PCP
PLANT & CELL PHYSIOLOGY

Regular Paper

Towards a Systems Approach for Lignin Biosynthesis in *Populus trichocarpa*: Transcript Abundance and Specificity of the Monolignol Biosynthetic Genes

Rui Shi[1,4], Ying-Hsuan Sun[1,4], Quanzi Li[1], Steffen Heber[2], Ronald Sederoff[1] and Vincent L. Chiang[1,3,*]

[1]Forest Biotechnology Group, Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, NC 27695, USA
[2]Bioinformatics Research Center, Department of Computer Science, North Carolina State University, Raleigh, NC 27695, USA
[3]Department of Wood and Paper Science, North Carolina State University, Raleigh, NC 27695, USA
[4]These authors contributed equally to this work.
*Corresponding author: E-mail, vincent_chiang@ncsu.edu; Fax, +1-919-515-7801

(Received September 25, 2009; Accepted December 2, 2009)

As a step toward a comprehensive description of lignin biosynthesis in *Populus trichocarpa*, we identified from the genome sequence 95 phenylpropanoid gene models in 10 protein families encoding enzymes for monolignol biosynthesis. Transcript abundance was determined for all 95 genes in xylem, leaf, shoot and phloem using quantitative real-time PCR (qRT-PCR). We identified 23 genes that most probably encode monolignol biosynthesis enzymes during wood formation. Transcripts for 18 of the 23 are abundant and specific to differentiating xylem. We found evidence suggesting functional redundancy at the transcript level for phenylalanineammonia-lyase(PAL),cinnamate4-hydroxylase (C4H), 4-coumarate:CoA ligase (4CL), *p*-hydroxycinnamoyl-CoA:quinate shikimate *p*-hydroxycinnamoyltransferase (HCT), caffeoyl-CoA *O*-methyltransferase (CCoAOMT) and coniferyl aldehyde 5-hydroxylase (CAld5H). We carried out an enumeration-based motif identification and discriminant analysis on the promoters of all 95 genes. Five core motifs correctly discriminate the 18 xylem-specific genes from the 77 non-xylem genes. These motifs are similar to promoter elements known to regulate phenylpropanoid gene expression. This work suggests that genes in monolignol biosynthesis are regulated by multiple motifs, often related in sequence.

Keywords: Lignin systems biology • Monolignol biosynthesis • *Populus trichocarpa* • Promoter motifs • Transcript abundance • Xylem-specific expression.

Abbreviations: CAD, cinnamyl alcohol dehydrogenase; CAld5H, coniferyl aldehyde 5-hydroxylase; CCoAOMT, caffeoyl-CoA *O*-methyltransferase; CCR, cinnamoyl-CoA reductase; CesA, cellulose synthase A; C3H, 4-coumarate 3-hydroxylase; C4H, cinnamate-4-hydroxylase; 4CL, 4-coumarate:CoA ligase; COMT, caffeic acid/5-hydroxyconiferaldehyde *O*-methyltransferase; EST, expressed sequence tag; HCT, *p*-hydroxycinnamoyl-CoA:quinate shikimate *p*-hydroxycinnamoyltransferase; LC-MS/MS, liquid chromatography–tandem mass spectrometry; ORF, open reading frame; PAL, phenylalanine ammonia-lyase; qRT-PCR, quantitative real-time PCR; RACE, rapid amplification of cDNA ends; SAD, sinapyl alcohol dehydrogenase; UTR, untranslated region.

The nucleotide sequences reported in this paper have been submitted to GenBank under accession numbers: EU603298, EU603299, EU603300, EU603301, EU603302, EU603303, EU603304, EU603305, EU603306, EU603307, EU603308, EU603309, EU603310, EU603311, EU603312, EU603313, EU603314, EU603315, EU603316, EU603317, EU603318, EU603319, EU603320, EU603321, EU603322.

Introduction

Lignin is one of the most abundant polymers in the terrestrial biosphere and plays an important role in the nature of trees and forests. Removal of lignin for processing of wood into paper is energy and chemical intensive (Chiang 2002, Ragauskas et al. 2006). The conversion efficiency of lignocellulosic biomass to ethanol is determined largely by lignin (Sarkanen 1976, Ragauskas et al. 2006, Chen and Dixon 2007). Lignin also affects the digestibility of forage (Jung and Deetz 1993). A systems approach to studying lignin biosynthesis would increase our understanding of secondary metabolism and improve the conversion of plants into energy, food and materials.

A systems approach involves building predictive models of complex biological processes (Kitano 2002, Albert 2007). Many aspects of the biosynthesis of lignin are not sufficiently understood or quantified for predictive model development. Understanding this pathway requires knowledge of the quantitative

Plant Cell Physiol. 51(1): 144–163 (2010) doi:10.1093/pcp/pcp175, available online at www.pcp.oxfordjournals.org

relationships of all components—genes, regulatory sequences, transcripts, proteins, metabolites, lignin structural units and linkages. The sequence of the genome of *Populus trichocarpa* (Nisqually-1) (Tuskan et al. 2006) and the available genomic, biochemical and chemical tools make possible the generation of much of the information needed.

Lignin is polymerized from three primary monomers, *p*-coumaryl alcohol (**20**, **Fig. 1**), coniferyl alcohol (**22**) and sinapyl alcohol (**24**), the H, G and S monolignols (Sarkanen 1971, Higuchi 1997, Ralph et al. 2008). In dicots, such as *Populus*, lignin is polymerized from S and G monolignols and low levels of H monolignols and other intermediates (light color, **Fig. 1**). Knowledge of key metabolites and enzymes stems from tracer studies by Brown, Neish and Higuchi (Brown and Neish 1955, Higuchi and Brown 1963). A metabolic grid (Higuchi 2003, Dixon et al. 2001, Ralph et al. 2008) of 10 enzyme families converts phenylalanine (**1**) to monolignols (**Fig. 1**), with a principal path within the grid (heavy color). Specific enzymes divert the G monolignol pathway at coniferaldehyde to form the S monolignol (Osakabe et al. 1999, Humphreys et al. 1999, Li et al. 2000, Dixon et al. 2001, Li et al. 2001). Then monolignols are transported to the lignifying zone (Sarkanen 1971, Wardrop 1981), where lignin is polymerized by oxidative free radical-based coupling of monomers (Harkin 1967, Freudenberg and Neish 1968, Higuchi 1997), following a combinatorial model (Ralph et al. 2008). The combinatorial model is the most accepted view of lignin biosynthesis, and readily explains the high variation in content, composition and linkage structure (Ralph et al. 2008).

Two studies have been carried out to identify *Populus* genes involved in phenylpropanoid metabolism, one using genome sequence, microarray analysis and expressed sequence tags (ESTs) from several species (Hamberger et al. 2007), and the other using semi-quantitative estimates of transcript abundance in a hybrid (*P. fremontii* × *angustifolia*) (Tsai et al. 2006). Hamberger et al. (2007) implicated 15 genes in lignin biosynthesis in *Populus*, based on tissue specificity of the ESTs. In Arabidopsis, 14 candidate genes were identified for vascular lignification by Raes et al. (2003).

In this study we focused on *P. trichocarpa* for quantifying the expression and regulation of genes involved in monolignol biosynthesis. We surveyed 95 gene models through transcript quantitation and tissue comparisons, and characterized and identified 23 genes potentially involved in monolignol biosynthesis during wood formation. We describe relative steady-state transcript abundance that is preferable to differentiating xylem as xylem-specific expression. We have sampled only four major tissues in this study; however, our quantitative inferences about abundance and specificity are consistent with the best characterized *Populus* monolignol genes (see Results and Vanholme et al. 2008). In these genes, we then identified promoter motifs that are associated with specificity and abundance in differentiating xylem.

Computational approaches to identify regulatory elements in plant promoters are derived from similar studies in animals (Hudson and Quail 2003, Rombauts et al. 2003). There are many programs available for sequence motif identification (Thompson et al. 2003, Bailey et al. 2006, D'haeseleer 2006). We chose DME-X, an enumeration algorithm-based program, because it can incorporate weighted parameters, such as transcript abundance or tissue specificity, to identify regulatory motifs (Smith et al. 2005). Knowledge of the genes and regulatory elements involved in lignin biosynthesis is an essential component of a comprehensive systems approach.

Results

Identification of genes associated with the phenylpropanoid biosynthetic pathway in the P. trichocarpa genome

A total of 45,555 *P. trichocarpa* gene models (annotation v1.1, http://genome.jgi-psf.org/Poptr1_1/) were surveyed to identify phenylpropanoid pathway genes, based on the annotated Arabidopsis genes. Of 169 gene models identified, 84 were truncated at the 5′ or 3′ end. Manual editing of the first and last open reading frames (ORFs) of these 84 genes allowed us to retrieve six additional possible full-gene models that may have been incorrectly truncated by the annotation (v1.1). These six are: one 4-coumarate:CoA ligase (4CL), one caffeoyl-CoA *O*-methyltransferase (CCoAOMT), one cinnamoyl-CoA reductase (CCR), two caffeic acid/5-hydroxyconiferaldehyde *O*-methyltransferases (COMTs) and one cinnamyl alcohol dehydrogenase (CAD). Information on the remaining 78 truncated models is listed in **Supplementary Table S1**. Four additional full-gene models for CCRs (CCR1, 3, 5 and 6), which were not in the annotation v1.1 but were identified by Tuskan et al. (2006) and Hamberger et al. (2007), were also included in our analysis. Overall, 95 full-gene models were identified as putative phenylpropanoid biosynthesis genes in *P. trichocarpa*. These 95 genes contain ORFs that encode five phenylalanine ammonia-lyases (PALs), three cinnamate-4-hydroxylases (C4Hs), 17 4CLs, seven *p*-hydroxycinnamoyl-CoA:quinate shikimate *p*-hydroxycinnamoyltransferases (HCTs), four 4-coumarate 3-hydroxylases (C3Hs), six CCoAOMTs, nine CCRs, three coniferyl aldehyde 5-hydroxylases (CAld5Hs), 25 COMTs and 16 CADs (**Table 1**). Phylogenetic relationships of the genes for each family are shown in **Supplementary Fig. S1**. The numerical IDs for the known gene family members (bold in **Table 1**) follow Tuskan et al. (2006). Newly identified members were numbered sequentially following the known genes.

Tissue-specific expression profiles of 95 putative P. trichocarpa phenylpropanoid biosynthetic pathway genes

To identify genes functionally associated with the monolignol biosynthesis, we carried out rigorous quantitative real-time PCR (qRT-PCR) to determine the abundance and specificity of transcripts for all 95 putative phenylpropanoid genes in four *P. trichocarpa* tissues, stem differentiating xylem (X), stem

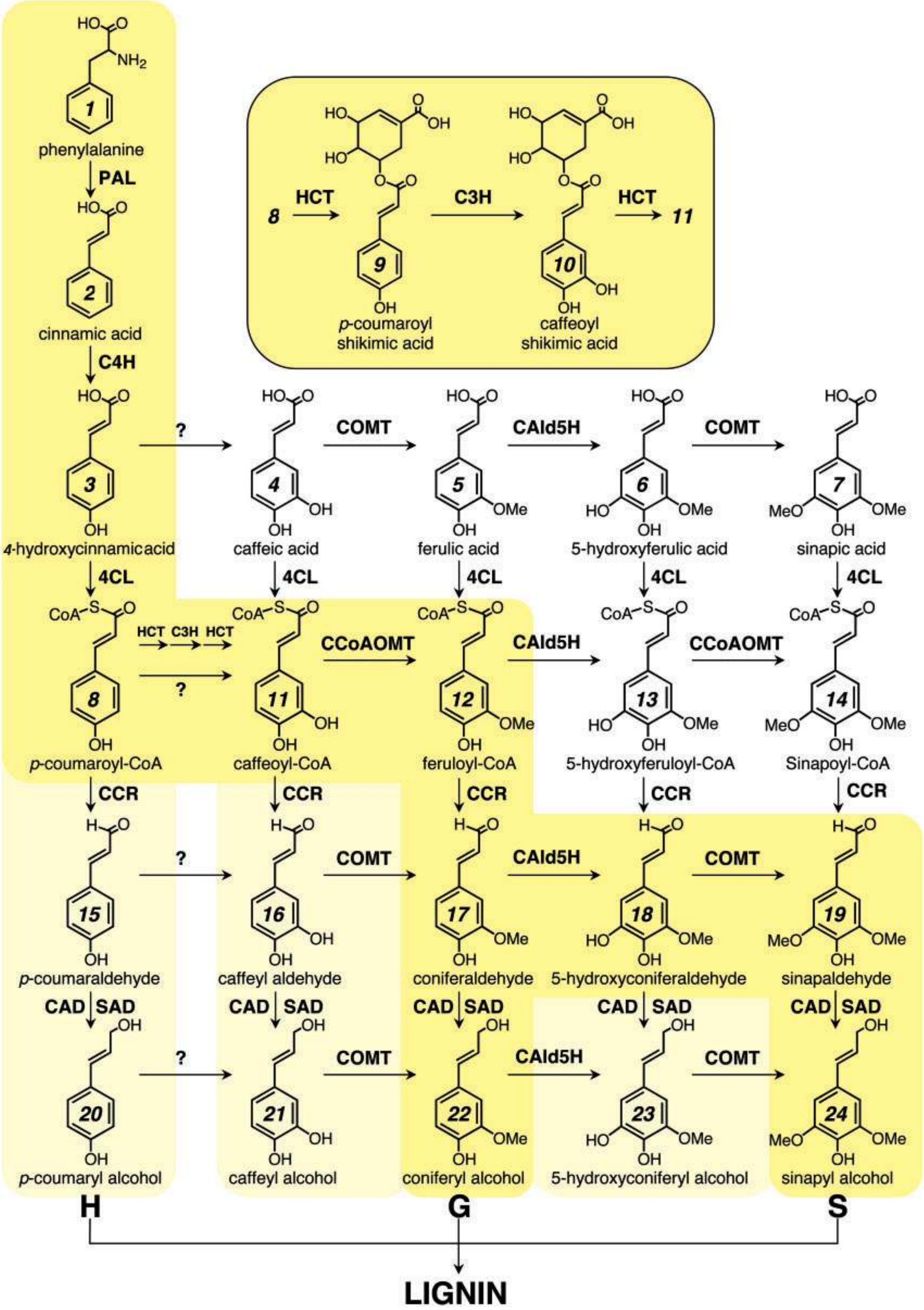


Fig. 1 Proposed monolignol biosynthesis pathway. PAL (phenylalanine ammonia-lyase), C4H (cinnamate-4-hydroxylase), 4CL (4-coumarate:CoA ligase), HCT (*p*-hydroxycinnamoyl-CoA:quinate shikimate *p*-hydroxycinnamoyltransferase), C3H (4-coumarate 3-hydroxylase), CCoAOMT (caffeoyl-CoA *O*-methyltransferase), CCR (cinnamoyl-CoA reductase), CAld5H (coniferyl aldehyde 5-hydroxylase), COMT (caffeic acid/5-hydroxyconiferaldehyde *O*-methyltransferase), CAD (cinnamyl alcohol dehydrogenase), SAD (sinapyl alcohol dehydrogenase).

Table 1 Ninety-five putative *P. trichocarpa* phenylpropanoid genes in 10 enzyme families

Gene family	Gene name[a]	JGI id	Gene model name
PAL	***PtrPAL1***	**739479**	**estExt_Genewise1_v1.C_280658**
	PtrPAL2	**820249**	**estExt_fgenesh4_pg.C_LG_VIII0293**
	PtrPAL3	**667815**	**grail3.0004045401**
	PtrPAL4	**822571**	**estExt_fgenesh4_pg.C_LG_X2023**
	PtrPAL5	770553[b]	fgenesh4_pg.C_LG_X002043
		228016	**gw1.X.2713.1**
C4H	***PtrC4H1***	**823837**	**estExt_fgenesh4_pg.C_LG_XIII0519**
	PtrC4H2	**665925**	**grail3.0094002901**
	PtrC4H3	786206[b]	fgenesh4_pg.C_scaffold_164000062
		584578	**eugene3.01640067**
4CL	***Ptr4CL1***	**827715**	**estExt_fgenesh4_pg.C_1210004**
	Ptr4CL2	**262277**	**gw1.XVIII.2818.1**
	Ptr4CL3	**639764**	**grail3.0100002702**
	Ptr4CL4	**665396**	**grail3.0099003002**
	Ptr4CL5	**758231**	**fgenesh4_pg.C_LG_III001773**
	Ptr4CL6	797174	fgenesh4_pm.C_LG_I000177
	Ptr4CL7	761575	fgenesh4_pg.C_LG_V001627
	Ptr4CL8	550798	eugene3.00020113
	Ptr4CL9	757240	fgenesh4_pg.C_LG_III000782
	Ptr4CL10	837028	estExt_fgenesh4_pm.C_1230033
	Ptr4CL11	831184	estExt_fgenesh4_pm.C_LG_IV0315
	Ptr4CL12	827131	estExt_fgenesh4_pg.C_640066
	Ptr4CL13	228198[c]	gw1.X.2895.1
	Ptr4CL14	768982	fgenesh4_pg.C_LG_X000472
	Ptr4CL15	834351	estExt_fgenesh4_pm.C_LG_XII0345
	Ptr4CL16	824740	estExt_fgenesh4_pg.C_LG_XV0666
	Ptr4CL17	662119	grail3.0015024001
HCT	***PtrHCT1***	**554899**	**eugene3.00031532**
	PtrHCT2	**835948**	**estExt_fgenesh4_pm.C_LG_XVIII0344**
	PtrHCT3	**825948**	**estExt_fgenesh4_pg.C_LG_XVIII0910**
	PtrHCT4	**578723**	**eugene3.00180947**
	PtrHCT5	**586045**	**eugene3.18780002**
	PtrHCT6	**587193**	**eugene3.02080010**
	PtrHCT7	**784746**	**fgenesh4_pg.C_scaffold_133000007**
C3H	***PtrC3H1***	**591136**	**eugene3.36160002**
	PtrC3H2	**576112**	**eugene3.00160247**
	PtrC3H3	**831788**	**estExt_fgenesh4_pm.C_LG_VI0096**
	PtrC3H4	582895	eugene3.14540001
CCoAOMT	***PtrCCoAOMT1***	**649581**	**grail3.0001059501**
	PtrCCoAOMT2	**829835**	**estExt_fgenesh4_pm.C_LG_I1023**
	PtrCCoAOMT3	**820698**	**estExt_fgenesh4_pg.C_LG_VIII1209**
	PtrCCoAOMT4	805093	fgenesh4_pm.C_LG_X000399
	PtrCCoAOMT5	837287	estExt_fgenesh4_pm.C_1450034

Continued

Table 1 Continued

Gene family	Gene name[a]	JGI id	Gene model name
	PtrCCoAOMT6	409538[c]	gw1.II.873.1
CCR	***PtrCCR1***	**828721**[d]	**estExt_fgenesh4_pg.C_2080034**
	PtrCCR2	**813757**	**estExt_fgenesh4_kg.C_LG_III0056**
	PtrCCR3	**787395**[d]	**fgenesh4_pg.C_scaffold_208000040**
	PtrCCR5	**828723**[d]	**estExt_fgenesh4_pg.C_2080041**
	PtrCCR6	**814849**[d]	**estExt_fgenesh4_pg.C_LG_I0389**
	PtrCCR8	297354	gw1.86.245.1
	PtrCCR9	722576	estExt_Genewise1_v1.C_LG_IX2899
	PtrCCR10	710668	estExt_Genewise1_v1.C_LG_II2105
	PtrCCR11	410792[c]	gw1.II.2127.1
CAld5H	***PtrCAld5H1***	679182[b]	grail3.0057011701
		836596	**estExt_fgenesh4_pm.C_570058**
	PtrCAld5H2	**563244**	**eugene3.00071182**
	PtrCAld5H3	557180	Eugene3.00090440
COMT	***PtrCOMT1***	**824484**	**estExt_fgenesh4_pg.C_LG_XV0035**
	PtrCOMT2	**834247**	**estExt_fgenesh4_pm.C_LG_XII0129**
	PtrCOMT3	731466	estExt_Genewise1_v1.C_LG_XIV1942
	PtrCOMT4	552360	Eugene3.00021675
	PtrCOMT5	799151	fgenesh4_pm.C_LG_II000840
	PtrCOMT6	806142	fgenesh4_pm.C_LG_XI000417
	PtrCOMT7	836247	estExt_fgenesh4_pm.C_280112
	PtrCOMT8	811844	fgenesh4_pm.C_scaffold_187000003
	PtrCOMT9	758840	fgenesh4_pg.C_LG_IV000468
	PtrCOMT10	555739	Eugene3.00040452
	PtrCOMT11	586060	Eugene3.01870010
	PtrCOMT12	780715	fgenesh4_pg.C_LG_XIX000854
	PtrCOMT13	829076	estExt_fgenesh4_pg.C_19310001
	PtrCOMT14	740577	estExt_Genewise1_v1.C_410003
	PtrCOMT15	200573[c]	gw1.IX.1038.1
	PtrCOMT16	588324	Eugene3.02390008
	PtrCOMT17	669875	grail3.1005000101
	PtrCOMT18	828771	estExt_fgenesh4_pg.C_2390003
	PtrCOMT19	676887	grail3.3096000101
	PtrCOMT20	582798	Eugene3.01420105
	PtrCOMT21	280232	gw1.239.28.1
	PtrCOMT22	571762	Eugene3.00131126
	PtrCOMT23	597596	Eugene3.09770002
	PtrCOMT24	827707	estExt_fgenesh4_pg.C_1200063
	PtrCOMT25	176981[c]	gw1.I.5581.1
CAD	***PtrCAD1***	**722315**	**estExt_Genewise1_v1.C_LG_IX2359**
	PtrCAD2	667694	grail3.0004034803
	PtrCAD3	804428	fgenesh4_pm.C_LG_IX000475
	PtrCAD4	587092	Eugene3.20690001

Continued

Table 1 Continued

Gene family	Gene name[a]	JGI id	Gene model name
	PtrCAD5	767919	fgenesh4_pg.C_LG_IX000970
	PtrCAD6	557759	Eugene3.00091019
	PtrCAD7	549334	eugene3.00011775
	PtrCAD8	831981	estExt_fgenesh4_pm.C_LG_VI0462
	PtrCAD9	735178	estExt_Genewise1_v1.C_LG_XVI2049
	PtrCAD10	815665	estExt_fgenesh4_pg.C_LG_I2533
	PtrCAD11	753324	fgenesh4_pg.C_LG_I002927
	PtrCAD12	550847	eugene3.00020162
	PtrCAD13	758155	fgenesh4_pg.C_LG_III001697
	PtrCAD14	825006	estExt_fgenesh4_pg.C_LG_XVI0159
	PtrCAD15	232836	gw1.XI.816.1
	PtrCAD16	417496[c]	gw1.VI.1869.1

[a] Gene information in bold is for those identified by Tuskan et al. (2006).
[b] Alternative gene model (JGI ID) descriptors for these loci were used by Tuskan et al. (2006).
[c] *P. trichocarpa* v1.1 models were truncated. Full models were obtained by manually editing the first and last ORFs.
[d] These gene models were not included in the *P. trichocarpa* v1.1 45,555 model set.

differentiating phloem (P), shoot tip (S) and fully expanded leaf (L) (**Fig. 2**). For each gene, 2–4 primer sets (**Supplementary Table 2**) were used to increase specificity, accuracy and precision of our transcript quantification. Absolute transcript levels were estimated for each member of each family as transcript copy numbers in units of total RNA (Suzuki et al. 2006; **Supplementary Table S2**). Our final primer sets have passed stringent dissociation curve analysis. All ‘xylem-specific’ monolignol genes selected show a single PCR product based on the dissociation profile and show an expected single sequence for the product from xylem tissue.

Gene-specific transcript abundance of the PAL family

The deamination of phenylalanine by PAL is the first step in monolignol biosynthesis (**Fig. 1**; Higuchi, 1997). *PAL* genes have been studied in many plant species (**Supplementary Table S3**). Some *PAL* genes are specifically expressed in differentiating xylem, while others are expressed in many tissues. In *P. trichocarpa*, five *PAL* gene models (*PtrPAL1–PtrPAL5*) were identified, consistent with other *Populus* species (**Supplementary Table S3**). Transcripts of all five *PtrPAL* genes were abundant in xylem (**Fig. 2a**; **Supplementary Tables S2, S3**). *PtrPAL1* and *PtrPAL3* were also expressed at moderate levels in shoot, phloem and leaf. *PtrPAL1* transcripts were most abundant in shoot, suggesting involvement in other pathways (Kao et al. 2002). *PAL* genes in other *Populus* species most similar to *PtrPAL1* or *PtrPAL3* were found to be expressed in many tissues, including non-woody tissues (**Supplementary Table S3**).

PtrPAL2, *4* and *5* are more xylem specific. Homologs of *PtrPAL2* and *PtrPAL4* in *P. fremontii* × *angustifolia* (Tsai et al. 2006) and in *Populus* ESTs (Sterky et al. 2004, Hamberger et al. 2007) were also suggested to be xylem specific. While these studies could not distinguish the transcripts of the homologs corresponding to *PtrPAL4* and *PtrPAL5* (98.4% protein sequence identity; **Supplementary Table S4**), Osakabe et al. (1995) detected expression of homologs of both *PtrPAL4* and *PtrPAL5* in developing xylem in *P. kitakamiensis*. Expression of homologs of *PtrPAL2*, *4* or *5* in other *Populus* species was also specific to xylem (**Supplementary Table S3**).

Gene-specific transcript abundance of the C4H family

C4H (CYP73) converts cinnamic acid into 4-hydroxycinnamic acid (**Fig. 1**), a precursor for many phenylpropanoids including flavonoids, phytoallexins and monolignols (Hahlbrock and Scheel 1989, Dixon et al. 2006, Lu et al. 2006). In the *P. trichocarpa* genome, there are three *C4H* gene models (**Table 1**). Transcripts of *PtrC4H1* and *PtrC4H2* (96.6% protein sequence identity; **Supplementary Table S4**) were abundant in differentiating xylem (**Fig. 2b**; **Supplementary Tables S2, S3**), suggesting that both are important in monolignol biosynthesis (Lu et al. 2006). *C4H1* and *C4H2* in *P. tremuloides* and two *Populus* xylem ESTs are orthologs of *PtrC4H2* and *PtrC4H1*, respectively. Previously studied *Populus C4H* genes that are similar to *PtrC4H1* or *PtrC4H2* are xylem specific (**Supplementary Table S3**). Transcripts of *PtrC4H3*, not previously characterized, had little or no expression in all examined tissues (**Fig. 2, Supplementary Table S2**).

Gene-specific transcript abundance of the 4CL family

4CL catalyzes the formation of CoA thioesters of several hydroxycinnamic acids, including 4-hydroxycinnamic acid, caffeic acid, ferulic acid, 5-hydroxyferulic acid and sinapic acid

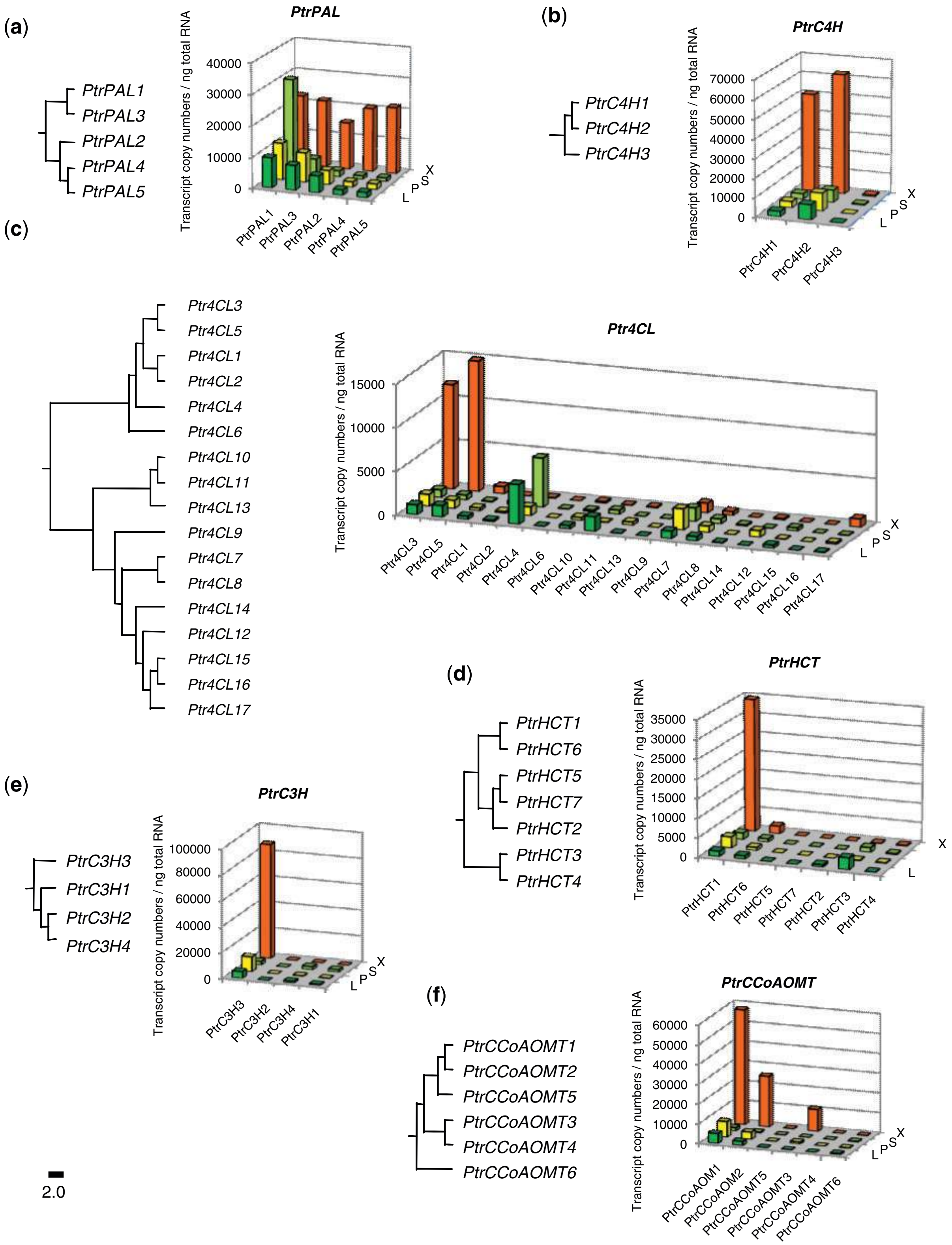


Fig. 2 qRT-PCR-based tissue-specific transcript abundances and phylogenetic analysis of 95 *P. trichocarpa* phenylpropanoid genes. Tissues were collected from three trees, which were pooled before RNA extraction for qRT-PCR. Three PCRs were carried out for each pooled sample. Mean transcript copy numbers and standard error of the means from three technical replicates are given in **Supplementary Table S2**. Neighbor–joining phylogenetic trees were generated using ClustalW with default settings for the protein sequences. Detailed phylogenetic trees with bootstrap values are shown in **Supplementary Fig. S1.**

continued

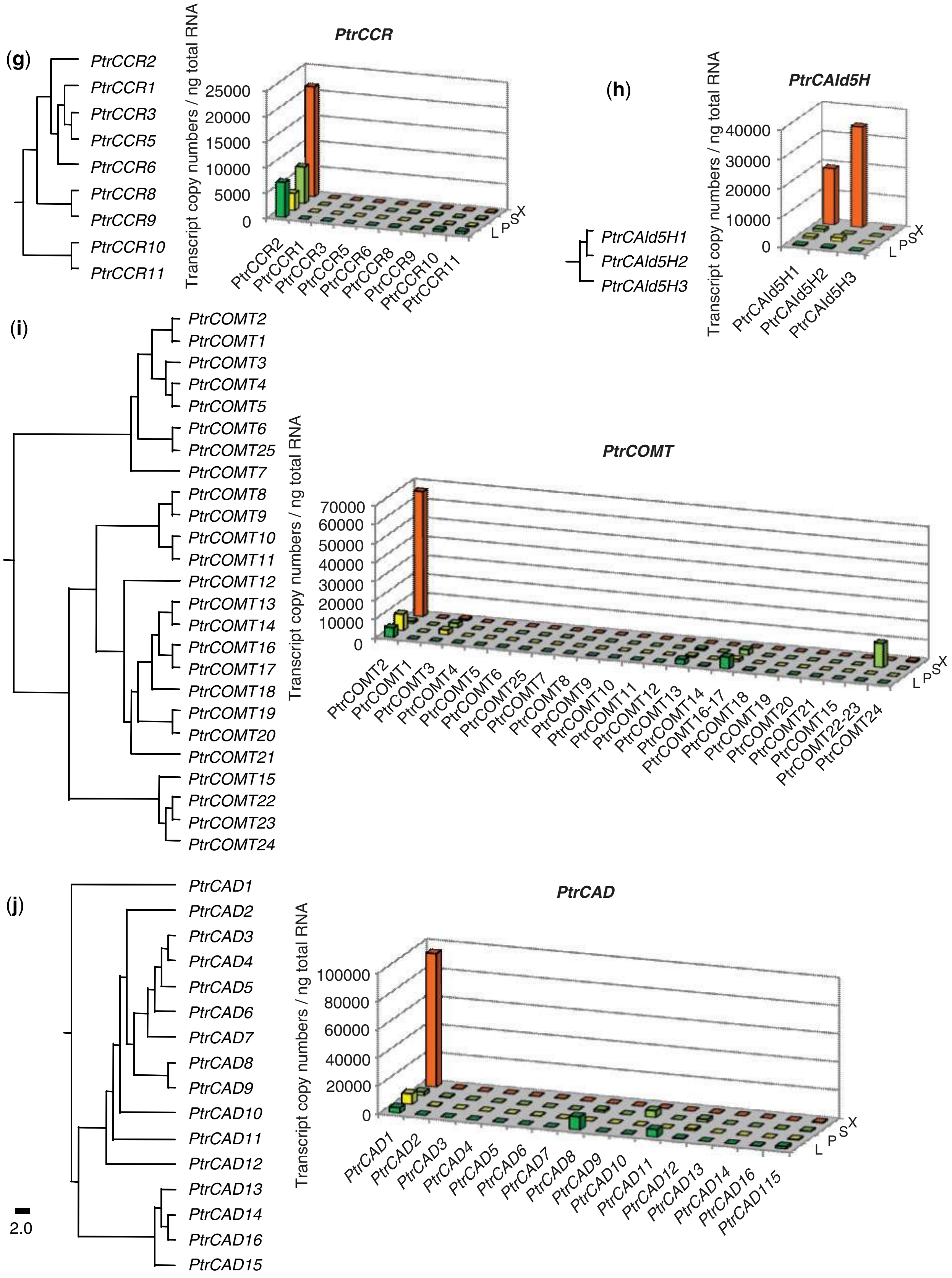


Fig. 2 Continued

(**Fig. 1**; Higuchi 1997, Lee et al. 1997, Hu et al. 1998), suggesting diverse functions for potential isozymes. Down-regulation of *4CL1* in *P. tremuloides* dramatically reduced lignin content in stem wood (Hu et al. 1999, Li et al. 2003), indicating that the level of enzyme activity and abundance of monolignol precursors affect the quantity of lignin.

In the genome of *P. trichocarpa*, 17 *4CL* gene models were identified (**Table 1**). Only two, *Ptr4CL3* and *Ptr4CL5*, had abundant transcripts in differentiating xylem (**Fig. 2c**; **Supplementary Tables S2, S3**). We previously identified a single *4CL* gene (*4CL1*) associated with monolignol biosynthesis in *P. tremuloides* (Hu et al. 1998, Hu et al. 1999). In *P. trichocarpa*, the two xylem-specific *4CL* genes (*Ptr4CL3* and *Ptr4CL5*) are most similar in sequence to the *P. tremuloides* monolignol *4CL1*, suggesting that both of these *P. trichocarpa* genes are important in monolignol biosynthesis. While additional *Ptr4CL* genes show some specificity for differentiating xylem (e.g. *Ptr4CL1* and *Ptr4CL17*), their total transcript levels were much lower than those of *Ptr4CL3* and *Ptr4CL5*. Transcripts of a *Ptr4CL1* homolog in *P. trichocarpa*×*deltoides* were more abundant in old leaves than in xylem (Allina et al. 1998) (**Supplementary Table S3**), indicating a different function.

Transcript levels of three *P. trichocarpa 4CL* genes (*Ptr4CL4, 7* and *11*) were preferentially abundant in tissues other than xylem. Transcripts of *Ptr4CL4*, which is most similar in sequence to the previously characterized *P. tremuloides 4CL2*, were highly abundant in shoots and leaves (Hu et al. 1998). *Ptr4CL4* may be involved in the biosynthesis of flavonoids and other soluble phenolics, as Hu et al. (1998) suggested for *P. tremuloides 4CL2*. The other two *P. trichocarpa 4CL* genes, *Ptr4CL7* and *Ptr4CL11*, not previously characterized, were more phloem or leaf specific.

Gene-specific transcript abundance of the HCT family

HCT, an acyltransferase, catalyzes the transfer of *p*-coumaroyl-CoA and caffeoyl-CoA to shikimate and quinate for the biosynthesis of corresponding shikimate and quinate esters (**Fig. 1**; Hoffmann et al. 2003). The reverse reaction for the formation of caffeoyl-CoA from caffeoyl shikimate or caffeoyl quinate esters is also catalyzed by HCT (**Fig. 1**). These reactions may be the entry point for the biosynthesis of methoxylated phenylpropanoids. Down-regulation of *HCT* in tobacco and alfalfa reduces lignin content and increases *p*-hydroxyphenyl units in lignin (Hoffmann et al. 2004, Shadle et al. 2007). *HCT* down-regulation in Arabidopsis redirects some phenylpropanoid precursors to flavonoids (Besseau et al. 2007).

Of the seven *HCT* genes in the *P. trichocarpa* genome (**Table 1**), *PtrHCT1* and *PtrHCT6* are more xylem specific (**Fig. 2d**; **Supplementary Tables S2, 3**) and most similar in sequence to those tobacco and alfalfa *HCT* genes involved in monolignol biosynthesis (Hoffmann et al. 2004, Shadle et al. 2007). *PtrHCT1* and *PtrHCT6* share 92.8% identity in protein sequences (**Supplementary Table S4**); however, the transcript level of *PtrHCT1* in differentiating xylem was 22 times that of *PtrHCT6*. Transcripts of a homolog of *P. trichocarpa PtrHCT6* were abundant in *P. fremontii*×*angustifolia* xylem (Tsai et al. 2006), but most probably they are encoded by the ortholog of *PtrHCT1*, the predominant *HCT* gene in differentiating xylem (**Fig. 2d**). Hamberger et al. (2007) identified seven *HCT* and eight *HCT-like* homologs. Only one, *HCT1*, the most similar in sequence to *P. trichocarpa PtrHCT1*, was abundant in differentiating xylem (**Supplementary Table S3**). Transcripts of *PtrHCT3* were more abundant in leaves. The transcript levels of the remaining *P. trichocarpa HCT* genes were insignificant in all tissues examined.

Gene-specific transcript abundance of the C3H family

C3H (CYP98A3) catalyzes the conversion of shikimate and quinate esters of *p*-coumarate into the corresponding caffeic acid conjugates (**Fig. 1**; Schoch et al. 2001). Reduction of *C3H* gene expression may therefore increase *p*-coumaryl (H) subunits in lignin (Ralph et al. 2006). A mutant with reduced C3H activity in Arabidopsis (*ref8*) is dwarfed, has reduced lignin content and is enriched in H subunits (Franke et al. 2002), a minor component in the lignin of wild-type plants. Reduction of *C3H* transcript levels led to reduced lignin content in *P. grandidentata*×*alba* (Coleman et al. 2008).

In *P. trichocarpa*, we identified four gene models for C3H (**Table 1**). Only one, *PtrC3H3*, was abundantly expressed in differentiating xylem (**Fig. 2e**; **Supplementary Tables S2, S3**). *Populus* EST analyses also revealed only one xylem abundant *C3H* gene (Hamberger et al. 2007) identical to *PtrC3H3*. The other *P. trichocarpa C3H* genes had low or insignificant transcript levels in all examined tissues.

Gene-specific transcript abundance of the CCoAOMT family

CCoAOMT catalyzes the methylation of caffeoyl-CoA to feruloyl-CoA as an early methyltransferase leading to the biosynthesis of coniferyl alcohol and sinapyl alcohol (**Fig. 1**; Ye et al. 1994, Li et al. 1999, Ye et al. 2001). Down-regulation of CCoAOMT activity in *P. tremula*×*alba* reduced lignin content and increased the S/G subunit ratio (Meyermans et al. 2000, Zhong et al. 2000).

We identified six *CCoAOMT* genes in *P. trichocarpa* (**Table 1**). *PtrCCoAOMT1* and *PtrCCoAOMT2* share 94.7% identity in protein sequence. Transcripts of these genes were both abundant and specific for differentiating xylem. All monolignol-related *Populus CCoAOMT* genes previously characterized are most similar to *PtrCCoAOMT1* or *PtrCCoAOMT2* (**Supplementary Table S3**). *PtrCCoAOMT3* and *PtrCCoAOMT4* have 80.8% protein sequence identity (**Supplementary Table S4**) but showed distinct expression (**Fig. 2f**; **Supplementary Tables S2, S3**). The transcripts of *PtrCCoAOMT3* were abundant and specific in differentiating xylem, whereas *PtrCCoAOMT4*, 5 and 6 transcripts were at background levels in all tissues tested.

Gene-specific transcript abundance of the CCR family

CCR catalyzes the conversion of cinnamoyl-CoA esters to their corresponding cinnamaldehydes (**Fig. 1**; Higuchi 1997). Down-regulation of CCR activity in *P. tremula*×*alba* reduced lignin content and increased incorporation of monolignol precursors such as ferulic acid into lignin (Leple et al. 2007).

We identified nine full-gene models for CCR (**Table 1**), whereas Tuskan et al. (2006) identified seven including two truncated genes (*PoptrCCR4* and *PoptrCCR7*, **Supplementary Table S3**), which were excluded in our study. Four *CCR* genes, *PtrCCR8–PtrCCR11*, identified in this study were not included in Tuskan et al. (2006). Hamberger et al. (2007) identified 13 *CCR* genes (**Supplementary Table S3**), of which *PoptrCCR4*, a truncated gene, and *PoptrCCRL1–PoptrCCRL5*, homologs of two Arabidopsis putative *CAD* genes (At5g19440 and At1g51410), were excluded in our study. The remaining seven are in our list of nine, which includes *PtrCCR9* and *PtrCCR10* that were not listed in Hamberger et al. (2007). Based on a *P. tremuloides CCR* cDNA sequence, Li et al. (2005) identified eight putative *CCR* loci in the *P. trichocarpa* genome. One locus, CCR-H1, contains the gene model for the *PtrCCR2* identified in this study, while the remaining loci contain truncated *CCR* genes. Of the nine *CCR* genes described in this study, *PtrCCR2* was the only one that had significant transcript levels in the tissues tested, and the level was highest in differentiating xylem and significantly less in the other tissues (**Fig. 2g**; **Supplementary Tables S2, S3**). All previous studies of *Populus* species reported expression of only one monolignol *CCR* gene related to *PtrCCR2* (**Supplementary Table S3**).

Gene-specific transcript abundance of the CAld5H/F5H family

CAld5H (CYP84), also known as F5H, catalyzes the hydroxylation primarily of coniferaldehyde, leading to the biosynthesis of sinapyl alcohol, the syringyl monolignol (**Fig. 1**; Humphreys et al. 1999, Osakabe et al. 1999). *CAld5H* from sweetgum (*Liquidambar styraciflua*) (Osakabe et al. 1999) and *F5H* from Arabidopsis (Humphreys et al. 1999) have been cloned and characterized. Overexpression of *CAld5H* in *P. tremuloides* (Li et al. 2003) and *F5H* in *P. tremula*×*alba* (Franke et al. 2000) increases the lignin S/G ratio.

We identified three gene models for CAld5H in *P. trichocarpa* (**Table 1**). Transcripts from two of these genes, *PtrCAld5H1* and *PtrCAld5H2* (91.4% protein sequence identity; **Supplementary Table S4**), were abundant and specific for differentiating xylem (**Fig. 2h**; **Supplementary Tables S2, S3**). *PtrCAld5H3* had little or no transcript in any tested tissues. While both *PtrCAld5H1* and *PtrCAld5H2* are likely to be the major 5-hydroxylase genes involved in monolignol biosynthesis during wood formation (**Supplementary Table 3**), their genetic and biochemical functions have not been verified.

Gene-specific transcript abundance of the COMT family

COMT was thought to catalyze the methylation of caffeate and 5-hydroxyferulate for the biosynthesis of monolignols (Higuchi 1997). Enzyme inhibition kinetics using recombinant COMT protein and different substrates for 10 angiosperm tree species showed that, instead, COMT catalyzed methylation of 5-hydroxyconiferyl aldehyde, which is the predominant path for the biosynthesis of the syringyl monolignol (**Fig. 1**) (Li et al. 2000). Reduced COMT activity in transgenic *Populus* species (Van Doorsselaere et al. 1995, Tsai et al. 1998, Jouanin et al. 2000, Ralph et al. 2001) reduced the lignin S/G ratio, and increased the incorporation of 5-hydroxyconiferyl alcohol into lignin.

We identified 25 gene models containing *COMT* coding sequences in *P. trichocarpa* (**Table 1**). Only *PtrCOMT2* transcripts were abundant and specific in differentiating xylem (**Fig. 2i**; **Supplementary Tables S2, S3**). Both *P. trichocarpa PtrCOMT2* and *PtrCOMT1* were identified by Hamberger et al. (2007) based on wood formation ESTs. Based on qRT-PCR, most *PtrCOMT*s had no significant expression in any tissue tested, including *PtrCOMT1*, which is 90.4% identical to *PtrCOMT2* in protein sequence (**Supplementary Table S4**). All the previously characterized *Populus COMT* genes implicated in monolignol biosynthesis in wood formation, based on substrate specificity, location or genetic regulation, are 98–99% identical in protein sequence to *PtrCOMT2* (**Supplementary Table S3**). Some *COMT* gene models encode identical proteins (*PtrCOMT16/17* and *PtrCOMT22/23*; **Supplementary Table S4**). Transcripts for *PtrCOMT16/17* were found with little or no expression in xylem, but were relatively abundant in leaf tissues. *PtrCOMT22/23* transcripts were highly shoot tip specific.

Gene-specific transcript abundance of the CAD family

CAD catalyzes the reduction of hydroxycinnamyl aldehydes to the corresponding monolignols (**Fig. 1**; Higuchi 1997). Down-regulation of CAD activity in *P. tremula*×*alba* had essentially no effect on either lignin content or the S/G ratio, but enhanced the incorporation of hydroxycinnamyl aldehydes into lignin, as both end groups and cross-coupling moieties (Baucher et al. 1996, Lapierre et al. 1999, Kim et al. 2002).

We found 16 *CAD* gene models in *P. trichocarpa* (**Table 1**). *PtrCAD1* showed a high transcript level and was very specific for differentiating xylem (**Fig. 2j**; **Supplementary Tables S2, S3**). *PtrCAD1* is most similar in sequence (98% protein identity) to a *P. tremuloides* monolignol-specific *CAD* that also had high expression in developing xylem (Li et al. 2001). The formation of sinapyl alcohol from sinapaldehyde is also catalyzed by another dehydrogenase, sinapyl alcohol dehydrogenase (SAD) (Li et al. 2001, Bomati and Noel 2005) (**Fig. 1**). In *P. trichocarpa*, *PtrCAD2* is a *SAD* and had low expression in all tissues tested. Hamberger et al. (2007) also found this *SAD* as a wood formation EST (**Supplementary Table S3**).

Candidate gene family members for monolignol biosynthesis in P. trichocarpa

Next, we applied quantitative criteria to assign genes that are most likely to be monolignol biosynthetic genes based on transcript abundance and specificity. We first selected genes whose absolute transcript levels (copy numbers per pg of total RNA) in differentiating xylem are >30% of the sum of transcripts in the four examined tissues (**Supplementary Tables S2, S3**). These genes were further screened for absolute transcript levels in differentiating xylem that are at least 2-fold greater that in any other tested tissues (**Supplementary Tables S2, S3**). Eighteen genes (*PtrPAL2–PtrPAL5*, *PtrC4H1* and 2, *Ptr4CL3* and 5, *PtrHCT1*, *PtrC3H3*, *PtrCCoAOMT1*, 2 and 3, *PtrCCR2*, *PtrCAld5H1* and 2, *PtrCOMT2* and *PtrCAD1*) meet these criteria (**Fig. 2**). In addition to these 18 core genes, we analyzed five more genes because they may lead to a more comprehensive understanding of the pathway. We included *PtrPAL1* because of its high transcript abundance in all tissues. Three additional genes (*Ptr4CL17*, *PtrHCT6* and *PtrCOMT24*) were expressed at high specificity for differentiating xylem, although at low levels. One more gene, *PtrCAD2* (an ortholog of *P. tremuloides SAD*), is important for the formation of sinapyl alcohol (Li et al. 2001, Bomati and Noel 2005). We selected a total of 23 putative monolignol biosynthesis genes. The transcripts (summed) from all the selected genes would account for >90% of the total transcripts for each gene family in differentiating xylem (**Supplementary Tables S2, S3**). Using liquid chromatography–tandem mass spectrometry (LC-MS/MS) we recently analyzed the trypsin-digested protein extracts from stem differentiating xylem of *P. trichocarpa* and identified specific tryptic peptides corresponding to the annotated protein sequences of all 23 genes (unpublished results), establishing that all are, to some extent, translated. We cloned all 23 of these candidate genes as cDNAs (from differentiating xylem), containing the 5′- and 3′-untranslated regions (UTRs) and the full coding sequence (**Supplementary Table S3**).

Identification of xylem-specific promoter sequence motifs

Many monolignol gene family homologs (particularly the phylogenetically paired genes) exhibit similar abundance and specificity in differentiating xylem. Such homologs are present in six of the 10 protein families (**Fig. 2a–f** and **h**), suggesting functional redundancy. Dot matrix analyses (Maizel and Lenk 1981) indicated that the overall promoter sequences of these paired homologs are quite divergent. For example, the pair of promoters for the genes *PtrC4H1* and *PtrC4H2*, whose protein sequences are 96.6% identical (**Supplementary Table S4**), show very low DNA sequence similarity (**Fig. 3**). Dot matrix plots for the other paired monolignol genes are shown in **Supplementary Fig. S2**.

Because the regulation of gene expression and transcript abundance is associated with the interaction of transcription factors and regulatory motifs, we searched for promoter sequence motifs for xylem specificity and abundance. We used DME-X (Smith et al. 2005) to search for 10 nt long motifs over-represented and abundant in differentiating xylem from the promoters of the 95 full-gene models (Materials and Methods). The $\log_2$-transformed transcript abundance within xylem (X) and the transcript abundance ratios of xylem to leaf (X/L), xylem to phloem (X/P) and xylem to shoot (X/S) were used as weighted parameters (**Supplementary Table S2**). Thirty motifs were identified for each weighted parameter (120 total) (**Supplementary Table S5**). We expect that motifs in the 5′-UTRs would be identified because the sequences searched included 2,000 bp upstream of the translation start site. A similar analysis of the 2,000 bp downstream of the translation stop sites did not yield any significant motifs over-represented for xylem expression. Analysis after randomization of the promoter sequences based on the Fisher–Yates shuffle (Fisher and Yates 1948) did not provide meaningful motifs.

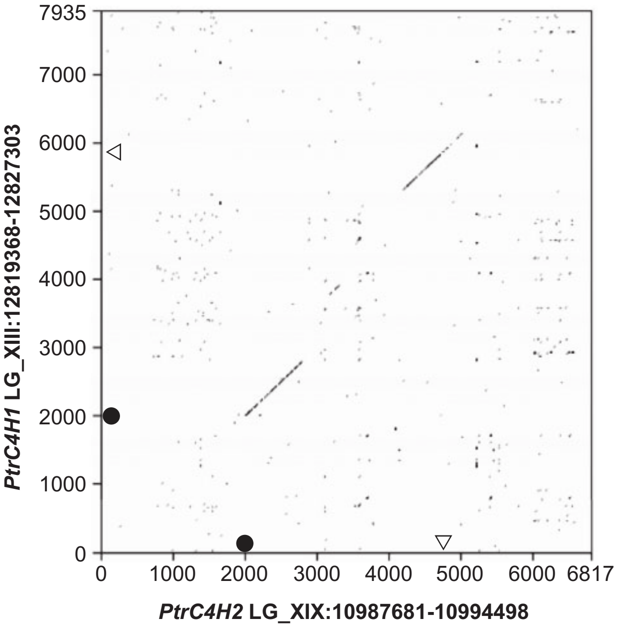


Fig. 3 Dot matrix sequence plot for *PtrC4H1* and *PtrC4H2* genes. The plot was created using the gene sequences from 2,000 bp upstream of the start codon (filled circle) to 2,000 bp downstream of the stop codon (open triangle). The program 'dottup' in the EMBOSS package (Rice et al. 2000) with word size of 10 nt was used.

Some combinations of promoter motifs discriminate transcript specificity and abundance in differentiating xylem

To identify how well specific motifs can discriminate the 18 core xylem-specific genes from the remaining 77 genes, we compiled the frequencies of the 120 motifs and performed stepwise discriminant analysis (PROC STEPDISC SAS). The analysis identified a combination of a minimum of five motifs, X/L-15, X/L-13, X/L-12, X/P-30 and X-3, that was sufficient to classify all 95 genes correctly into 18 xylem-specific and 77 non-xylem-specific groups.

Many of the 120 motifs resemble known *cis*-regulatory elements previously identified in lignin genes or vascular gene expression, such as AC elements (Raes et al. 2003, Rogers and Campbell 2004). We used STAMP (http://www.benoslab.pitt.edu/stamp/; Mahony and Benos 2007) to compare the five core motifs with known *cis*-elements. The *cis*-elements that best match these motifs were all found in PLACE (http://www.dna.affrc.go.jp/PLACE/; Higo et al. 1999) (**Fig. 4**). X/L-13 and X-3 resemble ACIIPVPAL2, the ACII element found in the promoters of PAL genes of many plant species (Hatton et al. 1995, Patzlaff et al. 2003, Gomez-Maldonado et al. 2004). X/L-15 resembles MYBPZM, the maize p element (Grotewold et al. 1994). X/L-12 resembles L4DCPAL1, a UVB-responsive element found in the promoter of the PAL1 gene of carrot (Takeda et al. 2002). X/P-30 best matches the SITEIIATCYTC element, an element in the promoter of cytochrome *c* genes involved in phosphorylation recognized by a TCP-domain transcription factor (Welchen and Gonzalez 2005, Welchen and Gonzalez 2006). The ability of the search to find known promoter elements of phenylpropanoid genes indicates a substantial robustness of the motif-finding algorithm.

Motifs are duplicated across gene promoters and preferentially located near the translation start sites

The 120 motifs are preferentially located within the proximal 600 bp upstream of the translation start site (**Fig. 5a**). The distribution bias is more obvious for the 18 core monolignol gene promoters than the non-xylem-specific promoters. The five core motifs are distributed across the 18 monolignol gene promoters, and each promoter has from four to eight copies of some, but not necessarily all, of the five motifs (**Table 2**). The total number of core motifs in the 18 monolignol genes is 91; therefore, the average number of core motifs in a promoter is five. The locations of the five core motifs in the xylem-specific gene promoters show stronger bias toward the translation start site (**Fig. 5b**), consistent with a functional role (Smith et al. 2006, Yamamoto et al. 2007).

Discriminating power of specific motifs

Any of the five motifs individually was able to classify correctly 9–12 of the 18 xylem-specific promoters and 64–76 of the 77 non-xylem-specific promoters (**Fig. 4**). Starting with one motif, each addition of a motif increased the number of correct assignments of genes into xylem- and non-xylem-specific groups. Any four of the five motifs were able to classify correctly from 16 to all of the 18 xylem-specific promoters and from 76 to all of the non-xylem-specific promoters. Each of the five individual motifs possesses a high degree of discriminating power and the power was increased in combination with other motifs.

To validate the discriminating power of the motifs, we carried out a leave-one-out cross-validation. Removing one promoter always resulted in groups of motifs that were similar in number of motifs (6.33±0.18, average±SE) and could correctly classify all promoters in their data set. DME-X analysis of the 95 cross-validation sets identified 607 motifs that discriminate xylem specificity (xylem from non-xylem). Of the 607 motifs, 319 (52.5%) were highly similar to the five core motifs, as denoted by the <1.0 Kullback–Leibler (KL) divergence score (Kullback and Leibler 1951, Schones et al. 2005). The 319 motifs included 233 that are identical (KL divergence = 0) to one of the five core motifs. The remaining 288 motifs had low to poor similarity with the

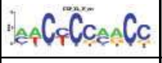
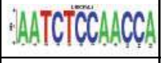
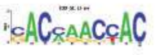
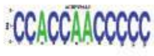
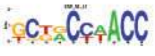
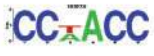
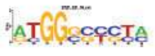
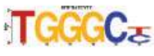
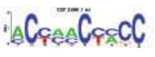
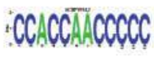
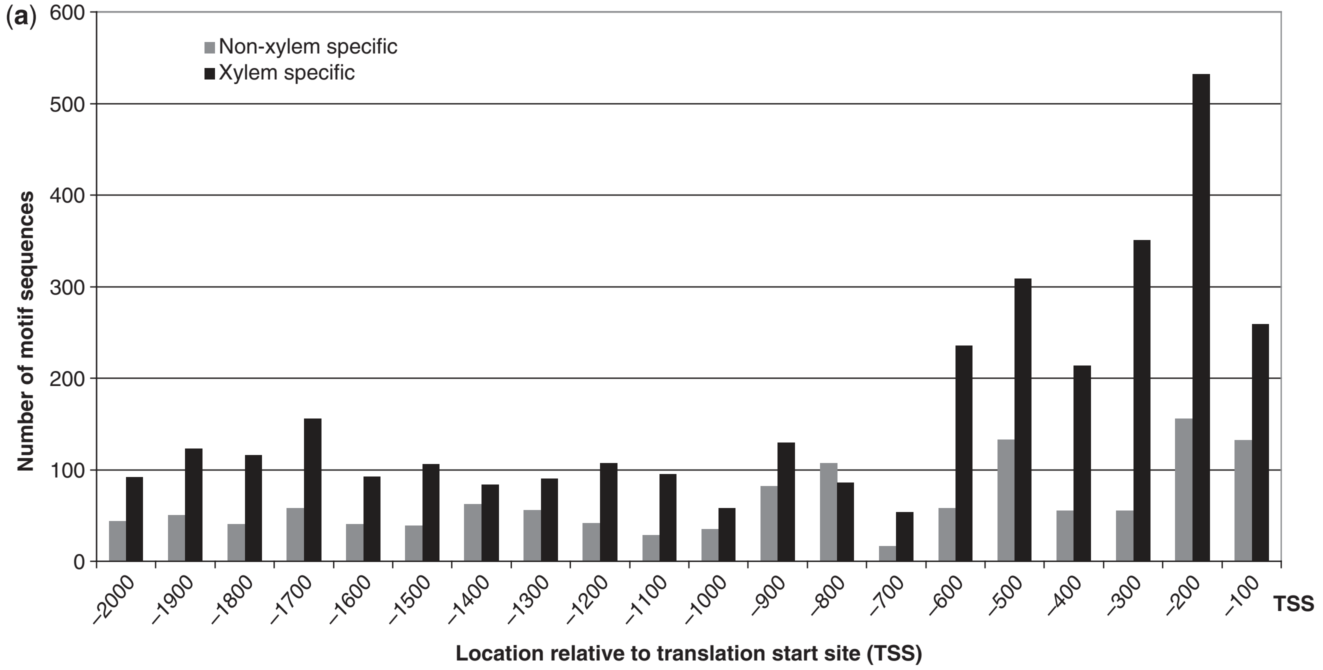
Five core motifs		Best match in PLACE database			Discriminating power of motif(s)													
Motif Name	Motif LOGO	Cis-element LOGO	Cis-element name	e-value	Motif(s) included in the discriminant analysis													
X/L-15*			L4DCPAL1	1.30E-08	√					√	√	√	√		√	√	√	√
X/L-13*			ACIIPVPAL2	1.19E-08		√				√	√	√	√	√		√	√	√
X/L-12			MYBPZM	4.85E-08			√				√	√	√	√	√		√	√
X/P-30*			SITEIIATCYTC	2.78E-06				√				√	√	√	√	√		√
X-3*			ACIIPVPAL2	7.21E-11					√				√	√	√	√	√	
False xylem					3	3	1	4	13	6	0	0	0	1	1	0	1	0
False non-xylem					6	6	8	9	6	0	3	1	0	2	0	1	2	1

Fig. 4 Identities and discriminating power of the five core motifs. Motif identities were determined by searching the most similar known plant *cis*-elements using STAMP. *The best *cis*-element match was a reverse complement. 'False xylem' or 'False non-xylem' refers to false positives in either tissue. The five core motifs were identified based on the stepwise discriminant analysis carried out using PROC STEPDISC in SAS, while the discriminating power and the false identification of promoters were analyzed using PROC DISCRIM in SAS.



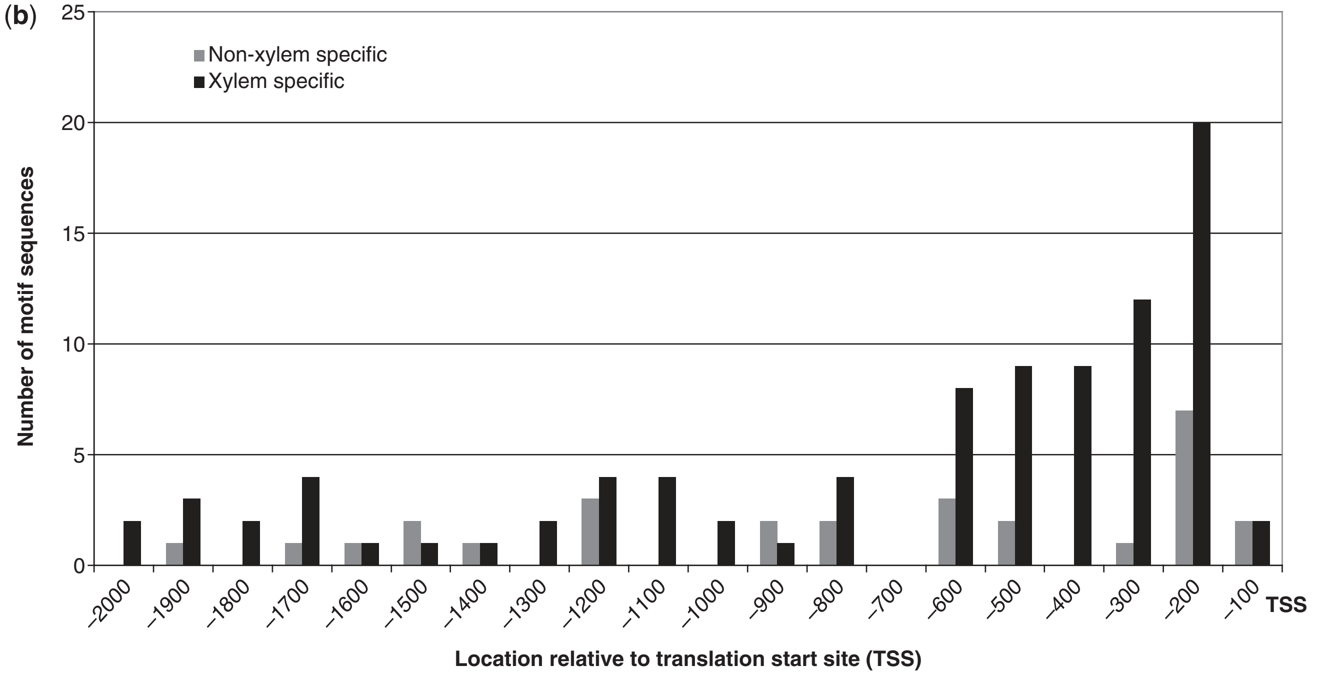


Fig. 5 Location of xylem-specific motifs in the promoters. The location of a motif was measured from its first nucleotide to the translation start site. The five core motifs were found in 38 (18 xylem specific and 20 non-xylem specific) of the 95 phenylpropanoid gene promoters. Distribution of (a) all 120 motifs in 74 of the 95 promoters and (b) the five core motifs in the 38 promoters.

five core motifs (KL divergence 1.0–4.65). Within the 288 motifs are two additional major groups, one resembling the PALBOXLPC element on the PAL promoter of parsley (Logemann et al. 1995) and the other resembling the CTRMCAMV35S element on the cauliflower mosaic virus (CaMV) 35S promoter (Pauli et al. 2004) in the PLACE database (Higo et al. 1999).

The power of discriminating tissue-specific genes varied depended on the stringency value set for the RSA tool (Thomas-Chollier et al. 2008) for scanning promoter motifs. Increasing the stringency value decreased the power for xylem genes and increased it for non-xylem genes (**Table 3**). The power is similar to that using inferred sequence modules to predict tissue specificity of genes in human (Smith et al. 2006).

Discussion

Genes associated with monolignol biosynthesis in wood formation

The sequence of the *Populus* genome makes possible a systematic and comprehensive study of the genes for lignin biosynthesis in wood formation. In Arabidopsis, wood is not readily obtained and correlation of gene expression with wood formation is limited. Arabidopsis has very little lignin, and the composition of that lignin is guaiacyl rich (Do et al. 2007, Nakatsubo et al. 2008), which is not typical of woody angiosperms, which have roughly equal amounts of G and S subunits. Many other phenylpropanoid pathways share monolignol biosynthetic genes. High levels of expression may be observed for members of gene families that are not related to wood formation, such as in the defense response (Tsai et al. 2006). Hamberger et al. (2007) carried out an in silico analysis of all of the *Populus* ESTs and, aided by microarray data, identified genes involved in monolignol biosynthesis during wood formation. An EST approach and results from different *Populus* species under different conditions may not be consistent or comprehensive for quantitation and comparison of gene expression. EST information is typically limited by incomplete depth and breadth of sequencing, or by normalization (if applied to library construction) that reduces expression information.

Table 2 The frequency of each of the five core motifs in the 18 monolignol gene promoters of *P. trichocarpa*

Genes	Core motifs					Sum of five core motifs
	X/L-15	X/L-13	X/L-12	X/P-30	X-3	
PtrPAL2	1	0	2	0	2	5
PtrPAL3	2	1	0	0	1	4
PtrPAL4	2	0	2	1	0	5
PtrPAL5	0	2	2	0	0	4
PtrC4H1	1	1	0	1	1	4
PtrC4H2	1	1	0	1	5	8
Ptr4CL3	2	2	0	0	2	6
Ptr4CL5	0	2	0	2	2	6
PtrHCT1	2	0	2	0	0	4
PtrC3H3	0	1	0	0	4	5
PtrCCoAOMT1	1	1	1	1	3	7
PtrCCoAOMT2	2	0	2	0	3	7
PtrCCoAOMT3	1	1	1	1	0	4
PtrCCR2	1	0	0	2	0	3
PtrCAld5H1	1	0	1	0	3	5
PtrCAld5H2	0	2	1	2	0	5
PtrCOMT2	0	1	1	0	2	4
PtrCAD1	0	1	0	3	1	5
Total	17	16	15	14	29	91
Average no. of motifs per gene	0.94	0.89	0.83	0.78	1.61	5.06

As a first step in a systematic analysis of monolignol gene expression, we analyzed 23 genes and their promoters to initiate the development of a comprehensive and predictive model of lignin biosynthesis. The functions of most of the 72 remaining phenylpropanoid genes (full-gene models) are not known. Many genes associated with monolignol biosynthesis have been characterized genetically and/or biochemically in one or more species (Vanholme et al. 2008). For seven of the 23 genes, genetic or biochemical verification of function in wood formation has been carried out for their homologs in other *Populus* species. These seven genes are *Ptr4CL3*, *PtrCCoAOMT1*, *PtrCCR2*, *PtrCAld5H1*, *PtrCOMT2*, *PtrCAD1* and *PtrCAD2*. The genetic and biochemical roles of the remaining 16 genes remain to be better defined. The availability of the cloned full-length transcript sequences of these genes (from this study) and of genetic transformation for *P. trichocarpa* (Song et al. 2006) now makes this possible.

Many of the 95 members of the phenylpropanoid gene set are not transcribed or are transcribed at low levels in the four tissues examined. Some gene families may have many genes with diverse functions in secondary metabolism in different cell types and tissues. In particular, *COMT* (25 genes), *4CL* (17) and *CAD* (16) represent 58 out of the 95 genes. Functions for many these genes have been proposed to be unrelated to monolignol biosynthesis based on the phylogenetic or expression analysis (Tsai et al. 2006, Souza et al. 2008, Barakat et al. 2009). Addition of methyl groups by OMT is widespread in primary and secondary metabolism. 4CL and CoA ligation are critical components for biosynthesis of flavonoids, anthocyanins and phytoalexins.

Table 3 Discriminating power determined by the leave-one-out cross-validation analysis

Motif scanning stringency	Correctly predicted from 18 xylem genes	Correctly predicted from 77 non-xylem genes	Overall correct rate (%)	Pearson's χ^2 *P*-value
4.0	17	23	52.1	0.0326
4.2	17	35	54.7	0.0017
4.5	15	46	64.2	0.0010
4.7	11	58	72.6	0.0028
5.0	8	72	84.2	0.0001

CAD is needed for lignan and norlignan biosynthesis, which occurs in non-woody tissues (Suzuki and Umezawa 2007).

These low level transcript genes could be abundantly expressed in rare cell types or tissues we did not examine. Alternatively, they could be expressed under specific conditions (e.g. abiotic or biotic stress responses). Transcripts that are highly unstable would not have been detected. The absence of expression for many genes in multigene families has been observed at the genome level in Arabidopsis, where 30% of the gene models have not shown evidence of transcription (Yamada et al. 2003).

Harmer et al. (2000), using an 8,000 gene microarray, found circadian variation of transcript abundance in about 480 genes in whole plants of Arabidopsis. A circadian response was found for transcripts of four monolignol genes (as annotated by Raes et al. 2004) peaking about 6 h before subjective dawn. Rogers et al. (2005), using Northern blot analysis also with whole plants of Arabidopsis, found an 'idealized, consensus trend' of diurnal cycle variation for 11 monolignol genes with a peak 4 h after dawn and a second peak at 20 h after dawn. Examination of the blots presented by Rogers et al. (2005) indicates substantial deviation from this 'trend'. Our qRT-PCR results for five monolignol genes (**Supplementary Fig. S3**) show transcript variation over a 24 h period, with a peak at 2 a.m., which declined gradually during the day. Our results are more consistent with the circadian pattern observed by Harmer et al. (2000), than those of Rogers et al. (2005). Diurnal variation is a significant issue for comparative analysis of transcript abundance and specificity. Similarly, seasonal variation (including effects of photoperiod and temperature), abiotic and biotic stresses could affect the relative transcript abundance of monolignol genes. Our results (**Supplementary Fig. S3**) indicate that sampling at a consistent time around 10 a.m. provides a reasonable estimate of transcript level and tissue specificity for monolignol genes in *P. trichocarpa*. For a comprehensive systems approach, the protein quantity and enzyme activity need to be determined and compared with the variation and tissue specificity of transcript levels for all the monolignol genes.

Functional redundancy in monolignol biosynthesis

Our genome-wide identification of *P. trichocarpa* monolignol genes and quantitation of the transcript abundance and specificity suggest substantial functional redundancy in monolignol biosynthesis. Members in six of the 10 monolignol enzyme families may have redundant functions. The extent of potential functional redundancy shows little relationship to the number of members of their gene families. Functional redundancy should be reflected in the response of plants to mutation or suppression in transgenics because redundancy would moderate the perturbations in gene expression.

All members of the *PAL* family (five members) are very similar in sequence (85–98% protein identity), and are abundantly transcribed in differentiating xylem. None of the *PAL* gene family members has been characterized for genetic function in tree species by knock-down or loss-of-function mutation. *PtrC4H1* and *PtrC4H2* (96.6% protein sequence identity) may be considered bona fide genes for monolignol biosynthesis based on biochemical properties of recombinant enzymes from their *P. tremuloides* orthologs (Lu et al. 2006). However, gene-specific knock-down experiments that would evaluate functional redundancy have not yet been done. The extent of redundancy for the paired *4CL* gene members, *Ptr4CL3* and *5*, with high xylem-specific expression also has not been resolved. Similarly, the *HCT* gene family has two highly similar predicted proteins (93% identity) expressed in differentiating xylem, *PtrHCT1* is highly abundant, while *PtrHCT6* is expressed at only 4.5% of that level.

For *CCoAOMT*, the most similar gene pair (*PtrCCoAOMT1* and *PtrCCoAOMT2*; 95% protein identity) has a 2-fold difference in differentiating xylem transcript abundance. *PtrCCoAOMT3* and *PtrCCoAOMT4* are also similar in protein sequence (81%), but have different tissue specificities. Only the ortholog of *PtrCCoAOMT1* has been shown genetically (Meyermans et al. 2000, Zhong et al. 2000) and biochemically (Meng and Campbell, 1998) to function in monolignol biosynthesis in *Populus*. *PtrCCoAOMT1*, *2* and *3* genes show both redundancy and specificity; therefore, they may all be involved in monolignol biosynthesis. *PtrCAld5H1* and *PtrCAld5H2*, sharing 91% protein identity, are abundantly expressed in differentiating xylem. Only the ortholog of one of these two *CAld5H* genes has been genetically (overexpression) and biochemically (recombinant protein) characterized in *Populus* (Osakabe et al. 1999, Franke et al. 2000, Li et al. 2003).

DNA motifs in monolignol pathway gene promoters and gene expression specificity

Many studies have addressed the role of specific promoters in regulation of gene expression in monolignol biosynthesis or in lignifying tissues (Higuchi, 1997, Rogers and Campbell 2004). These studies most often focus on identification of *cis*-regulatory elements by deletion analysis or by searching for previously identified elements, such as the AC element and the related P-box or L-box sequences. Using microarray hybridization and EST frequency information, Ko et al. (2006) associated the AC element and a novel ACAAAGAA motif with induced secondary xylem development in Arabidopsis. The ACAAAGAA motif was not found among the 120 motifs identified in our analysis. Promoter sequences of *CesA* (cellulose synthase A) genes in *Eucalyptus*, *Populus* and Arabidopsis have recently been analyzed by Creux et al. (2008). They identified 11 consensus promoter motifs for secondary cell wall *CesA* genes. However, the sequences of these 11 motifs and the sequences of our five core motifs are quite divergent, as indicated by the high KL divergence scores (Schones et al. 2005) ranging from 2.01 to 3.63 (scores <1 signify high similarity).

The core motifs appear to be redundant and quantitative, with each additional motif contributing to greater abundance and specificity. This may be tested by construction of synthetic promoters with varying numbers of specific motifs. We suggest that core motifs could be used to survey the *Populus* genome to predict quantitative expression and specificity from promoter sequences. The similarity and multiplicity of motifs suggest a parallel multiplicity and specificity of transcription factors regulating gene expression in monolignol biosynthesis.

Materials and Methods

Computational identification of P. trichocarpa phenylpropanoid genes

Phenylpropanoid biosynthesis genes were identified from the 45,555 gene model set of *P. trichocarpa* genome annotation v1.1 (http://genome.jgi-psf.org/Poptr1_1/) by reciprocal BLAST (Altschul et al. 1990) using protein sequences of the 63 Arabidopsis phenylpropanoid genes (http://cellwall.genomics.purdue.edu/). *Populus trichocarpa* gene models homologous to the 63 Arabidopsis phenylpropanoid genes were first searched based on the protein sequences using the program BLASTP with the e-value cut-off set at 1-E03. The resulting protein sequences of the positive *P. trichocarpa* gene model hits were then blasted against all 31,921 proteins in Arabidopsis (TAIR7, http://www.arabidopsis.org/). Only 169 *P. trichocarpa* gene models that had one of the 63 Arabidopsis phenylpropanoid genes as the top hit were retained for analysis. Truncated gene models were removed by examining the gene model for ORFs and by multiple protein sequence alignment using ClustalW (Thompson et al. 1994). The phylogenetic relationships of the gene models were evaluated using the Clustal alignments of the protein sequences and the neighbor–joining trees with 1,000 bootstrap trials. Phylogenetic trees were plotted using FigTree (http://tree.bio.ed.ac.uk/software/figtree/). Protein sequence identity matrices were analyzed by pair-wise alignment using JAligner (http://jaligner.sourceforge.net) with the default BLOSUM62 scoring matrix. JAligner implements the Smith–Waterman algorithm with Gotoh's improvement for biological local pair-wise sequence alignments. Sequence dot plots of the gene sequences from 2,000 bp upstream of the start codon to 2,000 bp downstream of the stop codon were created using the program 'dottup' in EMBOSS (Rice et al. 2000).

RNA extraction, primer design and qRT-PCR

Total RNA was extracted from fully expanded leaf (without the midvein), stem developing phloem, shoot tip (~1.5 cm from the top) and stem developing xylem of 1-year-old greenhouse-grown *P. trichocarpa* (Nisqually-1). Tissues were collected at about 10 a.m., approximately 4 h after dawn, from three trees, which were pooled before RNA extraction for qRT-PCR. Sampling of tissues over a 24 h period (**Supplementary Fig. S3**) indicates that tissue specificity is maintained throughout the 24 h period, and that sampling at 10 a.m. provides a reasonable estimate of relative transcript abundance. Three qRT-PCRs were carried out for each pooled sample. Total RNA extraction/purification, first-strand cDNA synthesis and qRT-PCR primer designs (**Supplementary Table S2**) were conducted as described (Lu et al. 2005, Lu et al. 2006, Suzuki et al. 2006). Several primer sets for nearly identical transcripts were designed by including at least one mismatch at the primer 3′ terminus to discriminate between homologs (Shi and Chiang 2005). For each targeted gene, 2–4 sets of primers were prepared, and the specificity of all primers was checked by BLAST with all annotated transcript sequences in the *P. trichocarpa* genome.

Real-time PCRs (Applied Biosystems 7900HT) were conducted as in previous work (Suzuki et al. 2006). Verification of amplification specificity was based on dissociation curve analysis (Shi and Chiang 2005) and sequences of the PCR products. Each reaction was repeated at least three times. Purified PCR products were used as standards for establishing a quantitative correlation between the transcript copy numbers and the C_T values (Suzuki et al. 2006).

To evaluate the effects of tissue pooling on our results, we used a simple analysis of variance (ANOVA) model, to analyze the variation in our diurnal experiment to estimate the relative effects of biological and technical replication. Biological variation accounts for 6.19% of the total variance while technical variation accounts for 0.23% (**Supplementary Table S7**). Pooling should average the biological variation. By pooling these samples and measuring the pools directly (**Supplementary Fig. S3**), we show that the transcript levels of the pools are not significantly different (*P*-value = 0.92) from the calculated average of the individual biological replicates.

Cloning of the whole transcript sequences for the identified xylem-specific monoliginol genes

The cDNA containing the full-length coding region for each putative monolignol gene was amplified by PCR using primer sets (**Supplementary Table S6**) based on the predicted sequences. The 5′- and 3′-UTR sequences were cloned by 5′- and 3′-RACE (rapid amplification of cDNA ends; RLM-RACE kit, Ambion, Austin, TX, USA) (primers in **Supplementary Table S2** were used). The whole transcript sequences were assembled from the sequences of the coding region and 5′- and 3′-RACE products and deposited in GenBank (**Supplementary Table S3**).

Xylem-specific promoter identification and motif discriminant analysis

The promoter regions, for each of the 95 phenylpropanoid genes, were defined as the 2,000 nt upstream of the translation start site based on the v1.1 annotation of the *P. trichocarpa* genome or the 5′-RACE results through cDNA cloning. Xylem-specific motifs of these promoters were inferred using the program DME-X (Smith et al. 2005) with the $\log_2$-transformed transcript abundance (copy number per ng of total RNA) within xylem and abundance ratios of xylem over phloem, xylem over shoot and xylem over leaf as weighted parameters. We have searched for 10 nt long motifs as was done by Smith et al. (2005), who evaluated different motif lengths of eight, 10 and 12 and found that while 10 is significantly better than eight, 12 is not significantly better than 10, and 12 is far more computationally intensive. For each of the four weighted parameters, we ranked potential motifs using the DME-X significance score. We examined up to 100 motifs selected by DME-X. Most of the motifs beyond the first 30 of the ranked motifs had low significance scores, when compared with motifs generated from random sequences, made by a shuffling algorithm from the 95 promoters and choosing a cut-off value that excluded random effects. The frequencies of the motifs in each promoter were compiled for stepwise discriminate analysis (PROC STEPDISC SAS, Cary, NC, USA) to identify the minimum number of motifs that could correctly classify the promoters into xylem and non-xylem groups. PROC DISCRIM (SAS) was used to analyze the discriminating power of individual and combinations of the motifs. 'WebLogo' (Crooks et al. 2004) was used to generate the motif logos.

Leave-one-out cross-validation analysis was conducted to evaluate the motifs identified by DME-X. A total of 95 cross-validation promoter sets were prepared so that each of the phenylpropanoid gene promoters was removed once from the cross-validation promoter set. Each cross-validation promoter set was analyzed in the same way as the complete 95 promoter sets to identify minimum numbers of motifs that discriminate the 94 promoters correctly. These minimum numbers of motifs were then scanned, using the RSA tool (Thomas-Chollier et al. 2008), on the removed promoter of each cross-validation set for discriminating power evaluated by PROC DISCRIM. Motif similarities were calculated for KL divergence using MatCompare (Schones et al. 2005).

Supplementary data

Supplementary data are available at PCP online.

Funding

This work was supported by the Forest Biotechnology Industrial Research Consortium at North Carolina State University; The National Science Foundation, Plant Genome Research Program [DBI-0922391].

Acknowledgements

The authors would like to thank Judy Jakobek and Ting-Feng Yeh for their aid in the preparation of the manuscript.

References

Albert, R. (2007) Network inference, analysis, and modeling in systems biology. *Plant Cell* 19: 3327–3338.

Allina, S.M., Pri-Hadash, A., Theilmann, D.A., Ellis, B.E. and Douglas, C.J. (1998) 4-coumarate:coenzyme A ligase in hybrid poplar—properties of native enzymes, cDNA cloning, and analysis of recombinant enzymes. *Plant Physiol.* 116: 743–754.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* 215: 403–410.

Bailey, T.L., Williams, N., Misleh, C. and Li, W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* 34: W369–W373.

Barakat, A., Bagniewska-Zadworna, A., Choi, A., Plakkat, U., DiLoreto, D.S., Yellanki, P. and Carlson, J.E. (2009) The cinnamyl alcohol dehydrogenase gene family in *Populus*: phylogeny, organization, and expression. *BMC Plant Biol.* 6: 9–26.

Baucher, M., Chabbert, B., Pilate, G., Van Doorsselaere, J., Tollier, M.T., Petit-Conil, M., et al. (1996) Red xylem and higher lignin extractability by down-regulating a cinnamyl alcohol dehydrogenase in poplar. *Plant Physiol.* 112: 1479–1490.

Besseau, S., Hoffmann, L., Geoffroy, P., Lapierre, C., Pollet, B. and Legrand, M. (2007) Flavonoid accumulation in Arabidopsis repressed in lignin synthesis affects auxin transport and plant growth. *Plant Cell* 19: 148–162.

Bomati, E.K. and Noel, J.P. (2005) Structural and kinetic basis for substrate selectivity in *Populus tremuloides* sinapyl alcohol dehydrogenase. *Plant Cell* 17: 1598–1611.

Brown, S.A. and Neish, A.C. (1955) Shikimic acid as a precursor in lignin biosynthesis. *Nature* 175: 948–962.

Chen, F. and Dixon, R. (2007) Lignin modification improves fermentable sugar yields for biofuel production. *Nat. Biotechnol.* 25: 759–761.

Chiang, V. (2002) From rags to riches. *Nat. Biotechnol.* 20: 557–558.

Coleman, H.D., Samuels, A.L., Guy, R.D. and Mansfield, S.D. (2008) Perturbed lignification impacts tree growth in hybrid poplar—a function of sink strength, vascular integrity and photosynthetic assimilation. *Plant Physiol.* 148: 1229–1237.

Creux, N.M., Ranik, M., Berger, D.K. and Myburg, A.A. (2008) Comparative analysis of orthologous cellulose synthase promoters from Arabidopsis, *Populus* and *Eucalyptus*: evidence of conserved regulatory elements in angiosperms. *New Phytol.* 179: 722–737.

Crooks, G., Hon, G., Chandonia, J. and Brenner, S. (2004) WebLogo: a sequence logo generator. *Genome Res.* 14: 1188–1190.

D'haeseleer, P. (2006) How does DNA sequence motif discovery work? *Nat. Biotechnol.* 24: 959–961.

Dixon, R., Chen, F., Guo, D. and Parvathi, K. (2001) The biosynthesis of monolignols: a 'metabolic grid', or independent pathways to guaiacyl and syringyl units? *Phytochemistry* 57: 1069–1084.

Dixon, R., Achnine, L., Deavours, B.E. and Naoumkina, M. (2006) Metabolomics and gene identification in plant natural product pathways. *In* Biotechnology in Agriculture and Forestry, Volume 57: Plant Metabolomics. Edited by Saito, K., Dixon, R.A. and Willmitzer, L. pp. 243–259. Springer-Verlag, Berlin.

Do, C.T., Pollet, B., Thevenin, J., Sibout, R., Denoue, D., Barriere, Y., et al. (2007) Both caffeoyl coenzyme A 3-O-methyltransferase 1 and caffeic acid *O*-methyltransferase 1 are involved in redundant functions for lignin, flavonoids and sinapoyl malate biosynthesis in Arabidopsis. *Planta* 226: 1117–1129.

Fisher, R.A. and Yates, F. (1948) Statistical Tables for Biological, Agricultural and Medical Research, 3rd edn. Oliver & Boyd, London.

Franke, R., Hemm, M.R., Denault, J.W., Ruegger, M.O., Humphreys, J.M. and Chapple, C. (2002) Changes in secondary metabolism and deposition of an unusual lignin in the *ref8* mutant of Arabidopsis. *Plant J.* 30: 47–59.

Franke, R., McMichael, C.M., Meyer, K., Shirley, A.M., Cusumano, J.C. and Chapple, C. (2000) Modified lignin in tobacco and poplar plants over-expressing the Arabidopsis gene encoding ferulate 5-hydroxylase. *Plant J.* 22: 223–234.

Freudenberg, K. and Neish, A.C. (eds) (1968) Constitution and Biosynthesis of Lignin. pp. 78–122. Springer-Verlag, Berlin.

Gomez-Maldonado, J., Avila, C., Torre, F., Canas, R., Canovas, F.M. and Campbell, M.M. (2004) Functional interactions between a glutamine synthetase promoter and MYB proteins. *Plant J.* 39: 513–526.

Grotewold, E., Drummond, B.J., Bowen, B. and Peterson, T. (1994) The *myb*-homologous *P* gene controls phlobaphene pigmentation in maize floral organs by directly activating a flavonoid biosynthetic gene subset. *Cell* 76: 543–553.

Hahlbrock, K. and Scheel, D. (1989) Physiology and molecular biology of phenylpropanoid metabolism. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 40: 347–369.

Hamberger, B., Ellis, M., Friedmann, M., Souza, C.D.A., Barbazuk, B. and Douglas, C.J. (2007) Genome-wide analyses of phenylpropanoid-related genes in *Populus trichocarpa*, *Arabidopsis thaliana*, and *Oryza sativa*: the *Populus* lignin toolbox and conservation and diversification of angiosperm gene families. *Can. J. Bot.* 85: 1182–1201.

Harkin, J.M. (1967) Lignin—a natural polymeric product of phenol oxidation. *In* Oxidative Coupling of Phenols. Edited by Taylor, W.I. and Battersby, A.R. pp. 243–321. Marcel Dekker, New York.

Harmer, S.L., Hogenesch, J.B., Straume, M., Chang H., Han, B., Zhu, T., et al. (2000) Orchestrated transcription of key pathways in *Arabidopsis* by the circadian clock. *Science* 290: 2110–2113.

Hatton, D., Sablowski, R., Yung, M.H., Smith, C., Schuch, W. andc Bevan, M. (1995) Two classes of cis sequences contribute to tissue-specific expression of a PAL2 promoter in transgenic tobacco. *Plant J.* 7: 859–876.

Higo, K., Ugawa, Y., Iwamoto, M. and Korenaga, T. (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.* 27: 297–300.

Higuchi, T. (1997) Biochemistry and Molecular Biology of Wood. pp. 131–181. Springer, New York.

Higuchi, T. (2003) Pathways for monolignol biosynthesis via metabolic grids: coniferyl aldehyde 5-hydroxylase, a possible key enzyme in angiosperm syringyl lignin biosynthesis. *Proc. Jpn. Acad., Ser. B, Phys. Biol. Sci.* 79: 227–236.

Higuchi, T. and Brown, S.A. (1963) Studies of lignin biosynthesis using isotopic carbon. XII. The biosynthesis and metabolism of sinapic acid. *Can. J. Biochem. Physiol.* 41: 613–620.

Hoffmann, L., Besseau, S., Geoffroy, P., Ritzenthaler, C., Meyer, D., Lapierre, C., et al. (2004) Silencing of hydroxycinnamoy-coenzyme A shikimate/quinate hydroxycinnamoyltransferase affects phenylpropanoid biosynthesis. *Plant Cell* 16: 1446–1465.

Hoffmann, L., Maury, S., Martz, F., Geoffroy, P. and Legrand, M. (2003) Purification, cloning, and properties of an acyltransferase controlling shikimate and quinate ester intermediates in phenylpropanoid metabolism. *J. Biol. Chem.* 278: 95–103.

Hu, W., Harding, S.A., Lung, J., Popko, J.L., Ralph, J., Stokke, D.D., et al. (1999) Repression of lignin biosynthesis promotes cellulose accumulation and growth in transgenic trees. *Nat. Biotechol.* 17: 808–812.

Hu, W., Kawaoka, A., Tsai, C.-J., Lung, J., Osakabe, K., Ebinuma, H. and Chiang, V.L. (1998) Compartmentalized expression of two structurally and functionally distinct 4-coumarate:coenzyme A ligase (4CL) genes in Aspen (*Populus tremuloides*). *Proc. Natl Acad. Sci. USA* 95: 5407–5412.

Hudson, M.E. and Quail, P.H. (2003) Identification of promoter motifs involved in the network of phytochrome A-regulated gene expression by combined analysis of genomic sequence and microarray data. *Plant Physiol.* 133: 1605–1616.

Humphreys, J., Hemm, M. and Chapple, C. (1999) New routes for lignin biosynthesis defined by biochemical characterization of recombinant ferulate 5-hydroxylase, a multifunctional cytochrome P450-dependent monooxygenase. *Proc. Natl Acad. Sci. USA* 96: 10045–10050.

Jouanin, L., Goujon, T., de Nadai, V., Martin, M.T., Mila, I., Vallet, C., et al. (2000) Lignification in transgenic poplars with extremely reduced caffeic acid *O*-methyltransferase activity. *Plant Physiol.* 123: 1363–1373.

Jung, H.G. and Deetz, D.A. (1993) Cell wall lignification and degradability. *In* Forage Cell Wall Structure and Digestibility. Edited by Jung, H.G., Buxton, D.R., Hatfield, R.D. and Ralph, J. pp. 315–340. ASA-CSSA-SSSA, Madison, WI.

Kao, Y.-Y., Harding, S.A. and Tsai, C.-J. (2002) Differential expression of two distinct phenylalanine ammonia-lyase genes in condensed tannin-accumulating and lignifying cells of quaking aspen. *Plant Physiol.* 130: 796–807.

Kim, H., Ralph, J., Lu, F.C., Pilate, G., Leple, J.C., Pollet, B. and Lapierre, C. (2002) Identification of the structure and origin of thioacidolysis marker compounds for cinnamyl alcohol dehydrogenase deficiency in angiosperms. *J. Biol. Chem.* 277: 47412–47419.

Kitano, H. (2002) Systems biology: a brief overview. *Science* 295: 1662–1664.

Ko, J.H., Beers, E.P. and Han, K.H. (2006) Global comparative transcriptome analysis identifies gene network regulating secondary xylem development in *Arabidopsis thaliana*. *Mol. Genet. Genomics* 276: 517–531.

Kullback, S. and Leibler, R.A. (1951) On information and sufficiency. *Ann. Math. Stat.* 22: 79–86.

Lapierre, C., Pollet, B., Petit-Conil, M., Toval, G., Romero, J., Pilate, G., et al. (1999) Structural alterations of lignins in transgenic poplars with depressed cinnamyl alcohol dehydrogenase or caffeic acid *O*-methyltransferase activity have an opposite impact on the efficiency of industrial kraft pulping. *Plant Physiol.* 119: 153–163.

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

Leple, J.C., Dauwe, R., Morreel, K., Storme, V., Lapierre, C., Pollet, B., et al. (2007) Downregulation of cinnamoyl-coenzyme a reductase in poplar: multiple-level phenotyping reveals effects on cell wall polymer metabolism and structure. *Plant Cell* 19: 3669–3691.

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

Li, L., Cheng, X., Lu, S., Nakatsubo, T., Umezawa, T. and Chiang, V.L. (2005) Clarification of cinnamoyl co-enzyme A reductase catalysis in monolignol biosynthesis of Aspen. *Plant Cell Physiol.* 46: 1073–1082.

Li, L., Osakabe, Y., Joshi, C.P. and Chiang, V.L. (1999) Secondary xylem-specific expression of caffeoyl-coenzyme A 3-O-methyltransferase plays an important role in the methylation pathway associated with lignin biosynthesis in loblolly pine. *Plant Mol. Biol.* 40: 555–565.

Li, L., Popko, J., Umezawa, T. and Chiang, V. (2000) 5-Hydroxyconiferyl aldehyde modulates enzymatic methylation for syringyl monolignol formation, a new view of monolignol biosynthesis in angiosperms. *J. Bio. Chem.* 275: 6537–6545.

Li, L., Zhou, Y.H., Cheng, X.F., Sun, J.Y., Marita, J.M., Ralph, J., et al. (2003) Combinatorial modification of multiple lignin traits in trees through multigene cotransformation. *Proc. Natl Acad. Sci. USA* 100: 4939–4944.

Logemann, E., Parniske, M. and Hahlbrock, K. (1995) Modes of expression and common structural features of the complete phenylalanine ammonia-lyase gene family in parsley. *Proc. Natl Acad. Sci. USA* 92: 5905–5909.

Lu, S., Sun, Y., Shi, R., Clark, C., Li, L. and Chiang, V. (2005) Novel and mechanical stress-responsive microRNAs in Populus trichocarpa that are absent from Arabidopsis. *Plant Cell* 17: 2186–2203.

Lu, S., Zhou, Y., Li, L. and Chiang, V.L. (2006) Distinct roles of cinnamate 4-hydroxylase genes in *Populus*. *Plant Cell Physiol.* 47: 905–914.

Mahony, S. and Benos, P.V. (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.* 35: W253–W258.

Maizel, J.V. and Lenk, R.P. (1981) Enhanced graphic matrix analysis of nucleic acid and protein sequences. *Proc. Natl Acad. Sci. USA* 78: 7665–7669.

Meng, H. and Campbell, W.H. (1998) Substrate profiles and expression of caffeoyl coenzyme A and caffeic acid *O*-methyltransferases in secondary xylem of aspen during seasonal development. *Plant Mol. Biol.* 38: 513–520.

Meyermans, H., Morreel, K., Lapierre, C., Pollet, B., De Bruyn, A., Busson, R., et al. (2000) Modifications in lignin and accumulation of phenolic glucosides in poplar xylem upon down-regulation of caffeoyl-coenzyme A *O*-methyltransferase, an enzyme involved in lignin biosynthesis. *J. Biol. Chem.* 275: 36899–36909.

Nakatsubo, T., Kitamura, Y., Sakakibara, N., Mizutani, M., Hattori, T., Sakurai, N., et al. (2008) At5g54160 gene encodes *Arabidopsis thaliana* 5-hydroxyconiferaldehyde *O*-methyltransferase. *J. Wood Sci.* 54: 312–317.

Osakabe, K., Tsao, C., Li, L., Popko, J., Umezawa, T., Carraway, D., et al. (1999) Coniferyl aldehyde 5-hydroxylation and methylation direct syringyl lignin biosynthesis in angiosperms. *Proc. Natl Acad. Sci. USA* 96: 8955–8960.

Osakabe, Y., Osakabe, K., Kawai, S., Katayama, Y. and Morohoshi, N. (1995) Characterization of the structure and determination of mRNA levels of the phenylalanine ammonia-lyase gene family from *Populus kitakamiensis*. *Plant Mol. Biol.* 28: 1133–1141.

Patzlaff, A., Newman, L.J., Dubos, C., Whetten, R.W., Smith, C., McInnis, S., et al. (2003) Characterisation of *PtMYB1*, an R2R3-MYB from pine xylem. *Plant Mol Biol.* 53: 597–608.

Pauli, S., Rothnie, H.M., Gang, C., He, X.Y. and Hohn, T. (2004) The cauliflower mosaic virus 35S promoter extends into the transcribed region. *J. Virol.* 78: 12120–12128.

Raes, J., Rohde, A., Christensen, J., Van de Peer, Y. and Boerjan, W. (2003) Genome-wide characterization of the lignification toolbox in Arabidopsis. *Plant Physiol.* 133: 1051–1071.

Ragauskas, A., Williams, C., Davison, B., Britovsek, G., Cairney, J., Eckert, C.A., et al. (2006) The path forward for biofuels and biomaterials. *Science* 311: 484–489.

Ralph, J., Akiyama, T., Kim, H., Lu, F., Schatz, P., Marita, J., et al. (2006) Effects of coumarate 3-hydroxylase down-regulation on lignin structure. *J. Biol. Chem.* 281: 8843–8853.

Ralph, J., Brunow, G., Harris, P.J., Dixon, R.A., Schatz, P.F. and Boerjan, W. (2008) Lignification: are lignins biosynthesized via simple combinatorial chemistry or via proteinaceous control and template replication? *In* Recent Advances in Polyphenol Research. Edited by Daayf, F., El Hadrami, A., Adam, L. and Ballance, G.M. pp. 36–66. Wiley-Blackwell Publishing, Oxford.

Ralph, J., Lapierre, C., Lu, F.C., Marita, J.M., Pilate, G., Van Doorsselaere, J., et al. (2001) NMR evidence for benzodioxane structures resulting from incorporation of 5-hydroxyconiferyl alcohol into lignins of *O*-methyltransferase-deficient poplars. *J. Agric. Food Chem.* 49: 86–91.

Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* 16: 276–277.

Rogers, L.A. and Campbell, M.M. (2004) The genetic control of lignin deposition during plant growth and development. *New Phytol.* 164: 17–30.

Rogers, L.A., Dubos, C., Cullis, I.F., Surman, C., Poole, M., Willment, J., et al. (2005) Light, the circadian clock, and sugar perception in the control of lignin biosynthesis. *J. Exp. Biol.* 56: 1651–1663.

Rombauts, S., Florquin, K., Lescot, M., Marchal, K., Rouze, P. and Van de Peer, Y. (2003) Computational approaches to identify promoters and cis-regulatory elements in plant genomes. *Plant Physiol.* 132: 1162–1176.

Sarkanen, K.V. (1971) Precursors and their polymerization. *In* Lignins, Occurrence, Formation, Structure and Reactions. Edited by Sarkanen, K.V. and Ludwig, C.H. pp. 95–163. Wiley-Interscience, New York.

Sarkanen, K. (1976) Renewable resources for production of fuels and chemicals. *Science* 191: 773–776.

Schoch, G., Goepfert, S., Morant, M., Hehn, A., Meyer, D., Ullmann, P., et al. (2001) CYP98A3 from *Arabidopsis thaliana* is a 3′-hydroxylase of phenolic esters, a missing link in the phenylpropanoid pathway. *J. Biol. Chem.* 276: 36566–36574.

Schones, D.E., Sumazin, P. and Zhang, M.Q. (2005) Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics* 21: 307–313.

Shadle, G., Chen, F., Reddy, M., Jackson, L., Nakashima, J. and Dixon, R. (2007) Down-regulation of hydroxycinnamoyl CoA:shikimate hydroxycinnamoyl transferase in transgenic alfalfa affects lignification, development and forage quality. *Phytochemistry* 68: 1521–1529.

Shi, R. and Chiang, V. (2005) Facile means for quantifying microRNA expression by real-time PCR. *Biotechniques* 39: 519–525.

Smith, A.D., Sumazin, P., Das, D. and Zhang, M.Q. (2005) Mining ChIP-chip data for transcription factor and cofactor binding sites. *Bioinformatics* 21: I403–I412.

Smith, A.D., Sumazin, P., Xuan, Z. and Zhang, M.Q. (2006) DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc. Natl Acad. Sci. USA* 103: 6275–6280.

Song, J., Lu, S., Chen, Z., Lourenco, R. and Chiang, V.L. (2006) Genetic transformation of *Populus trichocarpa* genotype Nisqually-1: a functional genomic tool for woody plants. *Plant Cell Physiol.* 47: 1582–1589.

Souza, C.A., Barbazuk, B., Ralph, S.G., Bohlmann, J., Hamberger, B. and Douglas, C.J. (2008) Genome-wide analysis of a land plant-specific *acyl:coenzyme A synthetase* (*ACS*) gene family in *Arabidopsis*, poplar, rice and *Physcomitrella*. *New Phytol.* 179: 987–1003.

Sterky, F., Bhalerao, R.R., Unneberg, P., Segerman, B., Nilsson, P., Brunner, A.M., et al. (2004) A *Populus* EST resource for plant functional genomics. *Proc. Natl Acad. Sci. USA* 101: 13951–13956.

Suzuki, S., Li, L., Sun, Y. and Chiang, V. (2006) The cellulose synthase gene superfamily and biochemical functions of xylem-specific cellulose synthase-like genes in *Populus trichocarpa*. *Plant Physiol.* 142: 1233–1245.

Suzuki, S. and Umezawa, T. (2007) Biosynthesis of lignans and norlignans. *J. Wood Sci.* 53: 273–284.

Takeda, J., Ito, Y., Maeda, K. and Ozeki, Y. (2002) Assignment of UVB-responsive *cis*-element and protoplastization-(dilution-) and elicitor-responsive ones in the promoter region of a carrot phenylalanine ammonia-lyase gene (*gDcPAL1*). *Photochem. Photobiol.* 76: 232–238.

Thomas-Chollier, M., Sand, O., Turatsinze, J., Janky, R., Defrance, M., Vervisch, E., et al. (2008) RSAT: regulatory sequence analysis tools. *Nucleic Acids Res.* 36: W119–W127.

Thompson, J., Higgins, D. and Gibson, T. (1994) Clustal-W—improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22: 4673–4680.

Thompson, W., Rouchka, E.C. and Lawrence, C.E. (2003) Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res.* 31: 3580–3585.

Tsai, C.-J., Harding, S., Tschaplinski, T., Lindroth, R. and Yuan, Y. (2006) Genome-wide analysis of the structural genes regulating defense phenylpropanoid metabolism in *Populus*. *New Phytol.* 172: 47–62.

Tsai, C.-J., Popko, J.L., Mielke, M.R., Hu, W.J., Podila, G.K. and Chiang, V.L. (1998) Suppression of *O*-methyltransferase gene by homologous sense transgene in quaking aspen causes red-brown wood phenotypes. *Plant Physiol.* 117: 101–112.

Tuskan, G., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313: 1596–1604.

Van Doorsselaere, J., Baucher, M., Chognot, E., Chabbert, B., Tollier, M.-T., Petit-Conil, M., et al. (1995) A novel lignin in poplar trees with a reduced caffeic acid 5-hydroxyferulic acid *O*-methyltransferase activity. *Plant J.* 8: 855–864.

Vanholme, R., Morreel, K., Ralph, J. and Boerjan, W. (2008) Lignin engineering. *Curr. Opin. Plant Biol.* 11: 278–285.

Wardrop, A.B. (1981) Lignification and xylogenesis. *In* Xylem Cell Development. Edited by Barnett, J.R. pp. 115–152. Castle House Publication, Tunbridge Wells, UK.

Welchen, E. and Gonzalez, D.H. (2005) Differential expression of the Arabidopsis cytochrome *c* genes *Cytc*-1 and *Cytc*-2. Evidence for the involvement of TCP-domain protein-binding elements in anther- and meristem-specific expression of the *Cytc*-1 gene. *Plant Physiol.* 139: 88–100.

Welchen, E. and Gonzalez, D.H. (2006) Overrepresentation of elements recognized by TCP-domain transcription factors in the upstream regions of nuclear genes encoding components of the mitochondrial oxidative phosphorylation machinery. *Plant Physiol.* 141: 540–545.

Yamada, K., Lim, J., Dale, J.M., Chen, H., Shinn, P., Palm, C.J., et al. (2003) Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science* 302: 842–846.

Yamamoto, Y.Y., Ichida, H., Matsui, M., Obokata, J., Sakurai, T., Satou, M., et al. (2007) Identification of plant promoter constituents by analysis of local distribution of short sequences. *BMC Genomics* 8: 67–90.

Ye, Z., Kneusel, R., Matern, U. and Varner, J. (1994) An alternative methylation pathway in lignin biosynthesis in *Zinnia*. *Plant Cell* 6: 1427–1439.

Ye, Z., Zhong, R., Morrison, W.H. and Himmelsbach, D.S. (2001) Caffeoyl coenzyme A *O*-methyltransferase and lignin biosynthesis. *Phytochemistry* 57: 1177–1185.

Zhong, R., Morrison, W.H., Himmelsbach, D.S., Poole, F.L. and Ye, Z. (2000) Essential role of caffeoyl coenzyme A *O*-methyltransferase in lignin biosynthesis in woody poplar plants. *Plant Physiol.* 124: 563–577.