmap. Current Opinion in Plant Biology 5, 224–229.

Humphreys, J.M., Hemm, M.R., Chapple, C., 1999. New routes for lignin biosynthesis defined by biochemical characterization of recombinant ferulate 5-hydroxylase, a multifunctional cytochrome P450-dependent monooxygenase. Proceedings of the National Academy of Sciences of the United States of America 96, 10045–10050.

Inoue, K., Sewalt, V.J.H., Ballance, G.M., Ni, W., Stürzer, C., Dixon, R.A., 1998. Developmental expression and substrate specificities of alfalfa caffeic acid 3-*O*-methyltransferase and caffeoyl coenzyme A 3-*O*-methyltransferase in relation to lignification. Plant Physiology 117, 761–770.

Jaeck, E., Martz, F., Stiefel, V., Fritig, B., Legrand, M., 1996. Expression of class I *O*-methyltransferase in healthy and TMV-infected tobacco. Molecular and Plant Microbe Interaction 9, 681–688.

Jin, H., Cominelli, E., Bailey, P., Parr, A., Mehrtens, F., Jones, J., Tonelli, C., Weisshaar, B., Martin, C., 2000. Transcriptional repression by AtMYB4 controls production of UV-protecting sunscreens in *Arabidopsis*. The EMBO Journal 19, 6150–6161.

Jones, L., Ennos, A.R., Turner, S.R., 2001. Cloning and characterization of *irregular xylem4* (*irx4*): a severely lignin-deficient mutant of *Arabidopsis*. The Plant Journal 26, 205–216.

Kawaoka, A., Ebinuma, H., 2001. Transcriptional control of lignin biosynthesis by tobacco LIM protein. Phytochemistry 57, 1149–1157.

Kawaoka, A., Kaothien, P., Yoshida, K., Endo, S., Yamada, K., Ebinuma, H., 2000. Functional analysis of tobacco LIM protein Ntlim1 involved in lignin biosynthesis. The Plant Journal 22, 289–301.

Kida, Y., Ohgiya, S., Mihara, K., Sakaguchi, M., 1998. Membrane topology of NADPH-cytochrome P450 reductase on the endoplasmic reticulum. Archives of Biochemistry and Biophysics 351, 175–179.

Kim, S.-J., Kim, M.-R., Bedgar, D.L., Moinuddin, S.G.A., Cardenas, C.L., Davin, L.B., Kang, C.-H., Lewis, N.G., 2003. Functional reclassification of the putative cinnamyl alcohol dehydrogenase (CAD) multigene family in *Arabidopsis*. Proceedings of the National Academy of Sciences of the United States of America (in press).

Klockars, T., Perheentupa, T., Dahl, H.-H.M., 2003. In silico analyses of mouse inner-ear transcripts. Journal of the Association for Research in Otolaryngology 4, 24–40.

Knight, M.E., Halpin, C., Schuch, W., 1992. Identification and characterization of cDNA clones encoding cinnamyl alcohol dehydrogenase from tobacco. Plant Molecular Biology 19, 793–801.

Koukol, J., Conn, E.E., 1961. The metabolism of aromatic compounds in higher plants. IV. Purification and properties of the phenylalanine deaminase of *Hordeum vulgare*. The Journal of Biological Chemistry 236, 2692–2698.

Kragler, F., Lametschwandtner, G., Christmann, J., Hartig, A., Harada, J.J., 1998. Identification and analysis of the plant peroxisomal targeting signal 1 receptor NtPEX5. Proceedings of the National Academy of Sciences of the United States of America 95, 13336–13341.

Kubasek, W.L., Shirley, B.W., McKillop, A., Goodman, H.M., Briggs, W., Ausubel, F.M., 1992. Regulation of flavonoid biosynthetic gene in germinating *Arabidopsis* seedlings. The Plant Cell 4, 1229–1236.

Lacombe, E., Hawkins, S., Doorsselaere, J.V., Piquemal, J., Goffner, D., Poeydomenge, O., Boudet, A.M., Grima-Pettenati, J., 1997. Cinnamoyl CoA reductase, the first committed enzyme of the lignin branch biosynthetic pathway: cloning, expression and phylogenetic relationships. The Plant Journal 11, 429–441.

Lauvergeat, V., Lacomme, C., Lacombe, E., Lasserre, E., Roby, D., Grima-Pettenati, J., 2001. Two cinnamoyl-CoA reductase (CCR) genes from *Arabidopsis thaliana* are differentially expressed during development and in response to infection with pathogenic bacteria. Phytochemistry 57, 1187–1195.

Lee, D., Ellard, M., Wanner, L.A., Davis, K.R., Douglas, C.J., 1995. The *Arabidopsis thaliana* 4-coumarate:CoA ligase (*4CL*) gene: stress and developmentally regulated expression and nucleotide sequence of its cDNA. Plant Molecular Biology 28, 871–884.

Lee, D., Meyer, K., Chapple, C., Douglas, C.J., 1997. Antisense suppression of 4-coumarate:coenzyme A ligase activity in *Arabidopsis thaliana* leads to altered lignin subunit composition. The Plant Cell 9, 1985–1998.

Lewis, N.G., Davin, L.B., 1994. Evolution of lignan and neolignan biochemical pathways. In: Nes, W.D. (Ed.), Isopentenoids and Other Natural Products: Evolution and Function, Vol. 562. ACS Symposium Series, Washington, DC, pp. 202–246.

Lewis, N.G., Davin, L.B., 1999. Lignans: biosynthesis and function. In: Barton, Sir D.H.R., Nakanishi, K., Meth-Cohn, O. (Eds.), Comprehensive Natural Products Chemistry, Vol. 1. Elsevier, London, pp. 639–712.

Lewis, N.G., Davin, L.B., Sarkanen, S., 1999. The nature and function of lignins. In: Barton, Sir D.H.R., Nakanishi, K., Meth-Cohn, O. (Eds.), Comprehensive Natural Products Chemistry, Vol. 3. Elsevier, London, pp. 617–745.

Lewis, N.G., Yamamoto, E., 1990. Lignin: occurrence, biogenesis and biodegradation. Annual Review of Plant Physiology and Plant Molecular Biology 41, 455–496.

Leyva, A., Jarillo, J.A., Salinas, J., Martinez-Zapater, J.M., 1995. Low temperature induces the accumulation of *phenylalanine ammonia lyase* and *chalcone synthase* mRNAs of *Arabidopsis thaliana* in a light-dependent manner. Plant Physiology 108, 39–46.

Li, L., Cheng, X.F., Leshkevich, J., Umezawa, T., Harding, S.A., Chiang, V.L., 2001. The last step of syringyl monolignol biosynthesis in angiosperms is regulated by a novel gene encoding sinapyl alcohol dehydrogenase. The Plant Cell 13, 1567–1585.

Li, L., Popko, J.L., Zhang, X.H., Osakabe, K., Tsai, C.J., Joshi, C.P., Chiang, V.L., 1997. A novel multifunctional *O*-methyltransferase implicated in a dual methylation pathway associated with lignin biosynthesis in loblolly pine. Proceedings of the National Academy of Sciences of the United States of America 94, 5461–5466.

Liang, X., Dron, M., Cramer, C.L., Dixon, R.A., Lamb, C.J., 1989. Differential regulation of phenylalanine ammonia-lyase genes

during plant development and by environmental cues. The Journal of Biological Chemistry 264, 14486–14492.

Lüderitz, T., Grisebach, H., 1981. Enzymic synthesis of lignin precursors. Comparison of cinnamoyl-CoA reductase and cinnamyl alcohol:NADP$^+$ dehydrogenase from spruce (*Picea abies* L.) and soybean (*Glycine max* L.). The European Journal of Biochemistry 119, 115–124.

Meinke, D.W., Cherry, J.M., Dean, C., Rounsley, S.D., Koornneef, M., 1998. *Arabidopsis thaliana*: a model plant for genome analysis. Science 282, 662, 679–682.

Meyer, K., Shirley, A.M., Cusumano, J.C., Bell-Lelong, D.A., Chapple, C., 1998. Lignin monomer composition is determined by the expression of a cytochrome P450-dependent monooxygenase in *Arabidopsis*. Proceedings of the National Academy of Sciences of the United States of America 95, 6619–6623.

Moyle, R., Moody, J., Phillips, L., Walter, C., Wagner, A., 2002. Isolation and characterization of a *Pinus radiata* lignin biosynthesis-related *O*-methyltransferase promoter. Plant Cell Reports 20, 1052–1060.

Moyle, R., Wagner, A., Walter, C., 1999. Nucleotide sequence of a caffeic acid *O*-methyltransferase (COMT) gene (Accession No. AF119225) from *Pinus radiata*. Plant Physiology 119, 1147.

Muzac, I., Wang, J., Anzellotti, D., Zhang, H., Ibrahim, R.K., 2000. Functional expression of an *Arabidopsis* cDNA clone encoding a flavonol 3′-*O*-methyltransferase and characterization of the gene product. Archives of Biochemistry and Biophysics 375, 385–388.

Nair, R.B., Xia, Q., Kartha, C.J., Kurylo, E., Hirji, R.N., Dalta, R., Selvaraj, G., 2002. *Arabidopsis* CYP98A3 mediating aromatic 3-hydroxylation. Developmental regulation of the gene and expression in yeast. Plant Physiology 130, 210–220.

Nielsen, H., Engelbrecht, J., Brunak, S., von Heijne, G., 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Engineering 10, 1–6.

Ohashi, H., Yamamoto, E., Lewis, N.G., Towers, G.H.N., 1987. 5-Hydroxyferulic acid in *Zea mays* and *Hordeum vulgare* cell walls. Phytochemistry 26, 1915–1916.

Ohl, S., Hedrick, S.A., Chory, J., Lamb, C.J., 1990. Functional properties of a phenylalanine ammonia-lyase promoter from *Arabidopsis*. The Plant Cell 2, 837–848.

Osakabe, K., Tsao, C.C., Li, L., Popko, J.L., Umezawa, T., Carraway, D.T., Smeltzer, R.H., Joshi, C.P., Chiang, V.L., 1999. Coniferyl aldehyde 5-hydroxylation and methylation direct syringyl lignin biosynthesis in angiosperms. Proceedings of the National Academy of Sciences of the United States of America 96, 8955–8960.

Pakusch, A.-E., Kneusel, R.E., Matern, U., 1989. *S*-Adenosyl-L-methionine:*trans*-caffeoyl-coenzyme A 3-*O*-methyltransferase from elicitor-treated parsley cell suspension cultures. Archives of Biochemistry and Biophysics 271, 488–494.

Pakusch, A.-E., Matern, U., 1991. Kinetic characterization of caffeoyl-coenzyme A specific 3-*O*-methyltransferase from elicited parsley cell suspensions. Plant Physiology 96, 327–330.

Saslowsky, D., Winkel-Shirley, B., 2001. Localization of flavonoid enzymes in *Arabidopsis* roots. The Plant Journal 27, 37–48.

Schmitt, D., Pakusch, A.-E., Matern, U., 1991. Molecular cloning, induction, and taxonomic distribution of caffeoyl-CoA 3-*O*-methyltransferase, an enzyme involved in disease resistance. The Journal of Biological Chemistry 266, 17416–17423.

Schoch, G., Goepfert, S., Morant, M., Hehn, A., Meyer, D., Ullmann, P., Werck-Reichhart, D., 2001. CYP98A3 from *Arabidopsis thaliana* is a 3′-hydroxylase of phenolic esters, a missing link in the phenylpropanoid pathway. The Journal of Biological Chemistry 276, 36566–36574.

Schröder, J., 1999. The chalcone/stilbene synthase-type family of condensing enzymes. In: Barton, Sir D.H.R., Nakanishi, K., Meth-Cohn, O. (Eds.), Comprehensive Natural Products Chemistry, Vol. 1. Elsevier, London, pp. 749–771.

Shimada, M., 1972. Biochemical studies on bamboo lignin and methoxylation in hardwood and softwood lignins. Wood Research 53, 19–65.

Shimada, M., Fushiki, H., Higuchi, T., 1973. Mechanism of biochemical formation of the methoxyl groups in softwood and hardwood lignins. Mokuzai Gakkaishi 19, 13–21.

Sterky, F., Regan, S., Karlsson, J., Hertzberg, M., Rhode, 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., Uhlén, M., Sundberg, B., Lundeberg, J., 1998. Gene discovery in the wood-forming tissues of poplar: analysis of 5692 expressed sequence tags. Proceedings of the National Academy of Sciences of the United States of America 95, 13330–13335.

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

The Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408, 796–815.

Urban, P., Mignotte, C., Kazmaier, M., Delorme, F., Pompon, D., 1997. Cloning, yeast expression, and characterization of the coupling of two distantly related *Arabidopsis thaliana* NADPH-cytochrome P450 reductases with P450 CYP73A5. Proceedings of the National Academy of Sciences of the United States of America 272, 19176–19186.

van Heerden, P.S., Towers, G.H.N., Lewis, N.G., 1996. Nitrogen metabolism in lignifying *Pinus taeda* cell cultures. The Journal of Biological Chemistry 271, 12350–12355.

White, J.A., Todd, J., Newman, T., Focks, N., Girke, T., Martínez de Ilárduya, O., Jaworski, J.G., Ohlrogge, J.B., Benning, C., 2000. A new set of *Arabidopsis* expressed sequence tags from developing seeds. The metabolic pathway from carbohydrates to seed oil. Plant Physiology 124, 1582–1594.

Wiermann, R., Gubatz, S., 1992. Pollen wall and sporopollenin. International Review of Cytology 140, 35–72.

Zhang, H., Wang, J., Goodman, H.M., 1997. An *Arabidopsis* gene encoding a putative 14-3-3-interacting protein, caffeic acid/5-hydroxyferulic acid *O*-methyltransferase. Biochimica et Biophysica Acta 1353, 199–202.