BMC Plant Biology

BioMed Central

Methodology article

Conservation and diversity of gene families explored using the CODEHOP strategy in higher plants

Marc Morant, Alain Hehn and Danièle Werck-Reichhart*

Address: Department of Plant Stress Response, Institute of Plant Molecular Biology, CNRS-UPR 2357, Université Louis Pasteur, 28 rue Goethe, 67083 Strasbourg, France

E-mail: Marc Morant - marc.morant@ibmp-ulp.u-strasbg.fr; Alain Hehn - alain.hehn@ibmp-ulp.u-strabg.fr; Danièle Werck-Reichhart* - daniele.werck@ibmp-ulp.u-strasbg.fr

*Corresponding author

Abstract

Background: Availability of genomewide information on an increasing but still limited number of plants offers the possibility of identifying orthologues, or related genes, in species with major economical impact and complex genomes. In this paper we exploit the recently described CODEHOP primer design and PCR strategy for targeted isolation of homologues in large gene families.

Results: The method was tested with two different objectives. The first was to analyze the evolution of the *CYP98* family of cytochrome P450 genes involved in 3-hydroxylation of phenolic compounds and lignification in a broad range of plant species. The second was to isolate an orthologue of the sorghum glucosyl transferase *UGT85B1* and to determine the complexity of the *UGT85* family in wheat. P450s of the *CYP98* family or closely related sequences were found in all vascular plants. No related sequence was found in moss. Neither extensive duplication of the *CYP98* genes nor an orthologue of *UGT85B1* were found in wheat. The *UGT85A* subfamily was however found to be highly variable in wheat.

Conclusions: Our data are in agreement with the implication of CYP98s in lignification and the evolution of 3-hydroxylation of lignin precursors with vascular plants. High conservation of the *CYP98* family strongly argues in favour of an essential function in plant development. Conversely, high duplication and diversification of the *UGT85A* gene family in wheat suggests its involvement in adaptative response and provides a valuable pool of genes for biotechnological applications. This work demonstrates the high potential of the CODEHOP strategy for the exploration of large gene families in plants.

Background

Plants have evolved extremely diversified gene families as tools to cope with a harsh environment. Some of these families such as cytochromes P450 and UDP-glucosyl-transferases (UGT) reflect the extraordinary biochemical versatility of plants and across plant species, and represent a very valuable source of genes for biotechnologies. Both gene families offer a huge potential for bioremediation and control of crop and weed pesticide tolerance [1–3], but obviously also for industrial applications. P450s, con-

sidered as the most versatile catalysts known [4], usually activate dioxygen and transfer one of its atoms into various substrates, but also catalyze a great diversity of reactions ranging from C-C and C=N bond cleavage, phenolic coupling, dehydration, dehydrogenation, isomerizations to reduction [5]. Many of these reactions are important for the biosynthesis of hormones, drugs, pigments, aromas, biopolymer building blocks and defense molecules [6,7]. Glycosyltransferases are also essential for the production of natural compounds since they control their solubility, stability, transport, storage and sometimes also their bioactivity [8,9]. Should some of this potential become directly accessible through genomewide sequencing, extensive information is restricted to model plants, usually with a small genome, or to plants with a major economical interest. Exploitation of this knowledge to target genes of other plants that need to be studied or engineered, or to explore gene families in plants with specific biosynthetic capacities is an objective for the next several years.

With the growing availability of gene sequences plus information regarding their diversity and phylogeny, increasingly sophisticated PCR techniques have been developed to target gene families. Plant P450s are low abundant membrane-bound and unstable proteins, usually difficult to purify. For this reason, early on, several groups attempted isolation of P450 genes on the basis of the most conserved consensus regions, after generating probes by conventional PCR at low stringency [10–13]. This approach was later refined and used by several other groups for isolation of P450 genes in various plant species [e.g. [14–18]]. It proved successful in many cases, although only leading to a small number of highly expressed and related P450 families. A significant step forward resulted from coupling degenerate PCR with a heme binding primer and differential display of the amplified fragments, an approach that allowed effective identification of nine P450 genes responsive to elicitor treatment of soybean cell cultures [19] and 21 unique P450 genes in *Taxus* cells induced for taxol production [20]. A carefully controlled and strongly differential system is however needed for such an approach. Another interesting improvement was recently reported that involves use of nested primers to increase PCR selectivity [21]. However, the major limitation of all the strategies reported so far is that they did not take into account the huge diversity and low conservation of P450s recently revealed by genome sequencing in higher plants, and allowed neither focused gene selection nor isolation of the most divergent P450 clades, i.e. no systematic exploration of the P450 superfamily in highly divergent species.

In this paper we report on the high potential of the recently described COnsensus-DEgenerate Hybrid Oligonucleotide Primers (CODEHOP) strategy [22] of primer design,

ensuring optimal match and PCR amplification focused on very short conserved sequences, for the isolation of orthologues in evolutionarily distant species and for the focused or systematic exploration of gene families in plants with a very large genome. The method was tested to analyze both the duplication and conservation of the CYP98 family of P450 genes in many plant species. This family was recently suggested to play an essential role in lignification and plant development [23–25]. The same approach was also used for the analysis of the UGT85 family in wheat.

Results
Chasing the CYP98 genes in wheat
The CYP98 family of cytochrome P450 genes encodes the 3'-hydroxylases of coumaroyl esters, which catalyze an essential step in the synthesis of lignin monomers and chlorogenic acid [23–25]. CYP98 activity is also needed for the biosynthesis of many phenolic flavouring compounds such as eugenol, safrole or vanilin. Engineering the expression of this family of P450s has important agro-industrial applications, including enhancement of plant defense and modification of lignin composition to improve forage digestibility and wood pulping [26,27]. Access to CYP98 genes from major crop and forage plants and most common woody species is a necessary step for modifying their expression. If some partial sequences are made available by EST sequencing of a limited number of species, they do not reflect the whole range of gene isoforms expressed in a plant or plant tissue, especially for large genome plants.

Our first aim was thus to test if it was possible to detect several genes belonging to the *CYP98* family expressed in the seedlings from wheat, a major crop plant with a very large genome. When this work was initiated, few *CYP98* sequences were available, some of them only partial sequences arising from ESTs. To optimize primer selection, with a bias in favour of monocot genes, available ESTs from rice and maize were aligned with the full-length sequences from sorghum [17], soybean [28] and *Arabidopsis thaliana*. In the latter case, *A. thaliana* genome sequencing had revealed three genes. Two of them (*CYP98A8* and *CYP98A9*: function unknown) were closely related and clearly divergent from the third, *CYP98A3*, recently shown to encode a coumaroyl esters 3'-hydroxylase. *CYP98A3* seemed to be the orthologue of the sequences isolated from sorghum, rice, maize and soybean. Primers were designed using the CODEHOP strategy. Three sense (P98a, d and c) and one reverse (P98cr) primer were selected (Figure 1) so as to avoid the strong consensus regions common to other P450 families such as the highly conserved PERF motif. To further ensure high primer selectivity, touch-down gene selection PCR was conducted starting with a high (70°C) annealing temperature.
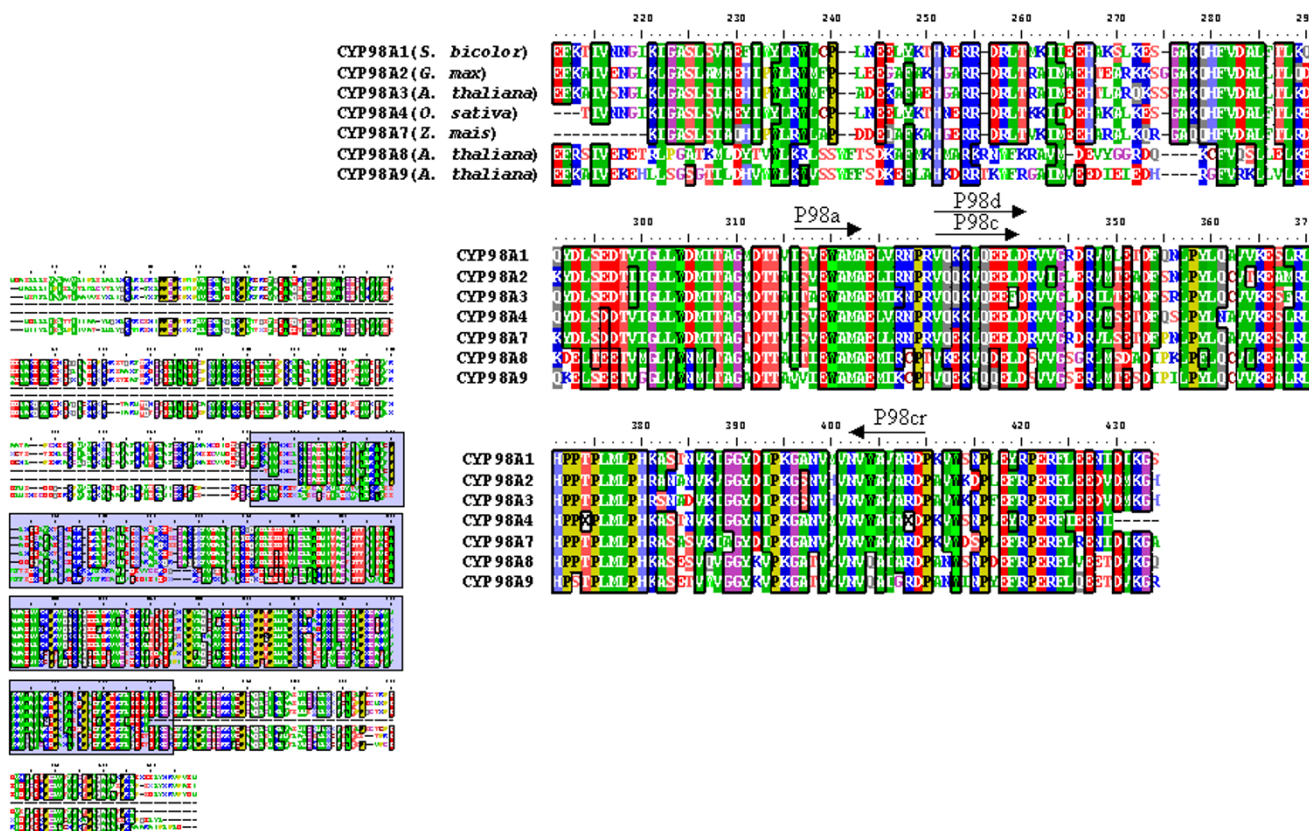
Figure I
Location of the P98 primers on the CYP98 alignment used for primer design. Only the region overlapping available monocots ESTs was used for primer selection in order to introduce a bias for monocot sequences. This overlapping region is shown on the full-length alignment (left) shaded in blue. All sequence alignments were performed using the BioEdit program [40].

BLASTp analysis, performed with the consensus protein sequence corresponding to each primer on a local plant P450 library, indicated that the sense primers were more specific than the reverse, and likely to control amplification selectivity (Table 1I). As predicted by the BLAST analysis, the P98c/P98cr pair was the most specific, and led to the amplification of two different but closely related *CYP98* fragments from wheat seedlings cDNA libraries. A single band of the expected size was obtained on an agarose gel that was eluted and subcloned in a 3'-T overhang vector. Out of 19 sequenced clones, 11 corresponded to *CYP98A11*, 7 to *CYP98A10*, and one to a non-CYP sequence (Figure 2). As predicted as well, the P98d/P98cr pair was the second most effective. In addition to *CYP98A10* and *CYP98A11*, it also amplified a clearly divergent *CYP98* gene, *CYP98A12*. Out of 14 amplified fragments sequenced 13 coded for CYP98s. The P98a primer is predicted to match a larger number of P450 families (Table 1). Used together with P98cr, it amplified more non-*CYP* and non-*CYP98 CYP* sequences than *CYP98*s

(*CYP98A11* and *CYP98A10*). Two of the *CYP* sequences however were closely related to *CYP98*s and *CYP76*s. Initiation of the touch-down PCR at a lower temperature and analysis of the amplified fragments did not reveal additional *CYP98* sequences.

The CODEHOP strategy thus appears well suited for the very focused to broader exploration of gene families in plants with a large genome, depending on primer selectivity. Selectivity of each set of primers can be predicted by a BLAST analysis. In young wheat seedlings, focused CODEHOP screening allowed the detection of three clearly different *CYP98* genes. All three genes are apparently related to *A. thaliana CYP98A3*.

Isolation of CYP98 ortho/homologues in other plant species
The second step of this investigation was aimed at testing the possibility to isolate *CYP98A3* orthologues in a broad range of distantly related species. The most selective prim-
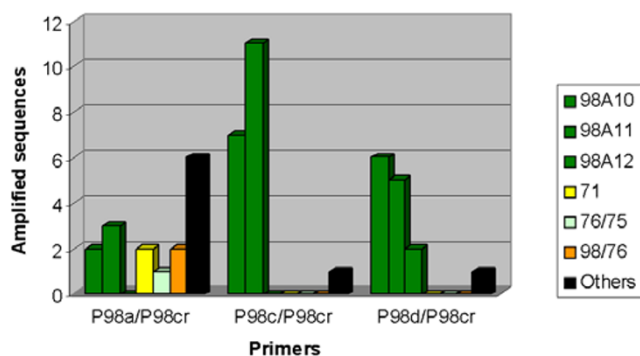
Figure 2
Proportion of CYP98 and other CYP sequences amplified with each couple of primers from the template wheat cDNA libraries. 76/75 and 98–76: BLASTp analysis of the amplified fragment assigns equal scores to CYP76 and CYP75, or CYP98 and CYP76 families. Others: non-CYP sequences.

er pair P98c/P98cr, without modification for codon usage, was first assayed with the same amplification programme or after shifting the initial annealing temperature from 70 to 65°C using various cDNA libraries, prepared from *Capsicum annuum* fruit (Solanaceae), *Ceratopteris richardii* (fern), *Coleus blumei* cell culture (Lamiaceae), *Eucalyptus globulus* xylem (Myrtaceae), *Helianthus annuus* stem and leaf (Asteraceae), *Lycopersicon esculentum* shoots (Solanaceae), *Picea abies* cell culture (Coniferales), *Pinus pinaster* stem and root (Coniferales), *Populus trichocarpa x deltoides* stem, root and leaf (Salicaceae), *Physcomitrella patens* protonemal tissue (moss). *CYP98*-related sequences were amplified from the libraries of *C. blumei*, *E. globulus*, *H. annuus*, *P. abies*, and *Populus*. After optimization of primer codon usage, a *CYP98*-like fragment was also am-

plified from the *C. richardii* cDNA library. No amplicon was obtained however for *P. patens*, *P. pinaster* or the Solanaceae, neither after optimizing codon usage, nor after further decreasing the initial temperature of the touchdown PCR.

Out of these ten representative species in a broad range of vascular plants, including fern, conifers, monocots and dicots, the CODEHOP strategy thus provided *CYP98*-related DNA sequences in seven cases (Figure 3). Due to the short size of the amplicons, it is not possible to unambiguously assign them all to the CYP98 family. Full-length sequences would be needed for such an assignment, if not catalytic activity for the proteins. To date, those were only obtained in the case of wheat, confirming gene identity and function (M. Morant, personal communication). Homology analysis of the amplified fragments is however consistent with evolutionary history of vascular plants (Figure 4). Combined with a BLAST analysis, it suggests that the *P. abies* and *C. richardii* amplified fragments could be either representative of ancestral forms of *CYP98*s or derived from a related A-type P450 family, possibly CYP81. No *CYP98*-related sequence was detected in the moss *P. patens*, which would be in agreement with the evolution of CYP98s with vascular plants. *CYP98* and related sequences are as yet also absent from *P. patens* ESTs, among which *CYP73* expected to code for cinnamate 4-hydroxylase can be found. Evolution of CYP73, involved in an upstream step in the phenylpropanoid pathway and the biosynthesis of flavonoids, is supposed to have preceded that of lignification [29].

Limits of the method
In a few cases, no amplification was obtained after decreasing the initial annealing temperature of the touchdown PCR and adapting codon usage. This occurred with libraries where CYP98 cDNAs were expected to be present,

Table 1: CODEHOP primers used for CYP98s amplification. For each primer the consensus clamp (XXX.XXX) is given in upper case and degenerate core (xxx.xxx) in lower case. y = [C, T], n = [A, G, C, T], s = [G, C], r = [A, G]. BLASTp analysis was performed with the consensus protein sequence corresponding to each primer. Hits: best scoring P450, with expectation values lower than 0.1, are listed in the order of scores.

Primer	Sequence	Protein consensus	BLASTp hits on P450 database [1]	Expected size of PCR product (bp)
Sense				
P98a	TC.ACC.GTC.GAG.tgg.gcn.atg.gc	ITVEWAMA	98/76/93/81/79/71/705	270
P98c	G.GTG.CAG.GAG.AAG.GTG.car.san.gar.yt	RVQEKVQEEL	98/78	220
P98d	G.GTG.CAG.CAG.AAG.GTC.CAG.san.gar.ytn.ga	RVQQKVQEELD	98/78/703/96/73	220
Reverse				
P98cr	C.CCG.GGC.GAT.Ggc.yyr.nac.rtt	NVWAIARD	98/75/93/76/71/78/83/92/706	

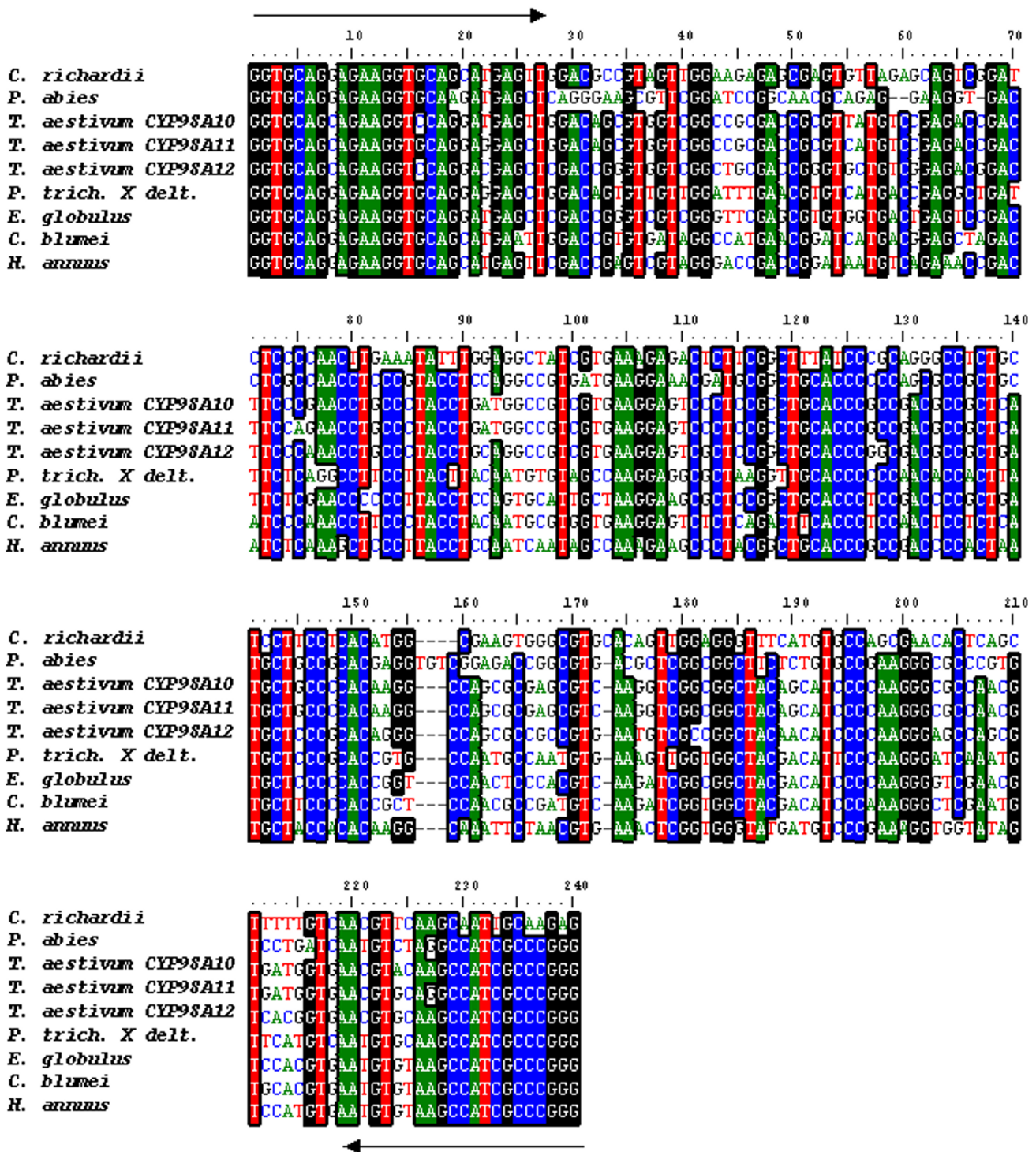[1] BLASTp parameters: Matrix PAM40/Expectation Value (E)10.

Figure 3
CYP98-like sequences amplified from the different plant species. The primer regions are shown in the alignment (delimited by arrows) to illustrate primer usage for each amplicon. In the case of *C. richardii*, a primer adapted to fern codon usage (P98Cr: 5' CTCTTGCAATTgcyyrnacrtt 3') was used for the amplification. These sequences are available from Genebank under the accessions: *C. richardii* AJ438346, *P. abies* AJ438350, *T. aestivum* CYP98A10 AJ439883, *T. aestivum* CYP98A11 AJ439884, *T. aestivum* CYP98A12 AJ439885, *Populus* AJ438351, *E. globulus* AJ438348, *C. blumei* AJ438347, *H. annuus* AJ438349.
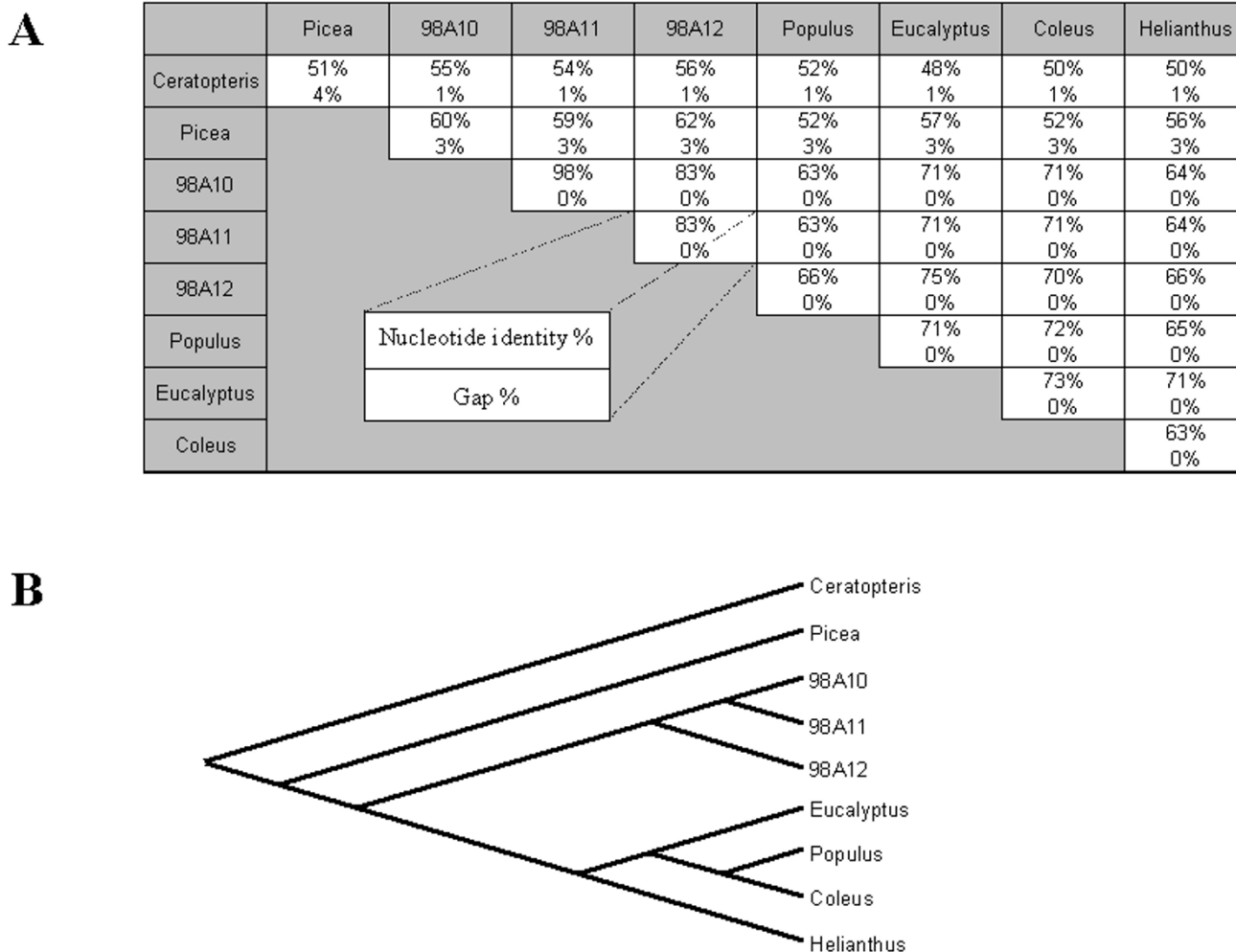
A

	Picea	98A10	98A11	98A12	Populus	Eucalyptus	Coleus	Helianthus
Ceratopteris	51% 4%	55% 1%	54% 1%	56% 1%	52% 1%	48% 1%	50% 1%	50% 1%
Picea		60% 3%	59% 3%	62% 3%	52% 3%	57% 3%	52% 3%	56% 3%
98A10			98% 0%	83% 0%	63% 0%	71% 0%	71% 0%	64% 0%
98A11				83% 0%	63% 0%	71% 0%	71% 0%	64% 0%
98A12					66% 0%	75% 0%	70% 0%	66% 0%
Populus		Nucleotide identity %				71% 0%	72% 0%	65% 0%
Eucalyptus		Gap %					73% 0%	71% 0%
Coleus								63% 0%

B

Figure 4
Identity and gap scores of the CYP sequences amplified with the P98c/cr pair in different plant species (A), and dendrogram deduced from their alignment (B). Identities and gaps were calculated using the software Genedoc [39]. 98A10, 98A11 and 98A12 are the three sequences amplified from *T. aestivum*.

for example in a library of pine stem and root, and libraries from bellpepper fruit or young tomato plants. In pine stem, CYP98 should be expressed at a high level for the synthesis of lignin monomers, while Solanaceae are described as accumulating large concentrations of hydroxycinnamic esters such as chlorogenic acid. The P98c and P98cr primers, initially designed with codon usage for wheat gene amplification, were compared to those designed with optimal codon usage for the other plant (Figure 5). P98c differed in 2 positions out of 16 in the clamp segment from the primer optimized for pine, and in 4 positions out of 16 from the primer predicted as optimal for Solanaceae. P98cr also differed in 4 positions out of 11 in the clamp for both pine and Solanaceae. Difference in co-

don usage thus seemed to provide an explanation for the failure of our first experiments. New primers using adapted codons were thus tested but did not lead to amplified fragment under any tested PCR condition.

To find an explanation, EST and cDNA sequences now available for pine and tomato, another Solanaceae, were examined and aligned with our primers and consensus sequences. This comparison to the authentic sequences revealed significant divergences from most frequent codon usage in addition to the selected consensus sequence. The assumption that the clamp segment of the primers has a very minor impact on the amplification [22] thus probably should be reconsidered. The dissimilarity present in
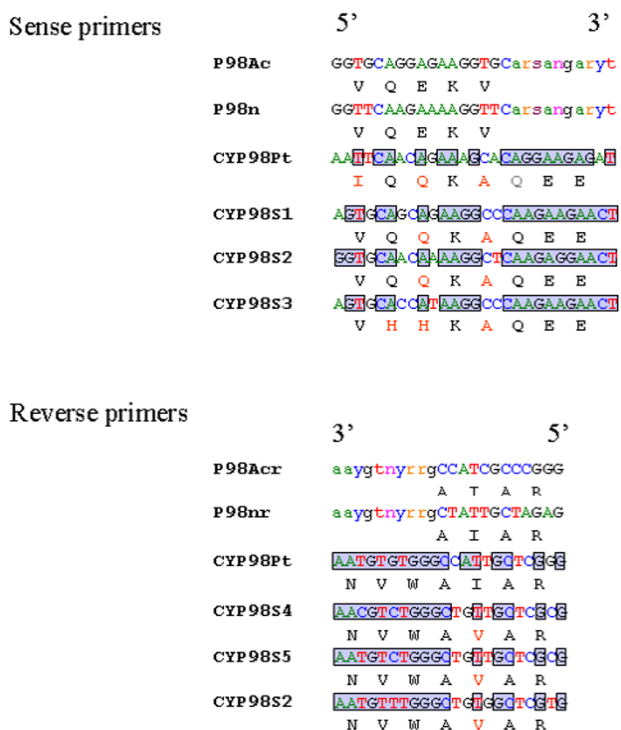
Figure 5
Deviations from most frequent codon usage and protein consensus sequences in *CYP98s* from pine and Solanaceae. Comparison of the original CODEHOP sense and reverse primers (CYP98c and CYP98cr) to the primers adapted to the codon usage of pine and Solanaceae (P98n and P98nr) and to the real sequences of recently available EST or cDNAs from *P. teada* (CYP98Pt: AY064170), or from Solanaceae (CYP98S1-6). Alignment of available CYP98 ESTs from tomato and potato reveals several subgroups with slightly different sequences in the primer regions: CYP98S1: **BG598096**/**BM11387I**/*BE450893*/*BE451611,* CYP98S2: *BE436335*/*BE432077*/*BE431773*/*BM535273,* CYP98S3: *BE451666,* CYP98S4: **BG594285**/**BG594451**/**BG598096**/ **BM11387I,** and CYP98S5: *BE450893* (*Solanum tuberosum* accessions are shown in bold characters, *L. esculentum* are italicized). Sequences matching both tested primers are boxed in blue. Divergence from amino acid consensus is shown in red.

this example is however likely to be local and thus successful amplification should be obtained by using multiple primers distributed along the full sequence.

Exploration of another gene family in wheat: UGT85

Glycosyl transferases (type 1) form another large gene family with important applications in agrochemistry, therefore, we tested the efficiency of the CODEHOP strategy for exploring the diversity of glycosyl transferases in hexaploid wheat. In this case, we decided to focus on UGTs related to the UDP-glucose:*p*-hydroxymandeloni-

trile-*O*-glucosyl transferase (*sbHMNGT* or *UGT85B1)* catalyzing the last step in the synthesis of cyanogenic glucosides, recently isolated from *S. bicolor*[30]. When this work was initiated BLASTp search revealed only 2 sequences, UGT85A2 and UGT85A3, significantly related to sbHMNGT (43% identity) in the genome of *A. thaliana.* CODEHOP primer design, based on the alignment of the 3 full-length sequences, provided five sense and five reverse primers likely to be selective of this group of UGTs (Table 2). PCR was conducted with all combinations of primers, starting touch-down PCR at 70°C. When a couple of primers amplified products of expected size, they were subcloned and analyzed. The primer couple was then discarded and the library was screened using the other primer combinations but with the initial temperature of the touch-down PCR decreased by 5°C. Successive temperature decreases led to successful amplification from 8 couples of primers. Glucosyl transferases were recently shown to be induced by the herbicide safener cloquintocet-mexyl in wheat [31]. In agreement with this report, a stronger amplification was usually obtained using a cDNA library prepared from safener-treated seedlings as a template (Figure 6). Analysis of 63 subclones led to the isolation of 18 distinct *UGT* sequences resulting from specific amplifications. Their sequences can be obtained from GeneBank under the accessions AJ438327, AJ438326, AJ438330, AJ438331, AJ438332, AJ438316, AJ438315, AJ438318, AJ438320, AJ438317, AJ438319, AJ438333, AJ438335 AJ438337, AJ438334, AJ438338, AJ438329, and AJ438328.

Not all amplified fragments are overlapping (Figure 7). It is thus not possible to determine if they correspond to more than 12 different wheat UGT genes or allelic variants. Their alignment and comparison with representative members of the different UGT families indicate that they are all phylogenetically related and derived from the same ancestor as the UGT85A subfamilly from *A. thaliana* and *sbHMNGT*. None of the fragments corresponded to an obvious orthologue of *sbHMNGT*.

Discussion

In this paper we investigated the potential of the CODEHOP strategy for targeted isolation of genes from organisms not yet submitted to extensive sequencing and for exploration of the complexity of selected gene families in large plant genomes. Special emphasis was given to wheat as representative of plants with a complex genome. The CODEHOP method proved extremely useful for the characterization of gene orthologues or homologues in a broad range of plants species. Focus of the gene search can be controlled by changing the degree of specificity of the primers (easily checked by a BLAST analysis) and by the choice of touch-down PCR temperature. In our hands, the method was successful where cDNA library screening with
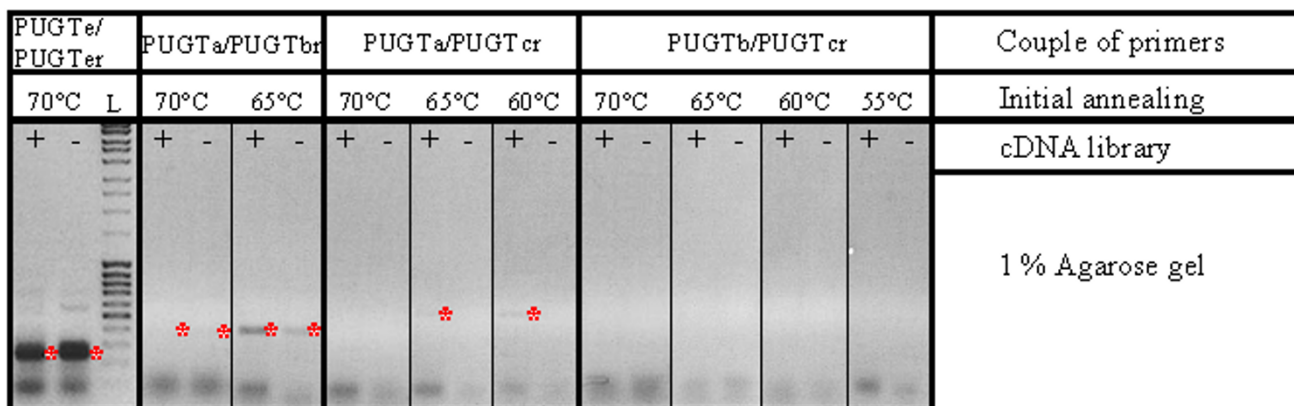
PUGTe/PUGTer			PUGTa/PUGTbr			PUGTa/PUGTcr				PUGTb/PUGTcr				Couple of primers
70°C	L		70°C	65°C		70°C	65°C	60°C		70°C	65°C	60°C	55°C	Initial annealing
+ −			+ −	+ −		+ −	+ −	+ −		+ −	+ −	+ −	+ −	cDNA library
														1 % Agarose gel

Figure 6
Result of a decrease in initial annealing temperature of the touch-down PCR on the amplification of UGT genes with different couples of primers. Location of amplicons of the expected size is indicated by an asterisk. The same couples of primers were tested with two cDNA libraries made out wheat seedlings treated (+) or not (-) with cloquintocet-mexyl and phenobarbital for inducing herbicide metabolism. Some couples of primers such as PUGTb/PUGTer did not provide any amplified fragment, even at low annealing temperatures. Some (e.g. PUGTe/PUGTer) were effective using stringent annealing temperature.

heterologous probes (e.g. cDNAs from maize for screening wheat libraries) had failed. The main advantage of this method is to rapidly provide a representative sample of either the allelic variants and recently evolved paralogues, or of the homologues of a given gene in complex genomes. Compared to other methods, a very high proportion of useful sequences (more than 90% with some primer couples) is obtained. Besides targeted search for specific expressed sequences, it allows complete exploration of the different P450 clades in a single organism, which was not possible using previously described methods. The main source of failure seems to be local sequence or codon divergence from the consensus or most frequent usage. This problem should be easily circumvented by using several primer couples chosen from different regions of the gene. Using the CODEHOP strategy on subgroups of phylogenetically related genes, families or clades, within large superfamilies such as UGTs or P450s is a powerful approach for exploring their complexity in various genomes. It is a very effective tool for the construction of expression libraries for agrochemical and other industrial applications. It can also be used for identifying genes from a given subgroup expressed at a specific stage of development.

Some plant P450 families result from extensive duplications, some of them forming clusters of up to 13 genes in *Arabidopsis*[7]. Some P450 families or subfamilies are found in only subsets of plant species. Extensive search for CYP98 genes expressed in wheat seedlings did not reveal more than 3 different sequences, all related to *A. thaliana* CYP98A3. The bread wheat genome is hexaploid and results of successive hybridizations and rearrangements

[32]. This suggests that the three *CYP98A10, A11 and A12* genes may correspond each to one of the three wheat genomes, with *CYP98A12* which is the most divergent possibly resulting from the most recent genome introduction. This hypothesis is currently investigated. In agreement with the low number of CYP98s found expressed in wheat, no extensive duplication of CYP98 was detected in other plant species. This observation, together with the high conservation of the CYP98As gene across evolution argue for essential functions of the CYP98 genes in higher plant development. Accordingly, strongly impaired growth and fertility are observed in *cyp98A3* mutants [[24], S. Goepfert, personal communication]. Similar conservation is observed for other P450 genes participating in the early phenylpropanoid pathway and hormone homeostasis.

No obvious homologue of the *sbHMNGT* gene was found in wheat. This may be connected to the fact that large amounts of dhurrin are not reported to accumulate in wheat, since sbHMNGT is described as showing a strong preference for mandelonitrile substrates [30]. A large number of related genes, all belonging to the UGT85A subfamily were however detected. This is not surprising considering that 6 UGT85A genes have been reported in the small *Arabidopsis* genome [33]. Such a duplication of genes in this family probably reflects an adaptative evolution and their implication in some type of stress/defense response rather than a developmentally essential function. Significantly, a large number of UGT85A-related sequences are found among wheat ESTs isolated from wheat challenged with pathogens. Recombinant expression of

Table 2: CODEHOP primers used for UGTs amplification. y = [C, T], n = [A, G, C, T], r = [A, G], w = [A, T], k = [G, T]. BLAST analysis of protein consensus sequences corresponding to each primer was performed as indicated in the legend of Table 1.

Primer	Sequence	Protein	BLASTp hits on arabidopsis UGT database
Sense			
PUGTa	TG.TGC.GAG.TCC.ACC.ATG.aar.may.tgy.yt	LCESTMKNCL	85A
PUGTb	GCC.GAG.GAG.CTC.GGC.rtn.ccn.gar.gt	AEELGVPEV	85A
PUGTc	CC.TTC.TGC.CGG.ACC.ACC.ray.ccn.gay.ga	SFCRTTNPDD	85A
PUGTd	CT.ATG.CAG.TCC.ATC.TTC.ccn.ccn.rtn.ta	AMQSIFPPVY	85A
PUGTe	G.CCG.GAC.GTG.GTG.GCn.ggn.gan.gar.g	LCESTMKNCL	85A
Reverse			
PUGTar	GC.GGA.GGT.GGT.CCA.GAA.Cwr.nac.ytc.ngg	PEVQFWTTS	85A
PUGTbr	GAA.GTT.GAC.CAT.GAC.GTC.Gtc.ngg.rty.ngt	TNPDDVMVSF	85A
PUGTcr	CGG.GCC.GAC.GGT.Gta.nay.ngg.ngg	PPVYSVGP	85A/71B/71C/72B
PUGTdr	CGG.CAC.CAT.GGC.Ctc.ntc.ncc.nkc	AGEEAMVP	85A
PUGTer	CCA.CTC.GTC.GCA.GCA.GAA.Cyk.rca.rtt.nrt	MNCKFCCDEW	85A

such genes should provide a valuable library for pharmacological and toxicological investigations, and for studying evolutionary ecology of plant pathogen interactions.

Conclusions

The CODEHOP strategy appears as a powerful method for exploring the complexity of gene families in plants with a large genome, and conservation of genes across evolution. CYP98s are genes evolved early and are highly conserved during evolution, as expected for genes with an essential role in homeostasis and development of vascular plants. Conversely, the UGT85s are more variable. No orthologue of the sorghum *UGT85B1* gene was detected in wheat, while UGT85As, found in many plants species, are also present in wheat. The great variability of this subfamily in wheat strongly suggests a role in environmental adaptation and plant defense.

Methods
cDNA libraries

Wheat cDNA libraries were constructed from poly(A)+ mRNA from 3–5 mm *Triticum aestivum* (L. cv. Darius) seedlings, both control or pre-coated with cloquintocet-mexyl (0.1% seed dry weight) and phenobarbital, as described previously [34], in λ ZipLox (GIBCOBRL) by T. Thomas (College Station, Texas). Other cDNA libraries were kindly provided by B. Camara, IBMP Strasbourg (*Capsicum annuum*), W. Faigl, MPI Cologne (*Ceratopteris richardii*), M. Petersen, Philipps Universität Marburg (*Coleus blumei*), J. Grimma-Pettenati, CNRS Toulouse (*Eucalyptus globulus*), J. L. Evrard, IBMP Strasbourg (*Helianthus annuus*), A. Schaller, ETH Zurich (*Lycopersicon esculentum*), D. Ernst, GSF Munich (*Picea abies*), J. M. Frigerio, INRA Gazinet (*Pinus pinaster*), C. Douglas, University of British

Columbia Vancouver (*Populus trichocarpa x deltoides*), and Leeds University (*Physcomitrella patens*).

Bioinformatics

CODEHOP primers are designed as to ensure a very high probability of annealing on the gene of interest with 11–12 completely degenerate core nucleotides at their 3'-end and efficient amplification with a consensus 18–25 nucleotide clamp sequence at the 5'-end. Design of the clamp is based on an alignment of available related sequences and codon usage of the target organism. Primers completely degenerate at their 3'-end and ensuring high probability of annealing to *CYP98s* or *UGTs* specific sequences were designed using the CODEHOP strategy [22] based on the multiply-aligned sequences: AF029856, AF022458, AA86449, C74921, D47937, AI881302, AI734373, AAG52369, AAG52373 for the P450 family CYP98, and AAF17077, AAF18537, BAA34687 for UGTs. The multiple alignment was first generated using ClustalW [35], then cut into blocks using the BlockMaker server [36]. Primers were designed using the default parameters of the CODEHOP server [37]. It was assumed that barley codon usage proposed by the server was close enough to that of wheat to obtain effective primer design.

From the primer solutions proposed by the server, five P450 primers were selected, which both provided the largest PCR fragments and avoided the most conserved consensus regions common to a large number of P450s. In the case of UGTs, a broader choice was offered by the server, thus 5 sense and 5 reverse primers could be selected on the same basis as for P450s.

Figure 7
Alignment of protein segments deduced from amplified wheat *UGT* sequences with sbHMNGT (UGT85B1).

Probe amplification and analysis

In the present work, the CODEHOP approach was not used with genomic DNA but with cDNA libraries, our main objective being to identify genes expressed in specific plant tissues. Screening can be directly performed on cDNA libraries but better results were obtained after preliminary extraction of cDNA from phages. For extraction, an aliquot of the cDNA library was heated 10 min at 70°C, then extracted with one volume of phenol-chloroform. PCR screening was then performed on 50 ng of this template using 2.6 U of HiFi (Expand High Fidelity, Roche), 1.5 mM Mg^{2+}, 0.3 mM dNTP and 0.5 μM primers in the polymerase manufacturer's buffer. The PCR program was designed according to the CODEHOP server's tips, including successively a touch down (A) and a classical (B) PCR as follows: first 3 min initial denaturation at 94°C, then (A) 20 cycles of 1 min at 94°C, 2 min at 70°C (-1°C/cycle), and 2 min at 72°C, then (B) 20 cycles of 1 min at 94°C, 2 min at 58°C, and 2 min at 72°C, and finally a 5 min extension. Elongation time was adapted to the largest expected fragment in each screening. When no amplified fragment was obtained with a primer pair, the touchdown starting temperature of annealing was decreased by 5°C until a successful amplification was achieved.

Analysis and cloning of PCR products

PCR products were analyzed on 1% agarose gels, and fragments of expected size were eluted by centrifugation on Ultrafree-DA columns (Millipore), precipitated and cloned into the pGEM-T vector (Promega). *E. Coli* XL1blue (Stratagene) was electroporated with 1/10 of the ligation volume. Inserts from five white colonies were sequenced.

Sequences were analyzed by BLASTx against a local database generated from the data found [33,38]. Non-P450 sequences were xblasted on the NCBI site. The CYP98A10, CYP98A11 and CYP98A12 names were assigned on the basis of the full-length coding sequences subsequently isolated from the cDNA library.

Authors' contributions

M. M. adapted and optimized the method, carried out sequence alignments, primers design and PCR fragments analysis. A. H. participated in the analysis of the UGT family. D. W.-R. was involved in the design and coordination, and drafted the manuscript.

Acknowledgements

References

1. Werck-Reichhart D, Hehn A, Didierjean L: **Cytochromes P450 for engineering herbicide tolerance.** *Trends Plant Sci* 2000, **5:**116-123
2. Cole D, Edwards R: **Secondary metabolism of agrochemicals in plants.** *In Metabolism of Agrochemicals in Plants (Edited by: Roberts T) Wiley* 2000, 107-154
3. Chaudry Q, Schröder PP, Werck-Reichhart D, Grajek W, Marecik R: **Prospects and limitations of phytoremediation for the removal of persistent pesticides in the environment.** *Environ Sci Pollut Res* 2002, **9:**4-17
4. Coon MJ, Vaz AD, Bestervelt LL: **Cytochrome P450 2: peroxidative reactions of diversozymes.** *FASEB J* 1996, **10:**428-434
5. Mansuy D: **The great diversity of reactions catalyzed by cytochromes P450.** *Comp Biochem Physiol C Pharmacol Toxicol Endocrinol* 1998, **121:**5-14
6. Kahn R, Durst F: **Function and evolution of plant cytochrome P450.** *Recent Adv Phytochem* 2000, **34:**151-189
7. Werck-Reichhart D, Bak S, Paquette S: **Cytochromes P450.** *In: The Arabidopsis Book (Edited by: Somerville CR, Meyerowitz EM) American Society of Plant Biologists, Rockville, MD* 2002 [http://www.aspb.org/publications/arabidopsis/]doi/10.1199/tab.0028
8. Vogt T, Jones P: **Glycosyltransferases in plant natural product synthesis: characterization of a supergene family.** *Trends Plant Sci* 2000, **5:**380-386
9. Ross J, Li Y, Lim E, Bowles DJ: **Higher plant glycosyltransferases.** *Genome Biol* 2001, **2:**REVIEWS3004
10. Meijer AH, Souer E, Verpoorte R, Hoge JH: **Isolation of cytochrome P-450 cDNA clones from the higher plant *Catharanthus roseus* by a PCR strategy.** *Plant Mol Biol* 1993, **22:**379-83
11. Holton TA, Brugliera F, Lester DR, Tanaka Y, Hyland CD, Menting JG, CY Lu, Farcy E, Stevenson TW, Cornish EC: **Cloning and expression of cytochrome P450 genes controlling flower colour.** *Nature* 1993, **366:**276-279
12. Toguri T, Tokugawa K: **Cloning of eggplant hypocotyl cDNAs encoding cytochromes P450 belonging to a novel family (CYP77).** *FEBS Lett* 1994, **338:**290-294
13. Urvardi MK, Metzger JD, Krishnapillai V, Peacock WJ, Dennis E: **Cloning and sequencing of a full-length cDNA from *Thlaspi arvense* L. that encodes a cytochrome P450.** *Plant Physiol* 1994, **105:**755-756
14. Frank MR, Deneyneka JM, Schuler MA: **Cloning of wound-induced cytochrome P450 monooxygenases expressed in pea.** *Plant Physiol* 1996, **110:**1035-1046
15. Akashi T, Aoki T, Takahashi T, Kameya N, Nakamura I, Ayabe S-I: **Cloning of cytochrome P450 cDNAs from cultured *Glycyrrhiza echinata* L. cells and their transcriptional activation by elicitor treatment.** *Plant Sci* 1997, **126:**39-47
16. Mizutani M, Ward E, Ohta D: **Cytochrome P450 superfamily in *Arabidopsis thaliana*: isolation of cDNAs, differential expression, and RFLP mapping of multiple cytochromes P450.** *Plant Mol Biol* 1998, **37:**39-52
17. Bak S, Kahn RA, Nielsen HL, Møller BL, Halkier BA: **Cloning of three A-type cytochromes P450, CYP71E1, CYP98, and CYP99 from *Sorghum bicolor* (L.) Moench by a PCR approach and identification by expression in *Escherichia coli* of CYP71E1 as a multifunctional cytochrome P450 in the biosynthesis of the cyanogenic glucoside dhurrin.** *Plant Mol Biol* 1998, **36:**393-405
18. Ralston L, Kwon ST, Schoenbeck M, Ralston J, Schenk DJ, Coates RM, Chappell J: **Cloning, heterologous expression, and functional characterization of 5-epi-aristolochene-1,3-dihydroxylase from tobacco (*Nicotiana tabacum*).** *Arch Biochem Biophys* 2001, **393:**222-235
19. Schopfer CR, Ebel J: **Identification of elicitor-induced cytochromes P450 of soybean (*Glycine max L.*) using differential display of mRNA.** *Mol Gen Genet* 1998, **258:**315-322
20. Schoendorf A, Rithner C, Williams RM, Croteau RB: **Molecular cloning of a cytochrome P450 taxane 10β-hydroxylase cDNA from *Taxus* and functional expression in yeast.** *Proc Natl Acad Sci USA* 2001, **98:**1501-1506
21. Fischer TC, Klattig JT, Gierl A: **A general cloning strategy for divergent plant cytochrome P450 genes and its application in *Lolium rigidum* and *Ocimum basilicum*.** *Theor Appl Genet* 2001, **103:**1014-1021
22. Rose TM, Schultz ER, Henikoff JG, Pietrovski S, McCallum C, Henikoff S: **Consensus-degenerate hybrid oligonucleotide primers for**

amplification of distantly related sequences. *Nucleic Acids Res* 1998, **26:**1628-1635

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

24. Franke R, Humphreys JM, Hemm MR, Denault JW, Ruegger MO, Cusumano JC, Chapple C: **The** *Arabidopsis REF8* **gene encodes the 3-hydroxylase of phenylpropanoid metabolism.** *Plant J* 2002, **30:**33-45

25. Franke R, Hemm MR, Denault JW, Ruegger MO, Humphreys JM, Chapple C: **Changes in secondary metabolism and deposition of an unusual lignin in the** *ref8* **mutant of** *Arabidopsis.* *Plant J* 2002, **30:**47-59

26. Baucher M, Monties B, Van Montaigu M, Boerjean W: **Biosynthesis and genetic engineering of lignin.** *Crit Rev Plant Sci* 1998, **17:**125-197

27. Boudet AM: **Lignins and lignification: selected issues.** *Plant Physiol Biochem* 2000, **38:**1-16

28. Siminsky B, Corbin FT, Ward ER, Fleischmann TJ, Dewey RE: **Expression of a soybean cytochrome P450 monooxygenase cDNA in yeast and tobacco enhances the metabolism of phenylurea herbicides.** *Proc Natl Acad Sci USA* 1999, **96:**1750-1755

29. Cooper-Driver GA, Bhattacharya M: **Role of phenolics in plant evolution.** *Phytochemistry* 1998, **49:**1165-1174

30. Jones PR, Møller BL, Høj PB: **The UDP-glucose:p-hydroxymandelonitrile-O-glucosyltransferase that catalyzes the last step in synthesis of the cyanogenic glucoside dhurrin in** *Sorghum bicolor.* **Isolation, cloning, heterologous expression, and substrate specificity.** *J Biol Chem* 1999, **274:**35483-35491

31. Brazier M, Cole DJ, Edwards R: *O***-Glucosyltransferase activities toward phenolic natural products and xenobiotics in wheat and herbicide-resistant and herbicide susceptible black-grass (***Alopecurus myosuroides***).** *Phytochemistry* 2002, **59:**149-156

32. Rieseberg LH: **Polyploid evolution: Keeping peace at genomic reunions.** *Current Biology* 2001, **11:**R925-R928

33. **The Arabidopsis P450, Cytochrome b$_5$, P450 Reductase, and Glycosyl Transferase Family 1 Database at PLACE** [http://www.biobase.dk/P450/]

34. Cabello-Hurtado F, Zimmerlin A, Rahier A, Taton M, DeRose R, Nedelkina S, Batard Y, Durst F, Pallet KE, Werck-Reichhart D: **Cloning and functional expression in yeast of a cDNA coding for an obtusifoliol 14α-demethylase (CYP51) in wheat.** *Biochem Biophys Res Commun* 1997, **230:**381-385

35. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680

36. **Blocks WWW Server** [http://blocks.fhcrc.org/blocks/]

37. **CODEHOP: COnsensus-DEgenerate Hybrid Oligonucleotide Primers** [http://blocks.fhcrc.org/codehop.html]

38. **David Nelson's Cytochrome P450 Homepage** [http://drnelson.utmem.edu/CytochromeP450.html]

39. **Genedoc: a tool for editing and annoting multiple sequence alignments** [www.psc.edu/biomed/genedoc]

40. Hall TA: **BioEdit: an user friendly biological sequence alignment editor and analysis program for Windows 95/98/NT.** *Nucl Acids Symp Ser* 1999, **41:**95-98