Plant Molecular Biology (2005) 58:229–245
DOI 10.1007/s11103-005-5346-5

Transcriptional co-regulation of secondary metabolism enzymes in *Arabidopsis*: functional and evolutionary implications

Claire M.M. Gachon, Mathilde Langlois-Meurinne, Yves Henry and Patrick Saindrenan*
*Institut de Biotechnologie des Plantes, CNRS–Université Paris–Sud, UMR8618, Bâtiment 630, 91405, Orsay Cedex, France (*author for correspondence; e-mail saindrenan@ibp.u-psud.fr)*

Received 19 January 2005; accepted in revised form 12 April 2005

Key words: *Arabidopsis*, functional genomics, gene duplication, glycosyltransferase, microarray

Abstract

The combined knowledge of the *Arabidopsis* genome and transcriptome now allows to get an integrated view of the dynamics and evolution of metabolic pathways in plants. We used publicly available sets of microarray data obtained in a wide range of different stress and developmental conditions to investigate the co-expression of genes encoding enzymes of secondary metabolism pathways, in particular indoles, phenylpropanoids, and flavonoids. We performed hierarchical clustering of gene expression profiles and found that major enzymes of each pathway display a clear and robust co-expression throughout all the conditions studied. Moreover, detailed analysis evidenced that some genes display co-regulation in particular physiological conditions only, certainly reflecting their modular recruitment into stress- or developmentally regulated biosynthetic pathways. The combination of these microarray data with sequence analysis allows to draw very precise hypotheses on the function of otherwise uncharacterized genes. To illustrate this approach, we focused our analysis on secondary metabolism glycosyltransferases (UGTs), a multigenic family involved in the conjugation of small molecules to sugars like glucose. We propose that UGT74B1 and UGT74C1 may be involved in aromatic and aliphatic glucosinolates synthesis, respectively. We also suggest that UGT75C1 may function as an anthocyanin-5-*O*-glucosyltransferase *in planta*. Therefore, this data-mining approach appears very powerful for the functional prediction of unknown genes, and could be transposed to virtually any other gene family. Finally, we suggest that analysis of expression pattern divergence of duplicated genes also provides some insight into the mechanisms of metabolic pathway evolution.

Abbreviations: CoA, coenzyme A; Glc, glucose; GT, glycosyltransferase; Myrs, million of years; WGD, whole genome duplication

Introduction

Transcriptional regulation has long been recognized as a major determinant of phenotypic variation, just like protein primary sequence modification (Wray *et al.*, 2003). Using classical molecular biology techniques, co-ordinate expression of secondary metabolite enzymes has been described in several instances in *Arabidopsis*, as in stress-inducible indole metabolism (Zhao and Last, 1996; Zhao *et al.*, 1998) or in flavonoid synthesis during seedling development (Pelletier *et al.*, 1999 and references therein). Likewise, key enzymes of the phenylpropanoid pathway, namely phenylalanine ammonia-lyase (PAL), coumarate-4-hydroxylase (C4H) and some 4-coumaroyl CoA ligase (4CL) isoforms, are expressed in a co-ordinated manner in stress conditions (Ehlting *et al.*, 1999). Extensive transcriptional perturbations resulting from the knock-out (or over-expression) of a single gene

of a pathway has also been demonstrated, showing both a co-ordination of gene transcription in a given pathway and a profound interdependency between pathways (see e.g. Martz *et al.*, 1998; Rohde *et al.*, 2004). In this respect, microarrays are an ideal tool to systematically investigate metabolic changes in a wide range of experimental conditions, whilst previous studies only focused on a restricted number of genes due to technical limitations. The compilation of microarray data obtained in numerous experimental conditions has the potential to give a comprehensive view of metabolism organization and plasticity during all aspects of plant life.

Moreover, many results of earlier studies on secondary metabolism have to be reinterpreted in the light of abundant duplication of the underlying genes. Indeed, extensive gene duplication processes seem to be a characteristic feature of angiosperms (Bowers *et al.*, 2003). Gene duplication is believed to be a major evolutionary mechanism leading to the emergence of new functions (Gu *et al.*, 2004), and this phenomenon might at least partially account for the extreme diversity of secondary metabolites synthesized in plants (Schwab, 2003). Two major mechanisms account for the emergence of gene families: tandem gene duplication, which can occur at any time, and large-scale duplications of chromosomal segments, or even of the whole genome (polyploidy). In *Arabidopsis*, at least two whole genome duplication (WGD) events have been identified (Blanc *et al.*, 2003), followed by extensive chromosomal rearrangements: a 'recent' WGD dated 24–40 Myrs occurred just after the emergence of the *Brassicaceae*, whereas an 'old' WGD most likely occurred 100 Myrs ago in the early evolution of Rosids. An older WGD putatively took place in the very early angiosperm history (Bowers *et al.*, 2003). The investigation of the mechanisms leading to the genesis and maintenance of multigenic families is still in its infancy, but already demonstrates that gene function influences both gene copy number and genome organization. In particular, it is striking that tandem-duplicated genes encoding secondary metabolism enzymes have a high probability to be maintained in the genome, whereas other protein families like WRKY transcription factors have expanded exclusively through WGD (Cannon *et al.*, 2004). Although the reasons of this phenomenon are not understood, it certainly originates from functional constraints, and must strongly impact on metabolic pathway evolution.

A widely accepted explanation for maintenance of a duplicated gene pair is functional specialization (Lynch and Conery, 2000). Paralogues may retain different subsets of the ancestral gene function, a process called sub-functionalization (Force *et al.*, 1999). In some rarer cases, one gene retains the ancestral function whereas the second undergoes accelerated evolution and acquires a new function (neo-functionalization) (Kellis *et al.*, 2004; Zhang *et al.*, 2003). Paralogue genes may also exhibit an overlapping function, as shown by the higher probability of functional compensation for null mutations of duplicated genes compared to singletons (Gu *et al.*, 2003). Regarding plant secondary metabolism, functional specialization of paralogue enzymes has been inferred from genetic data (e.g Niyogi and Fink, 1992). Differences in the expression pattern of duplicated genes have also widely been used as an argument supporting this hypothesis, as exemplified by Mobley *et al.* (1999) for chorismate mutase genes. Definite evidence of specialization of paralogue genes encoding key secondary metabolism enzymes has already been described in maize (Frey *et al.*, 2000). In this plant, three related enzymes catalysing the synthesis of indole (tryptophan synthase α, Bx1 and Igl) are recruited in three different physiological conditions (protein synthesis, phytoanticipin accumulation and herbivore stress response) and are unable to complement each other when mutated. Thus, identical metabolic pathways originating from duplicated genes can act in parallel *in planta* in distinct physiological conditions. Specific transcription patterns related to this specialized function can be expected, and should be detected in transcriptomic analyses. As such, whole genome microarray experiments allow to investigate functional specialization at the scale of an enzyme or of a complete pathway encoded by gene families. They should also provide insight into how duplicated genes are recruited into new metabolic pathways, which in turn will help to get a better description of their function.

In our laboratory, we are committed in the functional characterization of secondary metabolism glycosyltransferases (UGTs). This family of enzymes is responsible for the conjugation of small molecules to sugars (Jones and Vogt, 2001). They are encoded by about 120 genes in *Arabidopsis*

(Paquette *et al.*, 2003). Secondary metabolite glycosylation has long been recognized as a physiologically very important process. For instance, it contributes to hormone homeostasis (Lim and Bowles, 2004; Sembdner *et al.*, 1994), and increases the stability and solubility of molecules. For this reason, it is used for detoxification of toxins or xenobiotics (Jones and Vogt, 2001). Glycosylation was also suggested to be required for monolignol export into the apoplasm during lignification (Freudenberg and Harkin, 1963). Despite the interest in these enzymes, only a small fraction of them have been functionally characterized so far (Lim and Bowles, 2004).

In this study, we first intended to get an integrated view of secondary metabolism organization and dynamics in plant development and stress responses. We therefore compiled the information obtained with classical genetic and biochemical approaches and made out a comprehensive list of the functionally characterized *Arabidopsis* enzymes involved in the shikimate, phenylpropanoid, flavonoid, and indole pathways. We then retrieved the expression data of those genes from whole-genome microarray datasets. We used hierarchical clustering and bootstrap analysis to define robust clusters of genes displaying a co-ordinate expression. We further tried to take advantage of these data to predict a function for previously uncharacterized genes, focusing on members of the *Arabidopsis Ugt* gene family. We investigated if some of them were associated with one of the previously defined clusters in order to tentatively assign them to a particular metabolic pathway. Finally, we show that co-expression analysis also provides hints about metabolic pathway evolution and we propose a scenario for the evolution of glucosinolate biosynthesis in *Arabidopsis*.

Methods

Establishing a list of genes for microarray expression data clustering

In order to obtain reliable functional annotation, we restricted our analysis to *Arabidopsis* genes functionally characterized with biochemical or genetic approaches. Our analysis encompasses biosynthetic genes of phenylpropanoid, flavonoid, indole and glucosinolate metabolites (Table 1). AGI accession numbers were assigned according to the final TIGR *Arabidopsis* genome annotation (TIGR release 5.0). A BLAST analysis on the *Arabidopsis* genome was conducted to identify the closest paralogues of those genes. These were also included in our study in order to investigate the transcriptional specialization within gene families. A number of biotic or abiotic stress markers were also included to provide a wide range of transcriptional profiles. Some genes involved in isoprenoid metabolism are represented, mostly to provide a general view of biosynthetic pathways of glycosylated hormones (abscisic acid, gibberellins, brassinosteroids). Note that only a fraction of *Ugts* were included in our analysis. The complete list, including AGI number, Affymetrix™ probe number, functional description and relevant literature references is available as Supplementary Table 1.

Description of the microarray dataset

We analysed a dataset recently released by the AtGenExpress consortium (http://web.uni-frankfurt.de/fb15/botanik/mcb/AFGN/atgenex.htm), which encompasses a very wide range of experimental conditions. It was generated with the Affymetrix™ ATH1 GeneChip probe array (Redman *et al.*, 2004). The 'stress' time courses investigate the transcriptome of shoots and roots of plants submitted to cold, heat, drought, osmotic, saline, oxidative, UV and wounding stresses, or that were treated with elicitors. Pathogen infection is represented by two time-courses with the Oomycete *Phytophthora infestans* or various strains of the bacterium *Pseudomonas syringae*. The 'developmental' series contain expression data for all plant organs at successive development stages. Finally, a 'light' dataset investigates the plant responses to various light stimuli, in particular phytochrome- and cryptochrome-mediated responses. It was grouped with the developmental series for analysis.

Hierarchical clustering of microarray expression data

For the stress and light datasets, genes were clustered according to their expression fold change compared to the corresponding control experiment.

Table 1. Abridged list of genes included in this study.

AGI ID	Gene abbreviation	Protein name and function
Glycosyltransferases		
At2g36790	*Ugt73c6*	Flavonol-7-*O*-GT (UGT73C6)
At1g24100	*Ugt74b1*	UGT74B1 (glucosinolate synthesis)
At2g31790	*Ugt74c1*	UGT74C1
At2g31750	*Ugt74d1*	UGT74D1
At1g05670	*Ugt74e1*	UGT74E1
At1g05680	*Ugt74e2*	UGT74E2
At2g43840	*Ugt74f1*	UGT74F1
At2g43820	*Ugt74f2*	Anthranilate-GT (UGT74F2)
At1g05560	*Ugt75b1*	Callose-synthase associated-GT (UGT75B1)
At1g05530	*Ugt75b2*	UGT75B2
At4g14090	*Ugt75c1*	UGT75C1
At4g15550	*Ugt75d1*	UGT75D1
At1g30530	*Ugt78d1*	Flavonol-3-rhamnosyltransferase
At4g27560	*Ugt79b2*	UGT79B2
At4g27570	*Ugt79b3*	UGT79B3
At4g15480	*Ugt84a1*	Putative lignin precursor glucose ester GT (UGT84A1)
At2g23260	*Ugt84b1*	Auxin-GT (UGT84B1)
Stress markers, defence signalling		
At1g07160	*PP2C*	Phosphatase 2C
At4g28350	*LecRpk*	LecRPK (lectin receptor protein kinase)
At2g39200	*Mlo homolog*	MLO homolog
At2g19190	*Frk1*	FRK1 (flagellin-induced receptor kinase)
At1g72520	*Lox3*	Lipoxygenase 3 (putatively involved in JA synthesis)
At3g25760	*Aoc1*	Allene oxide cyclase 1 (JA synthesis)
At3g25770	*Aoc2*	Allene oxide cyclase 2 (JA synthesis)
At3g25780	*Aoc3*	Allene oxide cyclase 3 (JA synthesis)
At5g42650	*Aos*	Allene oxide synthase (CYP74, JA synthesis)
At3g28930	*Aig2*	AIG2 (AvrRpt2-induced gene)
At3g50970	*Xero2*	Cold-induced dehydrin
At2g42530	*Cor15 b*	COR15B
At5g52310	*Pr 78*	Low temperature-induced protein
At2g23120	*Cold-induced Pr*	Cold-induced protein
Shikimate pathway, tyrosine and folate synthesis		
At4g39980	*Dhs1*	DAHP synthase 1 (DHS1)
At4g33510	*Dhs2*	DAHP synthase 2 (DHS2)
At2g45300	*EPSP-synthase*	EPSP-synthase
At1g48860	*EPSP-like*	EPSP-synthase-like
At1g48850	*Cs*	Putative chorismate synthase (CS)
At3g29200	*Cm1*	Chorismate mutase 1 (CM1)
At5g10870	*Cm2*	Chorismate mutase 2 (CM2)
At1g69370	*Cm3*	Chorismate mutase 3 (CM3)
At2g28880	*ACS synthase*	4-Aminodeoxychorismate (ACS) synthase
At5g57850	*PABA synthase*	Aminodeoxychorismate lyase
Trp and auxin (IAA) synthesis, Trp-derived metabolism		
At5g05730	*Asa1* (*Trp5*)	Anthranilate synthase α1 (ASA1)
At2g29690	*Asa2*	Anthranilate synthase α2 (ASA2)
At1g25220	*Asb1* (*Trp4*)	Anthranilate synthase β1 (ASB1, TRP4)
At5g57890	*Asb2*	Anthranilate synthase β2
At1g25165	*Asb3*	Anthranilate synthase β3
At5g17990	*Pat*	Phosphoribosyl-anthranilate transferase (PAT)
At1g07780	*Pai*	Phosphoribosyl-anthranilate isomerase (PAI)
At3g54640	*Tsa*	Tryptophan synthase α (TSA)
At5g54810	*Tsb1*	Tryptophan synthase β1 (TSB1, TRP2)
At4g27070	*Tsb2*	Tryptophan synthase β2 (TSB2)

Table 1. Continued.

AGI ID	Gene abbreviation	Protein name and function
At3g26830	*Pad3*	CYP71B15 (camalexin synthesis)
At4g03060	*Aop2*	2-Oxoglutarate dependent dioxygenase (aliphatic glucosinolate synthesis)
At4g03050	*Aop3*	2-Oxoglutarate dependent dioxygenase (aliphatic glucosinolate synthesis)
At5g05260	*Cyp79a2*	CYP79A2 (benzylglucosinolate synthesis)
At4g39950	*Cyp79b2*	CYP79B2 (auxin and indolic glucosinolate synthesis)
At2g22330	*Cyp79b3*	CYP79B3 (auxin and indolic glucosinolate synthesis)
At1g16410	*Cyp79f1*	CYP79F1 (aliphatic glucosinolate synthesis)
At1g16400	*Cyp79f2*	CYP79F2 (aliphatic glucosinolate synthesis)
At4g13770	*Cyp83a1* (*Ref2*)	CYP83A1 (REF2, aliphatic glucosinolate synthesis)
At4g31500	*Cyp83b1*	CYP83B1 (indole glucosinolate synthesis
At2g20610	*Sur1* (*C-S lyase*)	C-S lyase (SUR1, glucosinolate synthesis)
At3g02875	*Ilr1*	ILR1 (auxin conjugate hydrolysis)
At1g51760	*Iar3*	IAA–Ala hydrolase (IAR3, auxin conjugate hydrolysis)
At3g44310	*Nit1*	Nitrilase (NIT1, auxin synthesis)
At3g44300	*Nit2*	Nitrilase (NIT2, auxin synthesis)
At3g44320	*Nit3*	Nitrilase (NIT3, auxin synthesis)
Flavonoid synthesis and regulation		
At5g13930	*Chs*	Chalcone synthase (CHS, TT4)
At3g55120	*Chi*	Chalcone isomerase (CHI, TT5)
At3g51240	*F3h* (*Tt6*)	Flavonol-3-hydroxylase (F3H, TT6)
At5g07990	*F3'h* (*Tt7*)	Flavonol-3'-hydroxylase (F3'H, TT7, CYP75B1)
At5g08640	*Fls*	Flavonol synthase (FLS)
At5g42800	*Dfr*	Dihydroflavonol reductase (DFR, TT3)
At1g61720	*Anr* (*Banyuls*)	Anthocyanidin reductase (ANR, BAN)
At5g17220	*Gst* (*Tt19*)	Glutathione-*S*-transferase (TT19)
At3g59030	*Tt12*	MATE transporter (TT12)
At4g22880	*Ldox* (*Tt18*)	Leucoanthocyanidin dioxygenase (LDOX, TT18, TDS4, ANS: anthocyanidin synthase)
At5g35550	*Tt2* (*Myb123*)	R2R3MYB transcription factor (MYB123)
At4g09820	*Tt8*	Basic helix-loop-helix domain protein
At5g23260	*Tt16* (*Abs*)	BSISTER MADS domain protein
At5g24520	*Ttg1*	WD40 repeat protein
At2g37260	*Ttg2*	WRKY transcription factor
Phenylpropanoid synthesis and regulation		
At2g37040	*Pal1*	Phenylalanine ammonia lyase (PAL1)
At3g53260	*Pal2*	Phenylalanine ammonia lyase (PAL2)
At5g04230	*Pal3*	Phenylalanine ammonia lyase (PAL3)
At3g10340	*Pal4*	Phenylalanine ammonia lyase (PAL4)
At2g30490	*C4h* (*Ref3*)	Cinnamate-4-hydroxylase (C4H, REF3, CYP73A5)
At2g40890	*C3h1* (*Ref8*)	Coumarate-3-hydroxylase (C3H1, REF8, CYP98A3)
At1g74540	*C3h2*	C3H-like (C3H2, CYP98A8)
At1g74550	*C3h3*	C3H-like (C3H3, CYP98A9)
At4g36220	*F5h-1* (*Fah1*)	Ferulate-5-hydroxylase (F5H1, FAH1, CYP84A1)
At5g04330	*F5h-2*	F5H-like (F5H2, CYP84A4)
At1g51680	*4Cl1*	4-Coumaroyl-CoA ligase (4CL1)
At3g21240	*4Cl2*	4-Coumaroyl-CoA ligase (4CL2)
At1g65060	*4Cl3*	4-Coumaroyl-CoA ligase (4CL3)
At3g21230	*4Cl5*	4-Coumaroyl-CoA ligase (4CL5)
At1g20510	*4Cl9*	By homology (4CL9)
At4g34050	*CCOmt1*	Putative caffeoyl-CoA *O*-methyltransferase (CCOMT1)
At4g26220	*CCOmt7*	Putative caffeoyl-CoA *O*-methyltransferase (CCOMT7)
At5g54160	*COmt1* (*Omt-1*)	*O*-methyltransferase 1 (COMT1, OMT1)
At1g33030	*COmt10* (*Omt10*)	Putative *O*-methyltransferase (COMT10, OMT10)
At1g15950	*Ccr1* (*Irx4*)	Cinnamyl-CoA reductase (CCR1, IRX4)
At2g23910	*Ccr6*	Putative cinnamyl-CoA reductase (CCR6)
At4g39330	*Cad-1*	Cinnamyl alcohol dehydrogenase (CAD1, AtCAD9)
At4g37990	*Cad-B2* (*Eli3-2*)	CAD-B2 (ELI3-2, AtCAD8)

Table 1. Continued.

AGI ID	Gene abbreviation	Protein name and function
At3g24503	*Aldh (Ref1)*	Aldehyde dehydrogenase (ALDH, REF1)
At5g48930	*Hct*	Hydroxycinnamoyltransferase (HCT)
At1g56650	*Pap1 (Myb75)*	MYB transcription factor (PAP1-D, MYB75)
At2g22990	*Sng1*	Sinapoylglucose:malate sinapoyltransferase (SNG1)
Terpenoid metabolism		
At1g05160	*Kao1*	Ent-kaurenoic acid oxidase (KAO1)
At5g05600	*G7Ox-2*	Putative gibberelin-7-oxidase (G7OX-2)
At5g51810	*Ga5-5*	Gibberelin-20-oxidase (G5A-5)
At3g50660	*Cyp90b1*	Steroid C22-hydroxylase (CYP90B1, brassinosteroid synthesis)
At5g05690	*Cyp90a1*	Steroid C2-hydroxylase (CYP90A1, brassinosteroid synthesis)
At3g30180	*Cyp85a2*	BR-6-oxidase 2 (CYP85A2, brassinosteroid synthesis)

Brief description of the genes and proteins discussed in the text or mentioned in the figures. The reader is invited to refer to Supplementary Table 1 for the complete list of genes taken into account in this study, Affymetrix™ probe names and relevant bibliographical references.

For developmental datasets, we took the median value of expression level in a given dataset as a reference for each gene. To avoid gene clustering based on unreliable very low expression data, the minimum expression level of each probe was set to 10 before calculating the expression fold change (See Supplementary Material 1 for a detailed presentation of this method). The latter was further log2-transformed and hierarchical clustering (average distance linkage) using the uncentered Pearson correlation coefficient as a distance measure was performed with the MeV software (Saeed *et al.*, 2003). The reliability of gene clusters was assessed with experiment bootstrapping (1000 replications). A similar exploitation of other microarray data available on public databases (NASCArrays: http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl; TAIR: http://www.arabidopsis.org) and clustering according to other commonly used methods (k-means, SOMs) was also attempted and led essentially to the same results.

Sequence analysis of cytochrome P450 and glycosyltransferases

The sequences and annotations were retrieved from the *Arabidopsis cytochrome P450 and glycosyltransferase family 1 Site at PlaCe* (http://www.biobase.dk/P450). The N-terminal part (until the LPPGP motif) of CYP79 protein sequences was not taken into account for the alignment because of its high divergence. UGTs were aligned over their complete sequence. Multiple alignments were generated using clustalw (BLOSUM62 matrix), and then manually edited. Neighbour-joining distance trees were built using the Jones–Taylor–Thornton matrix provided in the Phylip package (Felsenstein, 1989). Bootstrap analysis (1000 replications) was performed to assess the robustness of tree nodes. The identification of genes originating from WGD and the dating of these polyploidization events were determined according to the analysis of Blanc *et al.* (2003), using the related website (http://wolfe.gen.tcd.ie/athal/dup). Tandemly-duplicated genes were evidenced both through their close relationships on phylogenetic trees and their physical proximity on *Arabidopsis* chromosomes, as revealed by their AGI accession number.

Results

Clustering of gene expression profiles using bootstrap

An exhaustive literature survey was conducted to establish a comprehensive list of *Arabidopsis* genes functionally characterized as involved in the shikimate, phenylpropanoid, flavonoid, indole pathways (Table 1, Suppl. Table 1). Hierarchical clustering of microarray data released by the AtGenExpress consortium (http://web.uni-frankfurt.de/fb15/botanik/mcb/AFGN/atgenex.htm) was used to group these genes according to their expression pattern. In total, the expression fold

change of 266 probes was calculated in 563 experimental conditions, covering all stages of plant development, multiple light conditions, and a wide range of biotic and abiotic stresses. This dataset was further subdivided into stress and developmental (including light experiments) series to specifically assess gene co-expression in those conditions. The high number of experiments in each dataset allowed to perform bootstrap analysis to estimate the robustness of tree nodes. The major clusters revealed by this analysis are represented on Figure 1. The exhaustive results can be downloaded as Supplementary Figures 1, 2, and 3.

To facilitate data interpretation, our study is focused on genes involved in secondary metabolism. This might distort our results towards the artefactual clustering of some of those genes together. In order to validate our approach, we therefore included in our analysis some other marker genes previously characterized as acting together in various stress responses; jasmonate biosynthetic genes are co-ordinately induced in stress and developmental processes (Turner *et al.*, 2002), and indeed fall into the same cluster (Figure 1A). Cold-induced genes (Kreps *et al.*, 2002) are clustered together with bootstrap values close to 100% (Figure 1B). These examples show that gene associations supported by high bootstrap values can be interpreted as biologically relevant. However, bootstrap values are very often lower than on usual phylogenetic trees. This can be ascribed to the nature of the data under analysis, which are subject to both biological variation and microarray technical limitations. We empirically determined that bootstrap values over 50% can most often be considered significant, whereas they would only reflect poor tree node support on a classical phylogenetic tree.

Co-ordinate expression of genes involved in major metabolic pathways

Flavonoids

Strikingly, the most robust clusters specifically contain key genes of each metabolic pathway that we took into consideration (Figure 1). Figure 2 summarizes the biological significance of the most prominent ones. For instance, the flavonoid biosynthetic genes are grouped in three distinct clusters (Figure 1A). One comprises genes encoding the first enzymes of the pathway: 4-coumaroyl CoA ligase 3 (4CL3), chalcone synthase (CHS), chalcone isomerase (CHI), flavonol-3-hydroxylase (F3H), flavonol-3′-hydroxylase (F3′H) and flavonol synthase (FLS), as well as a cinnamoyl-CoA reductase homologue (CCR6) and two glycosyltransferases (UGT78D1, UGT84A1). The second cluster is made of genes encoding dihydroflavonol reductase (DFR), leucoanthocyanidin dioxygenase (LDOX, also called anthocyanidin synthase), a glutathione-*S*-transferase (TT19) and the glycosyltransferase UGT75C1. It reflects the onset of anthocyanin synthesis in flowers and is repressed in stems and leaves, where flavonol synthesis is known to be very active. The latter cluster also contains *Pap1* and *Ttg2* (in developmental conditions, Figure 1C), which encode two known regulators of the flavonoid pathway (Borevitz *et al.*, 2000; Johnson *et al.*, 2002). Intriguingly, *Ugt84a1*, a putative sinapoylglucose ester glucosyltransferase (Lim *et al.*, 2001), is strongly associated with the flavonol cluster in the developmental datasets (Figure 1C). Finally, the *Banyuls* gene encoding anthocyanin reductase (ANR) is isolated in a cluster of genes specifically expressed in developing seeds, reflecting the re-orientation of metabolic fluxes towards tannin synthesis in this organ (Figure 1A). Remarkably, *Banyuls* is grouped with *Tt2* and *Tt8*, two genes encoding transcription factors which physically interact to regulate its expression (Baudry *et al.*, 2004). TTG1, a more generalist transcription factor also involved in this regulatory complex, displays a wider expression pattern and is logically rejected outside this cluster. Despite its clear biological significance, this cluster appears less stable than the two others, most probably because the genes are detected only in a limited number of experiments corresponding to the first stages of seed development.

Trp and trp-derived metabolism

The indole biosynthetic genes encoding phosphoribosyl-anthranilate transferase (PAT), anthranilate synthase α1 (ASA1), tryptophan synthase α (TSA), and *Cyp71b15* (also known as *Pad3*), are co-expressed in all the conditions studied (Figure 1A). Even if the ambiguous design of Affymetrix™ probes does not allow to distinguish between isoforms of anthranilate synthase β (ASB1, ASB2) and tryptophan synthase β (TSB1,

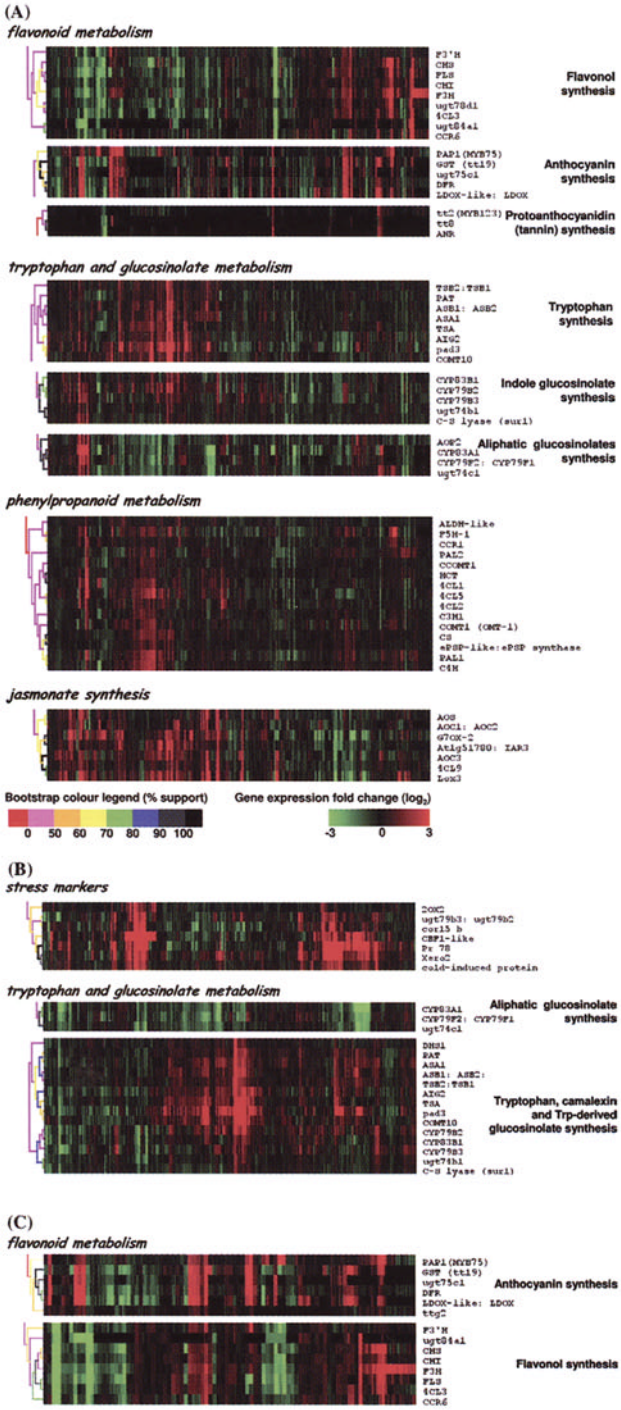

Figure 1. Major groups of co-expressed genes revealed by hierarchical clustering of expression profiles. The expression profiles of about 250 *Arabidopsis* genes involved in secondary metabolism were regrouped using average linkage hierarchical clustering. The tree robustness was determined by bootstrap analysis. Bootstrap values associated to each node are represented by a colour legend. Some biologically relevant clusters supported by significant bootstrap values can be recognized. Gene abbreviations are defined in Table 1. Genes represented by an ambiguous probe are separated by a colon. The reader is invited to refer to Supplementary Figures 1–3 for the complete description of clusters and experimental conditions. (A) all experimental conditions (563 experiments), (B) stress datasets (290 experiments), (C) developmental conditions (273 experiments).

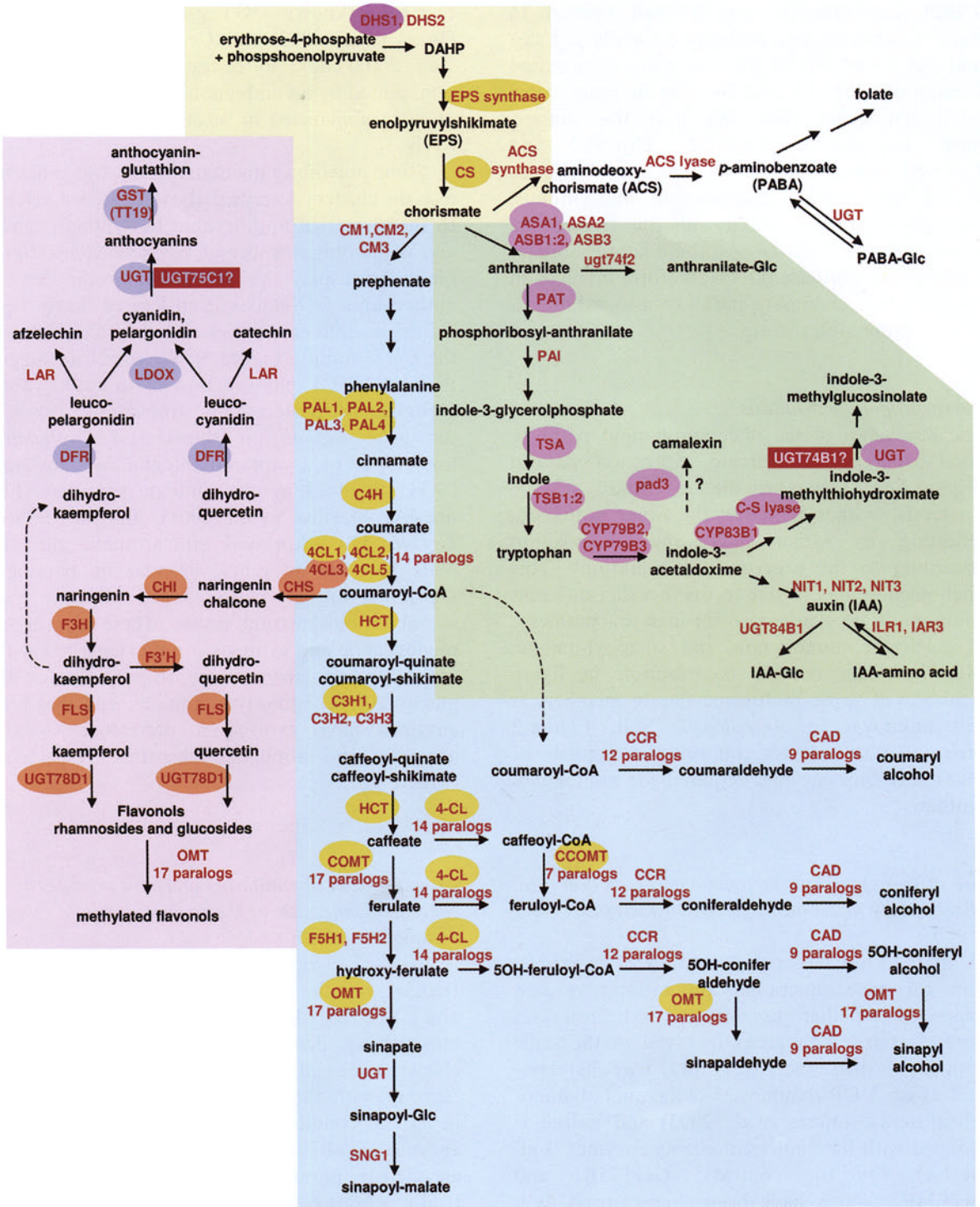

Figure 2. Co-expression of genes encoding enzymes of the shikimate (yellow), indole (green), phenylpropanoid (blue) and flavonoid (purple) pathways in *Arabidopsis*. This scheme summarizes the organization and relationships between these pathways in *Arabidopsis*. The principal enzymes taken into account in this study are noted in red. Group of genes that were co-expressed in our study are highlighted in the same colour.

TSB2), these genes are also grouped together. In stress conditions, this pathway is clearly induced and the robustness of the clustering is increased (compare Figure 1A and 1B). Furthermore, some additional genes then fall into this cluster, most importantly *Cyp83b1*, *Cyp79b2* and *Cyp79b3*, which encode cytochromes P450 involved in indole glucosinolate biosynthesis. Though not tightly linked with this cluster, the stress-inducible 3-deoxy-D-arabino-heptulosonate-7-phosphate-synthase (DHS1) isoform involved in the upstream shikimate pathway also exhibits a similar expression pattern.

Phenylpropanoid synthesis

The key genes of the phenylpropanoid pathway also display a co-ordinate expression pattern (Figure 1A). However, the robustness of the cluster is weaker than for the other pathways, reflecting its rather fluctuating composition according to the experimental conditions. This phenomenon may be due to the fact that different compounds are synthesized through this pathway, in particular monolignols and sinapoyl-malate. Moreover, the relative contribution in lignin synthesis of some multigenic family members is still unknown in *Arabidopsis*. Still, Figure 2 shows that this cluster contains the majority of genes encoding enzymes required for monolignol synthesis.

Use of co-expression for functional prediction of secondary metabolism glycosyltransferases

In order to test the predictive value of co-expression for gene functional annotation, we first addressed whether known glycosyltransferases were clustered with genes involved in the same pathway. For instance, UGT78D1 was characterized as an UDP-rhamnose:3-*O*-flavonol rhamnosyltransferase (Jones *et al.*, 2003) and indeed is grouped with flavonol synthesizing enzymes (Figure 1A). On the contrary, UGT75B1 and UGT74F2, which have been characterized as a callose synthase-associated GT (Hong *et al.*, 2001) and an anthranilate-GT (Quiel and Bender, 2003) respectively do not exhibit a transcriptional co-regulation with any gene involved in the same cellular process (supplementary Figure 1). Finally, two other known UGT genes, namely *Ugt73c6* (Jones *et al.*, 2003) and *Ugt84b1* (Jackson *et al.*, 2001, 2002) could not be tested, because they were represented by an ambiguous Affymetrix™ probe or were undetected in most experiments, respectively.

Some previously uncharacterized *Ugt* genes fall into the clusters described above: *Ugt75c1* belongs to the cluster of anthocyanin biosynthetic genes, and is the only member of the *Arabidopsis Ugt75* family to display this expression profile. Several anthocyanin-5-*O*-glucosyltransferases have been cloned in different species and they all belong to the *Ugt75* family (Figure 3). We therefore suggest that UGT75C1 may function as an anthocyanin-5-*O*-glucosyltransferase in *Arabidopsis*. Likewise the gene *Ugt74b1* is the closest *Arabidopsis* homologue of a rapeseed thioglucosyltransferase (SGT) involved in glucosinolate synthesis (Figure 4A, Marillia *et al.*, 2001). In our dataset, *Ugt74b1* is co-expressed with aromatic glucosinolate biosynthetic genes, whereas its paralogue *Ug74c1* is strongly clustered with aliphatic glucosinolate biosynthetic genes. These converging phylogenetic and expression data suggest that the corresponding enzymes are responsible for the glucosylation of thiohydroximates, and that both enzymes have evolved a preference towards aromatic and aliphatic substrates, respectively (Figure 4B).

Co-expression of similarly composed modules suggests duplication at the scale of a whole metabolic pathway

Remarkably, the genes encoding CYP79B2, -79B3 and CYP83B1, all involved in indole glucosinolate biosynthesis, define a very stable co-expression cluster over all experimental conditions, and regroup with the tryptophan biosynthetic genes in stress conditions (Figure 1A and B). The enzymes CYP79F1, -79F2 and CYP83A1 are encoded by paralogue genes. They catalyse identical reactions compared to CYP79B2, -79B3 and CYP83B1, but only metabolize methionine-derived substrates (Bak and Feyereisen, 2001; Hansen *et al.*, 2001; Chen *et al.*, 2003; Hemm *et al.*, 2003). They display a different expression pattern, but co-expression between members of

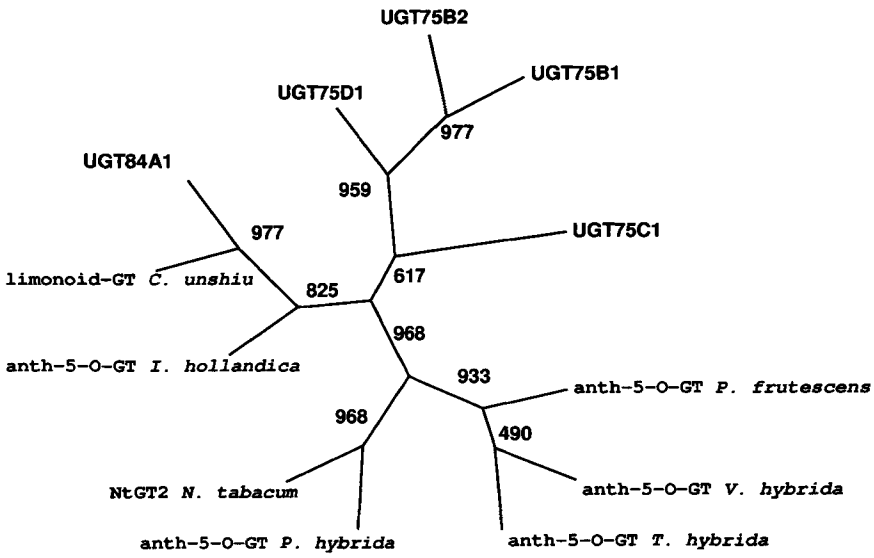

Figure 3. Phylogenetic affinity of UGT75C1 towards known anthocyanin-5-*O*-glucosyltransferases from other species. Neighbour-joining tree representing the *Arabidopsis* UGT75 family and anthocyanin-5-*O*-GTs from other species. Bootstrap values (1000 replications) are indicated for each node. The Genbank accession numbers of these sequences are: *Perilla frutescens* and *Verbena hybrida*: AB013596 and AB013598, respectively (Yamazaki *et al.*, 1999), *Torenia hybrida*: AB076698, *Petunia hybrida*: AB027455, *Iris hollandica*: AB113664 (Imayama *et al.*, 2004). *Nicotiana tabacum* NTGT2 is not an anthocyanin-5-*O*-GT but glucosylates flavonols on their 7-OH group (Taguchi *et al.*, 2003). The closely related protein sequences of UGT84A1 and the limonoid-GT from *Citrus unshiu* (AB033758, Kita *et al.*, 2000) are also indicated.

this latter pathway is retained (Figure 1A and B). Furthermore, each of these clusters contains one of the paralogue glucosyltransferase genes *Ugt74b1* and *Ugt74c1*, which we propose to be involved in the desulfoglucosinolate biosynthetic step. The reconstruction of the duplication events in the *Ugt74* and *Cyp79* families shows that the genes belonging to these two clusters originate from duplications older than 24–40 Myrs and 100 Myrs respectively (Figure 4A). Given their level of sequence divergence, the *Cyp83a1* and *Cyp83b1* genes also probably result from an old duplication, which could not be ascribed to either of the two known *Arabidopsis* polyploidization events (data not shown). Taken together, these results are compatible with the hypothesis of an ancient duplication of the glucosinolate pathway as a whole (Figure 4B). Sequence analysis of the *Arabidopsis Cyp* family further suggests that the tandem duplication of *Cyp79f1* and *-f2* later resulted in a substrate specialization into short-chain and long lateral chain methionine derivatives (Chen *et al.*, 2003; Reintanz *et al.*, 2001), whereas a marked divergence between *Cyp79a2* and the *Cyp*79b- genes accounts for their reported specialization into phenolic or indole substrates, respectively (Figure 4B).

Discussion

Assessing the reliability of co-expressed gene clusters with bootstrap

Statistical exploitation of microarray data raises a number of mathematical concerns, mostly because expression data are treated as if they were an accurate measure of the expression level and the effects of measurement errors are not addressed. Bootstrapping is widely used for the construction of phylogenetic trees and can be transposed to microarray analysis to specifically address this concern (Zhang and Zhao, 2000; Van der Laan and Bryan, 2001). However it is not useful unless a high number of experiments is taken into account. The restricted number of hybridizations in microarray datasets has prevented the systematic use of this method in plant biology until recently, but the

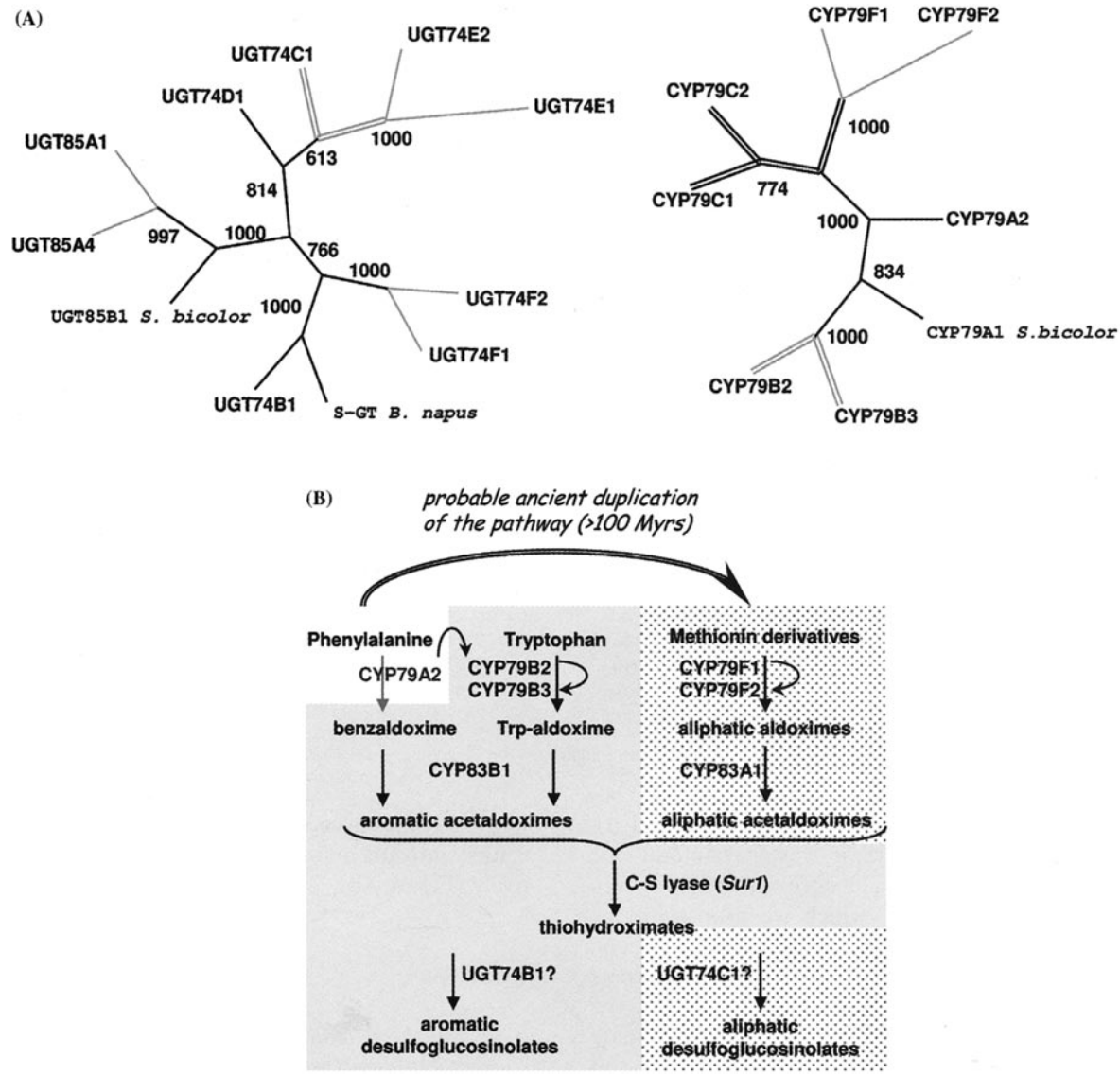

Figure 4. Conserved transcriptional co-regulation suggests pathway duplication. (A) Phylogenetic relationships between enzymes involved in aliphatic and aromatic glucosinolate synthesis. Neighbour-joining trees of the *Arabidopsis* CYP79, UGT74 and UGT85 protein families. Bootstrap values (1000 replications) are indicated for each node. Some sequences were omitted for clarity. Homologue genes from *Sorghum bicolor* and *Brassica napus* are indicated in a different font. Whenever known, the nature of gene duplication events is indicated (single black line: undetermined; single grey line: recent tandem duplication; double grey line: recent polyploidization, 24–40 Myrs; double black line: ancient polyploidization, around 100 Myrs). (B) Model of glucosinolate synthesis evolution inferred from phylogenetic and microarray data. *Arabidopsis* contains two parallel pathways specialized in aromatic and methionine-derived glucosinolate synthesis. They are composed of paralogue enzymes, which display a strong co-expression pattern (light gray and dotted background, respectively). The gene encoding CYP79A2 is very lowly expressed and unfortunately was not detected in the majority of the microarray experiments. This organization suggests an ancient duplication of the metabolic pathway as a whole, followed by several rounds of gene duplications (small curved arrows, see Figure 4A). Whether the *Sur1* gene was never duplicated or if its paralogue was subsequently lost remains unknown. See text for details.

public release of larger datasets now allows to do it. The straightforward biological significance of the most robust clusters revealed in this study demonstrate that it is a very powerful tool to facilitate the interpretation of microarray data. Finally, our study is restricted to hierarchical clustering because of the small number of genes under consideration, but there is no theoretical

obstacle to implement bootstrap in non-hierarchical clustering procedures.

Co-ordinate gene expression reflects the modular organization of metabolic pathways

Co-expression of genes involved in a particular metabolic pathway has been suggested by numerous studies, but most of them focused on a restricted number of genes and experimental conditions. The high number of experimental conditions in the AtGenExpress dataset allows to generalize this notion and to show that this co-ordinate expression is basically maintained throughout plant development and responses to stress. For instance, we show here that flavonoid biosynthetic genes form three distinct clusters. Earlier studies have evidenced such a co-ordination of flavonoid biosynthetic enzymes at the gene and protein levels (Pelletier and Shirley, 1996; Pelletier *et al.*, 1997; Pelletier *et al.*, 1999). They have led to a distinction between 'early' (*Chs*, *Chi*, *F3h* and *Fls*) and 'late' genes (*Ldox*, *Dfr*), which is consistent with the way the genes are grouped in our analysis. According to this classification, *F3'h* also belongs to 'early' genes, whereas *Tt19* is a 'late' gene. Finally, we define clusters that (i) contain all known enzymes of the pathway, and (ii) reflect the reorientation of metabolic fluxes from flavonol to anthocyanin biosynthesis during plant development. The third cluster, which contains both *Banyuls* and its regulators, is indicative of the onset of protoanthocyanidin synthesis in developing seeds. Therefore, the transcriptional changes of genes encoding secondary metabolism enzymes provides an integrative view of metabolic plasticity during plant development.

A co-ordinate expression of tryptophan biosynthetic genes is also apparent from our analysis. Interestingly, the genes involved in tryptophan-derived glucosinolates and camalexin synthesis fall into the same cluster in stress conditions only, pointing out a modular recruitment of the corresponding enzymes for the synthesis of these well-known stress-inducible compounds (Tsuji *et al.*, 1992; Brader *et al.*, 2001; Mikkelsen *et al.*, 2003). This observation certainly reflects that the majority of the metabolic flux in the indole pathway is re-oriented towards inducible camalexin and glucosinolate biosynthesis in stress conditions, whereas it is not true in standard growth conditions.

Co-ordinate gene expression may reflect functional specialization of paralogue genes

The interpretation of the expression pattern of the flavonoid genes is relatively straightforward, because they are not duplicated. For other pathways like phenylpropanoids, the existence of numerous paralogues certainly accounts for a more fluctuating composition of gene clusters (as evidenced by lower bootstrap values). However, phenylpropanoid biosynthetic genes also exhibit a co-ordinate expression throughout plant development and responses to stress. Moreover, the presence of only certain isoforms in given clusters gives hints about functional specialization of duplicated genes, as exemplified by 4-coumaroyl CoA ligases (4CL): the phenylpropanoid cluster contains 4CL1 and 4CL5, whereas 4CL3 is grouped with 'early' flavonoid genes. As already proposed by Ehlting *et al.* (1999), these results point out 4CL as key regulatory enzymes, which are responsible for the orientation of metabolic fluxes into either lignin precursors or flavonoid biosynthesis.

Co-ordinate expression with known enzymes allows functional prediction of uncharacterized genes

We are interested in the functional characterization of *Ugt*s, a multigenic family encoding enzymes involved in the conjugation of secondary metabolites to sugars. A major difficulty for assigning them a function is the diversity of substrates that are glycosylated *in planta*, and the lack of an obvious relationship between the protein primary sequence and its substrate specificity. In the last few years, a systematic *in vitro* approach has been undertaken, based on heterologous expression of all the recombinant enzymes of the family, and systematic testing of the most relevant substrates on each clone (Jackson *et al.*, 2001; Lim *et al.*, 2001; Lim *et al.*, 2002). Although fruitful, this strategy is labour-intensive and the broad substrate specificity displayed by the enzymes *in vitro* sometimes hampers the identification of their real target *in vivo*. To by-pass this, we looked for co-regulation of *Ugt*s with genes previously demonstrated to act in metabolic pathways of special interest. In itself, gene co-expression is not a proof of co-regulation, which in turn does not necessarily imply the involvement in the same function. However, in the light of our results, co-expression

of uncharacterized genes and known enzymes involved in the same pathway can be expected, and be used as a first indication to predict gene function. The combination of this information with primary sequence analysis should allow to reinforce this hypothesis, which should further be validated using reverse genetics approaches.

To illustrate this approach, we tentatively assign a putative role for three *Ugt* genes. *Ugt75c1* is co-ordinately expressed with anthocyanin biosynthetic genes, and displays a high sequence similarity with the known 5-*O*-anthocyanin glucosyltransferases from other species. As a general rule, glycosyltransferases belonging to the same clade often exhibit comparable regiospecificity (Vogt, 2002). Furthermore, the most abundant anthocyanins in *Arabidopsis* stems and leaves are glucosylated on their 5-OH position (Bloor and Abrahams, 2002), which implies that an enzyme catalyzing this reaction must be present. These converging data strongly suggest that *Ugt75c1* may encode an anthocyanin-5-*O*-glucosyltransferase.

Very recently, Grubb *et al.* (2004) reported in an independent study very strong *in vitro* and *in vivo* evidences showing that UGT74B1 is involved in desulfoglucosinolate synthesis in planta. In particular, they found that *Ugt74b1* knock-out only partially depletes the plant in glucosinolates, which is in favour of another glucosyltransferase acting on thiohydroximates in planta. However, they did not investigate this question any further. Our analysis points out UGT74C1 as the only member of the UGT74 group susceptible to fulfil this role. Furthermore, the co-regulation pattern of *Ugt74b1* and *Ugt74c1* suggests that the corresponding enzymes may preferentially metabolize aromatic or methionine-derived substrates, respectively. Finally, the knock-out of *Ugt74b1* also affects basal levels of aliphatic glucosinolates, which at first sight is not consistent with our hypothesis of a substrate specialization between UGT74B1 and –74C1. However, our conclusions are based on expression pattern divergence of both genes, and we do not exclude partial mutual complementation in tissues where both pathways (aliphatic and aromatic glucosinolates synthesis) co-exist. In fact, in our developmental dataset, all glucosinolate-related genes fall into a unique cluster. Subsequently, these results obtained with non-elicited seedlings are fully compatible with the possibility that UGT74B1 is the major enzyme responsible for stress-induced aromatic glucosinolate synthesis *in planta* and that UGT74C1 could have specialized into the developmentally regulated aliphatic glucosinolate synthesis (Petersen *et al.*, 2002).

Evidence for different evolutionary scenarios of metabolic pathways

Gene duplication is a major process ensuring the diversification of metabolic pathways, and it certainly accounts for the incredible diversity of plant secondary metabolites. Theoretically, a new metabolic pathway can result from different mechanisms: (i) duplicated genes may encode consecutive steps of a same pathway (Figure 5A), as illustrated by the *Bx* genes of maize involved in benzoxazinoid synthesis (Frey *et al.*, 1997); (ii) paralogues can be recruited to catalyse a new reaction (Figure 5B), as exemplified by the *Aop2* and *Aop3* genes (Kliebenstein *et al.*, 2001). Independent repetitions of this process may yield a new pathway; (iii) a whole pathway can be duplicated, followed by divergence either of substrate specificity or gene expression pattern (Figure 5C). Though not demonstrated yet, the existence of a

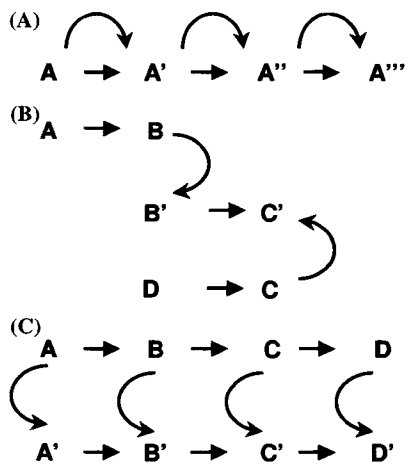

Figure 5. Evolving new pathways from duplicated genes: possible mechanisms. Paralogue enzymes are designed by the same letter and distinguished by apostrophes. Straight arrows represent chemical reactions, whereas curved arrows represent duplications of the corresponding genes. (A) Recruitment of paralogues in successive reactions, (B) Recruitment of independently duplicated genes to form a new pathway, (C) Duplication of the pathway as a whole, followed by substrate (or expression) divergence.

so-called modular evolution of metabolic pathways is highly expectable in plants because of their propensity to polyploidy. In this case, one would expect to find clusters of co-expressed genes, each containing one isoform of each enzyme, as is apparent for the glucosinolate biosynthetic genes.

Co-expression in itself is a poor indicator of the involvement in the same function, but a genome-wide analysis of duplications in *Caenorhabditis elegans* and *Drosophila melanogaster* show that the encoded proteins of co-expressed genes pairs are very likely (97%) to be part of the same pathway after parallel gene duplication (van Noort *et al.*, 2003). In *Arabidopsis*, groups of genes duplicated through polyploidy and displaying concerted divergence of their expression profiles have been evidenced, but their functional significance is not established (Blanc and Wolfe, 2004). Conversely, we propose here that a set of co-expressed paralogue genes involved in parallel metabolic pathways potentially derives from a single large-scale duplication event encompassing all genes of the pathway, rather than from the sequential recruitment of independently duplicated genes (Figure 4B). This large-scale event was later followed by several rounds of duplications, among which only a few copies have been retained. For instance, the tandem-duplicated *Cyp79f1* and *Cyp79f2* acquired a new substrate specificity towards the lateral chain length of glucosinolates (Chen *et al.*, 2003). This illustrates that plant metabolic pathway evolution most likely takes place by combining all possible mechanisms described above.

Concluding remarks

More generally, our work illustrates the potential of microarray data-mining as a tool to draw precise hypotheses on gene function and to orient functional genomics studies. Indeed, a major frustration emerging from reverse genetics is the lack of any obvious phenotype of many knock-out mutants (Bouché and Bouchez, 2001). Although functional redundancy of highly similar genes could account for some of these failures, a highly probable explanation is that the paucity of biological knowledge on the targeted genes prevents adequate prediction and testing of truly existing phenotypes. We show here that the combination of classical biochemical and genetic knowledge with microarray data allows to point out a few candidate genes the function of which can be tested in depth. A similar approach can be developed on any aspect of plant physiology to assign a putative function to uncharacterized genes and contribute significant advances in gene annotation.

Note addea in prrof

After the submission of this manuscript, T. Tohge *et al.* demonstrated that UGT75C1 encodes a 5-*O*-anthocyanin-GT using a combined reverse genetics and metabolomics approach (Plant J. 42: 218–235).

Acknowledgements

We are very grateful to the AtGenExpress consortium and other researchers who made their microarray data freely available to the community. V. Thareau is acknowledged for his help with handling the data files. We also thank Dr G. Tcherkez and M.-S. Remigereau for very stimulating discussions, and Dr F.C. Küpper (The Scottish Association for Marine Science, U.K.) and Prof M. Dron for critical reading of the manuscript.

References

Bak, S. and Feyereisen, R. 2001. The involvement of two P450 enzymes, CYP83B1 and CYP83A1, in auxin homeostasis and glucosinolate biosynthesis. Plant Physiol. 127: 108–118.

Baudry, A., Heim, M.H., Dubreucq, B., Caboche, M., Weisshaar, B. and Lepiniec, L. 2004. TT2, TT8, and TTG1 synergistically specify the expression of *Bayals* and proto-anthocyanidin synthesis in *Arabidopsis thaliana*. Plant J. 39: 366–380.

Blanc, G., Hokamp, K. and Wolfe, K.H. 2003. A recent polyploidy superimposed in older large-scale duplication events in the *Arabidopsis* genome. Genome Res 13: 137–144.

Blanc, G. and Wolfe, K.H. 2004. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. Plant Cell 16: 1679–1691.

Bloor, S.J. and Abrahams, S. 2002. The structure of the major anthocyanin in *Arabidopsis thaliana*. Phytochemistry 59: 343–346.

Borevitz, J.O., Xia, Y., Blount, J., Dixon, R.A. and Lamb, C. 2000. Activation tagging identifies a conserved MYB regulator of phenylpropanoid biosynthesis. Plant Cell 12: 2383–2393.

Bouché, N. and Bouchez, D. 2001. *Arabidopsis* gene knock-out: phenotypes wanted. Curr. Opin. Plant Biol. 4: 111–117.

Bowers, J.E., Chapman, B.A., Rong, J. and Paterson, A.H. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422: 433–438.

Brader, G., Tas, E. and Palva, E.T. 2001. Jasmonate-dependent induction of indole glucosinolates in *Arabidopsis* by culture filtrates of the nonspecific pathogen *Erwinia amylovora*. Plant Physiol. 126: 849–860.

Cannon, S.B., Mitra, A., Baumgarten, A., Young, N.D. and Georgiana, M. 2004. The role of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. BMC Plant Biol. 4: 10.

Chen, S., Glawischnig, E., Jorgensen, K., Naur, P., Jorgensen, B., Olsen, C.E., Hansen, C.H., Rasmussen, H., Pickett, J. and Halkier, B.A. 2003. CYP79F1 and CYP79F2 have distinct functions in the biosynthesis of aliphatic glucosinolates in *Arabidopsis*. Plant J. 33: 923–937.

Ehlting, J., Büttner, D., Wang, Q., Douglas, C.J., Somssich, I.E. and Kombrink, E. 1999. Three coumarate:coenzyme A ligases in *Arabidopsis thaliana* represent two evolutionarily divergent classes in Angiosperms. Plant J. 19: 9–20.

Felsenstein, J. 1989. PHYLIP – Phylogeny inference package (Version 3.2). Cladistics 5: 164–166.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.-l., and Postlethwait, J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151: 1531–1545.

Freudenberg, K. and Harkin, J.M. 1963. The glucosides of cambial sap of spruce. Phytochemistry 2: 189–193.

Frey, M., Chomet, P., Glawischnig, E., Stettner, C., Grün, S., Winklmair, A., Eisenreich, W., Bacher, A., Meeley, R.B., Briggs, S.P., Simcox, K. and Gierl, A. 1997. Analysis of a chemical plant defense mechanism in grasses. Science 277: 696–699.

Frey, M., Stettner, C., Schmelz, E.A., Tumlinson, J.H. and Gierl, A. 2000. An herbivore elicitor activates the gene for indole synthesis emission in maize. Proc. Natl. Acad. Sci. USA 97: 14801–14806.

Grubb, C.D., Zipp, B., Ludwig-Müller, J., Masuno, M.N., Molinski, T.F. and Abel, S. 2004. *Arabidopsis* glucosyltransferase ugt74b1 functions in glucosinolate biosynthesis and auxin homeostasis. Plant J. 40: 893–908.

Gu, Z., Rifkin, S.A., White, K.P. and Li, W.-H. 2004. Duplicate genes increase gene expression diversity within and between species. Nat. Genet. 36: 577–579.

Gu, Z., Steinmetz, L.M., Gu, X., Scharfe, C., Davis, R.W. and Li, W.-H. 2003. Role of duplicate genes in genetic robustness against null mutations. Nature 421: 63–66.

Hansen, C.H., Wittstock, U., Olsen, C.E., Hick, A.J., Pickett, J.A. and Halkier, B.A. 2001. Cytochrome P450 CYP79F1 from *Arabidopsis* catalyzes the conversion of dihomomethionine and trihomomethionine to the corresponding aldoximes in the biosynthesis of aliphatic glucosinolates. J. Biol. Chem. 276: 11078–11085.

Hemm, M.R., Ruegger, M.O. and Chapple, C. 2003. The *Arabidopsis ref2* mutant is defective in the gene encoding CYP83A1 and shows both phenylpropanoid and glucosinolate phenotypes. Plant Cell 15: 179–194.

Hong, Z., Zhang, Z., Olson, J.M. and Verma, D.P.S. 2001. A novel UDP-glucose transferase is part of the callose synthase complex and interacts with phragmoplastin at the forming cell plate. Plant Cell 13: 769–779.

Imayama, T., Yoshihara, N., Fukuchi-Mizutani, M., Tanaka, Y., Ino, I. and Yabuya, T. 2004. Isolation and characterization of a cDNA clone of UDP-glucose:anthocyanin 5-*O*-glucosyltransferase in *Iris hollandica*. Plant Sci. 167: 1243–1248.

Jackson, R.G., Kowalczyk, M., Li, Y., Higgins, G., Ross, J., Sandberg, G. and Bowles, D.J. 2002. Over-expression of an *Arabidopsis* gene encoding a glucosyltransferase of indole-3-acetic acid: phenotypic characterisation of transgenic lines. Plant J. 32: 573–583.

Jackson, R.G., Lim, E.-K., Li, Y., Kowalczyk, M., Sandberg, G., Hoggett, J., Ashford, D.A. and Bowles, D.J. 2001. Identification and biochemical characterization of an *Arabidopsis* indole-3-acetic acid glucosyltransferase. J. Biol. Chem. 276: 4350–4356.

Johnson, C.S., Kolevski, B. and Smyth, D.R. 2002. *TRANSPARENT TESTA GLABRA 2*, a trichome and seed coat development gene of *Arabidopsis*, encodes a WRKY transcription factor. Plant Cell 14: 1359–1375.

Jones, P.R., Messner, B., Nakajima, J., Schäffner, A.R. and Saito, K. 2003. UGT73C6 and UGT78D1, glycosyltransferases involved in flavonol glycoside biosynthesis in *Arabidopsis thaliana*. J. Biol. Chem. 278: 43910–43918.

Jones, P.R. and Vogt, T. 2001. Glycosyltransferases in secondary plant metabolism: tranquilizers and stimulant controllers. Planta 213: 164–174.

Kellis, M., Birren, B.W. and Lander, E.S. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature 428: 617–624.

Kita, M., Hirata, Y., Moriguchi, T., Endo-Inagaki, T., Matsumoto, R., Hasegawa, S., Suhayda, C.G. and Omura, M. 2000. Molecular cloning and characterization of a novel gene encoding limonoid UDP-glucosyltransferase in *Citrus*. FEBS Lett. 469: 173–178.

Kliebenstein, D.J., Lambrix, V.M., Reichelt, M., Gershenzon, J. and Mitchell-Olds, T. 2001. Gene duplication in the diversification of secondary metabolism: tandem 2-oxoglutarate-dependent dioxygenases control glucosinolate biosynthesis in *Arabidopsis*. Plant Cell 13: 681–693.

Kreps, J.A., Wu, Y., Chang, H.-S., Zhu, T., Wang, X. and Harper, J.F. 2002. Transcriptome changes for *Arabidopsis* in response to salt, osmotic, and cold stress. Plant Physiol. 130: 2129–2141.

Lim, E.-K. and Bowles, D.J. 2004. A class of plant glycosyltransferases involved in cellular homeostasis. EMBO J. 23: 2915–2922.

Lim, E.-K., Doucet, C.J., Li, Y., Elias, L., Worrall, D., Spencer, S.P., Ross, J. and Bowles, D.J. 2002. The activity of *Arabidopsis* glucosyltransferases toward salicylic acid, 4-hydroxybenzoic acid, and other benzoates. J. Biol. Chem. 277: 586–592.

Lim, E.-K., Li, Y., Parr, A., Jackson, R., Ashford, D.A. and Bowles, D.J. 2001. Identification of glucosyltransferase genes involved in sinapate metabolism and lignin synthesis in *Arabidopsis*. J. Biol. Chem. 276: 4344–4349.

Lynch, M. and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1155.

Marillia, E.-F., MacPherson, J.M., Tsang, E.W.T., Van Audenhove, K., Keller, W.A. and GrootWassink, J.W.D. 2001. Molecular cloning of a *Brassica napus* gene and its expression in *Escherichia coli*. Physiol. Plant 113: 176–184.

Martz, F., Maury, S., Pinçon, G. and Legrand, M. 1998. cDNA cloning, substrate specificity and expression study of tobacco caffeoyl-CoA 3-*O*-methyltransferase, a lignin biosynthetic enzyme. Plant Mol. Biol. 36: 427–437.

Mikkelsen, M.D., Petersen, B.L., Glawischnig, E., Jensen, A.B., Andreasson, E. and Halkier, B.A. 2003. Modulation of

CYP79 genes and glucosinolates profiles in *Arabidopsis* by defense signaling pathways. Plant Physiol. 131: 298–308.

Mobley, E.M., Kunkel, B.N. and Keith, B. 1999. Identification, characterization and comparative analysis of a novel chorismate mutase gene in *Arabidopsis thaliana*. Gene 240: 115–123.

Niyogi, K.K. and Fink, G.R. 1992. Two anthranilate synthase genes in *Arabidopsis*: defense-related regulation of the tryptophan pathway. Plant Cell 4: 721–733.

Paquette, S., Moller, B.L. and Bak, S. 2003. On the origin of family 1 glycosyltransferases. Phytochemistry 62: 399–413.

Pelletier, M.K., Burbulis, I.E. and Winkel-Shirley, B. 1999. Disruption of specific flavonoid genes enhances the accumulation of flavonoid enzymes and end-products in *Arabidopsis* seedlings. Plant Mol. Biol. 40: 45–54.

Pelletier, M.K., Murrell, J.R. and Shirley, B.W. 1997. Characterization of flavonol synthase and leucoanthocyanidin dioxygenase genes in *Arabidopsis*. Plant Physiol 113: 1437–1445.

Pelletier, M.K. and Shirley, B.W. 1996. Analysis of flavanone 3-hydroxylase in *Arabidopsis* seedlings. Plant Physiol 111: 339–345.

Petersen, B.L., Chen, S., Hansen, C.H., Olsen, C.E. and Halkier, B.A. 2002. Composition and content of glucosinolates in developing *Arabidopsis thaliana*. Planta 214: 562–571.

Quiel, J.A. and Bender, J. 2003. Glucose conjugation of anthranilate by the *Arabidopsis* UGT74F2 glucosyltransferase is required for tryptophan mutant blue fluorescence. J Biol Chem 278: 6275–6281.

Redman, J.C., Haas, B.J., Tanimoto, G. and Town, C.D. 2004. Development and evaluation of an *Arabidopsis* whole genome Affymetrix probe array. Plant J. 38: 545–561.

Reintanz, B., Lehnen, M., Reichelt, M., Gershenzon, J., Kowalczyk, M., Sandberg, G., Godde, M., Uhl, R. and Palme, K. 2001. *bus*, a bushy *CYP79F1* knock-out mutant with abolished synthesis of short-chain aliphatic glucosinolates. Plant Cell 13: 351–367.

Rohde, A., Morreel, K., Ralph, J., Goeminne, G., Hostyn, V., de Rycke, R., Kushnir, S., Van Doorsselaere, J., Joseleau, J.-P., Vuylsteke, M., Van Driessche, G., Van Beumen, J., Messens, E. and Boerjan, W. 2004. Molecular phenotyping of the *pal1* and *pal2* mutants of *Arabidopsis thaliana* reveals far-reaching consequences on phenylpropanoid, amino acid, and carbohydrate metabolism. Plant Cell 16: 2749–2771.

Saeed, A., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M., Sturn, A., Snuffin, M., Rezantsev, A., Popov, D., Ryltsov, A., Kostukovich, E., Borisovsky, I., Liu, Z., Vinsavich, A., Trush, V. and Quackenbush, J. 2003. TM4: a free, open-source system for microarray data management and analysis. Biotechniques 24: 374–378.

Schwab, W. 2003. Metabolome diversity: too few genes, too many metabolites?. Phytochemistry 62: 837–849.

Sembdner, G., Atzorn, R. and Schneider, G. 1994. Plant hormone conjugation. Plant. Mol. Biol. 26: 1459–1481.

Taguchi, G., Ubukata, T., Hayashida, N., Yamamoto, H. and Okazaki, M. 2003. Cloning and characterization of a glucosyltransferase that reacts on 7-hydroxyl groups of flavonol and 3-hydroxyl group of coumarin from tobacco cells. Arch. Biochem. Biophys. 420: 95–102.

Tsuji, J., Jackson, E.P., Douglas, A.G., Hammerschmidt, R. and Somerville, S.C. 1992. Phytoalexin accumulation in *Arabidopsis thaliana*. Plant Physiol. 98: 1304–1309.

Turner, J.G., Ellis, C. and Devoto, A. 2002. The jasmonate signal pathway. Plant Cell Suppl. S153–S164.

Van der Laan, M.J. and Bryan, J. 2001. Gene expression analysis with the parametric bootstrap. Biostatistics 2: 445–461.

van Noort, V., Snel, B. and Huynen, M.A. 2003. Predicting gene function by conserved co-expression. Trends. Genet. 19: 238–242.

Vogt, T. 2002. Substrate specificity and sequence analysis define a polyphyletic origin of betanidin 5- and 6-*O*-glucosyltransferase from *Dorotheanthus bellidiformis*. Planta 214: 492–495.

Wray, G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V. and Romano, L.A. 2003. The evolution of transcriptional regulation in Eukaryotes. Mol. Biol. Evol. 20: 1377–1419.

Yamazaki, M., Gong, Z., Fukuchi-Mizutani, M., Fukui, Y., Tanaka, Y., Kusumi, T. and Saito, K. 1999. Molecular cloning and biochemical characterization of a novel anthocyanin 5-*O*-glucosyltransferase by mRNA differential display for plant forms regarding anthocyanin. J. Biol. Chem. 274: 7405–7411.

Zhang, K. and Zhao, H. 2000. Assessing reliability of gene clusters from gene expression data. Funct. Integr. Genom. 1: 156–173.

Zhang, P., Gu, Z. and Li, W.-H. 2003. Different evolutionary patterns between young duplicate genes in the human genome. Genome Biol. 4: R56.

Zhao, J. and Last, R.L. 1996. Coordinate regulation of the tryptophan biosynthetic pathway and indolic phytoalexin accumulation in *Arabidopsis*. Plant Cell 8: 2235–2244.

Zhao, J., Williams, C.C. and Last, R.L. 1998. Induction of *Arabidopsis* tryptophan pathway enzymes and camalexin by amino acid starvation, oxidative stress, and an abiotic elicitor. Plant Cell 10: 359–370.