Gene Duplication Leads to Altered Membrane Topology of a Cytochrome P450 Enzyme in Seed Plants

Hugues Renault, Minttu de Marothy, Gabriella Jonasson, Paricia Lara, David R Nelson, Ingmarie Nilsson, François André, Gunnar von Heijne, Danièle Werck-Reichhart

Gene Duplication Leads to Altered Membrane Topology of a Cytochrome P450 Enzyme in Seed Plants

Hugues Renault,*[,1] Minttu De Marothy,[2,3] Gabriella Jonasson,[4] Patricia Lara,[2] David R. Nelson,[5] IngMarie Nilsson,[2] François André,[4] Gunnar von Heijne,[2,3] and Danièle Werck-Reichhart*[,1]

[1]Centre National de la Recherche Scientifique, Institute of Plant Molecular Biology, University of Strasbourg, Strasbourg, France

[2]Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden

[3]Science for Life Laboratory, Stockholm University, Solna, Sweden

[4]Institute for Integrative Biology of the Cell (I2BC), DRF/Joliot/SB[2]SM, CEA, CNRS, Université Paris Sud, Université Paris-Saclay, Gif-sur-Yvette, France

[5]Department of Microbiology, Immunology and Biochemistry, University of Tennessee Health Science Center, Memphis, TN

***Corresponding authors:** E-mails: hugues.renault@ibmp-cnrs.unistra.fr; werck@unistra.fr.

Associate editor: Song Ge

Abstract

Evolution of the phenolic metabolism was critical for the transition of plants from water to land. A cytochrome P450, CYP73, with cinnamate 4-hydroxylase (C4H) activity, catalyzes the first plant-specific and rate-limiting step in this pathway. The *CYP73* gene is absent from green algae, and first detected in bryophytes. A *CYP73* duplication occurred in the ancestor of seed plants and was retained in Taxaceae and most angiosperms. In spite of a clear divergence in primary sequence, both paralogs can fulfill comparable cinnamate hydroxylase roles both in vitro and in vivo. One of them seems dedicated to the biosynthesis of lignin precursors. Its *N*-terminus forms a single membrane spanning helix and its properties and length are highly constrained. The second is characterized by an elongated and variable *N*-terminus, reminiscent of ancestral CYP73s. Using as proxies the *Brachypodium distachyon* proteins, we show that the elongation of the *N*-terminus does not result in an altered subcellular localization, but in a distinct membrane topology. Insertion in the membrane of endoplasmic reticulum via a double-spanning open hairpin structure allows reorientation to the lumen of the catalytic domain of the protein. In agreement with participation to a different functional unit and supramolecular organization, the protein displays modified heme proximal surface. These data suggest the evolution of divergent C4H enzymes feeding different branches of the phenolic network in seed plants. It shows that specialization required for retention of gene duplicates may result from altered protein topology rather than change in enzyme activity.

Key words: **plant metabolism, membrane protein, metabolic complexity, cinnamic acid 4-hydroxylase, evolution of lignin metabolism.**

Article

Introduction

Land colonization was a major step of plant evolution. Occurring about 450 million year ago (MYA), it was a founding event for the formation of terrestrial ecosystems (Kenrick and Crane 1997; Bowman 2013). For their transition from water to land, plants had to face major challenges. One of them was to prevent loss of water and solutes. Another was to keep protected from UV and strong solar radiation, and from other challenging environmental conditions. The first critical step was most likely to shield the gametophyte from desiccation and other environmental stresses, such as light and temperature (Rensing et al. 2008; Renault et al. 2017). Vascular tissues were another important innovation in land plants, allowing long-distance water and nutrient transport, and providing rigid structure for erect growth (Weng and Chapple 2010). Acquisition of new metabolic functions was thus required.

Evolution of the phenolic metabolism is associated with early plant transition from water to land, and was most likely one of the prerequisites for adaptation to a more constrained environment. In higher plants, this pathway leads to the synthesis of antioxidants, UV-screens, defense compounds and precursors of biopolymers such as sporopollenin, lignin and suberin, essential for water conduction in vascular tissues and for protection against desiccation, UV exposure and pests. Lignin is in addition the second most abundant biopolymer on Earth after cellulose (Boerjan et al. 2003) and thus constitutes a major carbon sink.

The plant phenolic pathway is highly channeled via supramolecular organization of the soluble enzymes around membrane-anchored oxygenases belonging to superfamily of cytochromes P450 to form functional units called metabolons (Achnine et al. 2004; Chen et al. 2011; Bassard et al. 2012b). This spatial organization is thought to be important to provide a sustained flux of precursors in a pathway that processes up to 30% of the carbon fixed by photosynthesis, but also to provide efficient switches into branch pathways. The cinnamate 4-hydroxylase, or CYP73, is one of these membrane-anchored P450 oxygenases (Teutsch et al. 1993). It catalyzes the first phenolic ring hydroxylation, second step

of the core segment, common to all branches of the pathway. It also serves as an anchor for the most upstream functional unit or metabolon (Bassard et al. 2012b). Although sequencing of the first plant genome revealed the presence of a single *CYP73* gene copy in *Arabidopsis thaliana* (Bak et al. 2011), earlier investigations based on protein purification indicated the existence of a *CYP73* duplication in French bean (*Phaseolus vulgaris*) (Nedelkina et al. 1999). A *CYP73* gene duplication was later confirmed in other plant species (Ehlting et al. 2006). On the basis of the large set of recently sequenced genomes, we reinvestigate here the evolution and complexity of the CYP73 family in plants.

Our large-scale phylogenomic analysis points to a conserved gene duplication in seed plants, which resulted in one CYP73 copy with a highly constrained *N*-terminal membrane anchoring segment, and a second with an elongated and more variable *N*-terminus. We investigate the impact of this specific trait on the subcellular localization and membrane topology of the protein using as a model CYP73A94 from the monocotyledon *Brachypodium distachyon*. This investigation suggests a modification of the membrane topology of this protein, unprecedented for a plant P450 enzyme. The implications of this duplication and resulting modification of membrane topology are discussed.

Results

Members of the CYP73 Family Can Be Detected Only in Land Plants

It is usually assumed that the emergence of the phenolic metabolism was required to support the transition of plants from water to land. Homologs of the first enzyme in the pathway, the phenylalanine ammonia-lyase (PAL), are present in microorganisms (Barros et al. 2016). CYP73 is thus expected to be the next key determinant allowing for the evolution of a phenylpropanoid pathway. Based on the mining of the first genomic data, CYP73 homologs were found in the most basal group of land plants (i.e., bryophytes), but not in the green algae *Chlamydomonas reinhardtii* (Ehlting et al. 2006). In an attempt to more precisely date the first emergence of the basal phenolic pathway, we reinvestigated the largely extended data recently made available from genome and transcriptome sequencing, including the 1kP project (www.onekP.com; last accessed May 22, 2017). Whereas CYP73 homologs were found in all investigated taxa of land plants, no related sequences were detected by BLAST in green algae, including Klebsormidiales and Zygnematales (supplementary tables S1–S3, Supplementary Material online), the sister group to land plants (Wickett et al. 2014; Chang et al. 2016). This finding was confirmed using the hidden Markov model based pHMMER software (http://hmmer.org; last accessed May 22, 2017), which failed to detect any remote homology in green algae and essentially recapitulated BLAST analysis (supplementary tables S1–S3, Supplementary Material online). Absence of CYP73 homologs in green algae was further indicated by a phylogenetic analysis including all cytochromes P450 from *Chlamydomonas reinhardtii* and *Klebsormidium nitens*, and representative members of the CYP73 and CYP98 families from land plants (supplementary fig. S1, Supplementary Material online). This analysis pointed to algal sequences that might derive from the same common ancestor as CYP73s and CYP98s from higher plants, but none of them can be reliably associated with the CYP73 clade and rather group with the CYP98 clade (supplementary fig. S1, Supplementary Material online).

To ascertain that bryophyte *CYP73*-like genes already encoded true cinnamic acid 4-hydroxylase (C4H), a putative homolog from *Physcomitrella patens* was heterologously expressed in yeast, and the activity of the resulting recombinant protein was assessed. PpCYP73A48 efficiently converted cinnamic acid into *p*-coumaric acid, indicating that CYP73 evolved in bryophytes most likely for cinnamic acid hydroxylation as an early step of the phenolic metabolism (supplementary fig. S2, Supplementary Material online).

Together, our data would date the emergence of the CYP73 family after the split of land and water obligate plants around 450 MYA.

Two Classes of Conserved *CYP73* Genes Evolved in Seed Plants

The evolutionary history of the *CYP73* family was then reconstructed with a maximum likelihood (ML) approach, using 204 *CYP73* sequences derived from 118 land plant species (supplementary table S4, Supplementary Material online). The phylogeny of the *CYP73* family essentially matched the organismal phylogeny, asserting robustness, and accuracy of the phylogenetic reconstruction (fig. 1). The tree structure also revealed an early duplication in the *CYP73* family in seed plants, occurring prior to the split between gymnosperms and angiosperms. The two copies of *CYP73s*, hereafter referred to as class I and class II genes, were retained in most fully sequenced genomes of seed plants (fig. 1; supplementary table S4, Supplementary Material online). A notable exception is detected in gymnosperms for which class II genes could be retrieved only from Taxaceae using transcriptome data (supplementary fig. S3 and table S4, Supplementary Material online), whereas none was found in the *Picea abies* genome and transcriptome of nontaxaceae species. This suggests that the class II copy may have been lost in some conifer lineages. In addition, a recent loss of class II gene has occurred in Brassicaeae since no homolog could be retrieved from the fully sequenced Brassicaceae genomes, including *Arabidopsis sp.* and the basal *Aethionema arabicum*, whereas genomes of the earlier diverging *Carica papaya* (*Capparaceae*) and *Tarenaya hassleriana* (*Cleomaceae*) species include a class II *CYP73* (supplementary fig. S3, Supplementary Material online). The class II *CYP73* gene loss event in *Brassicaceae* can therefore be dated between 65 and 55 MYA (Beilstein et al. 2010).

The CYP73 Family Has Evolved under Strong Purifying Selection

With one to five copies per genome (supplementary table S4, Supplementary Material online), the *CYP73* genes have been kept in low-copy number in vascular plants compared with other plant P450 families. The main clade, most likely

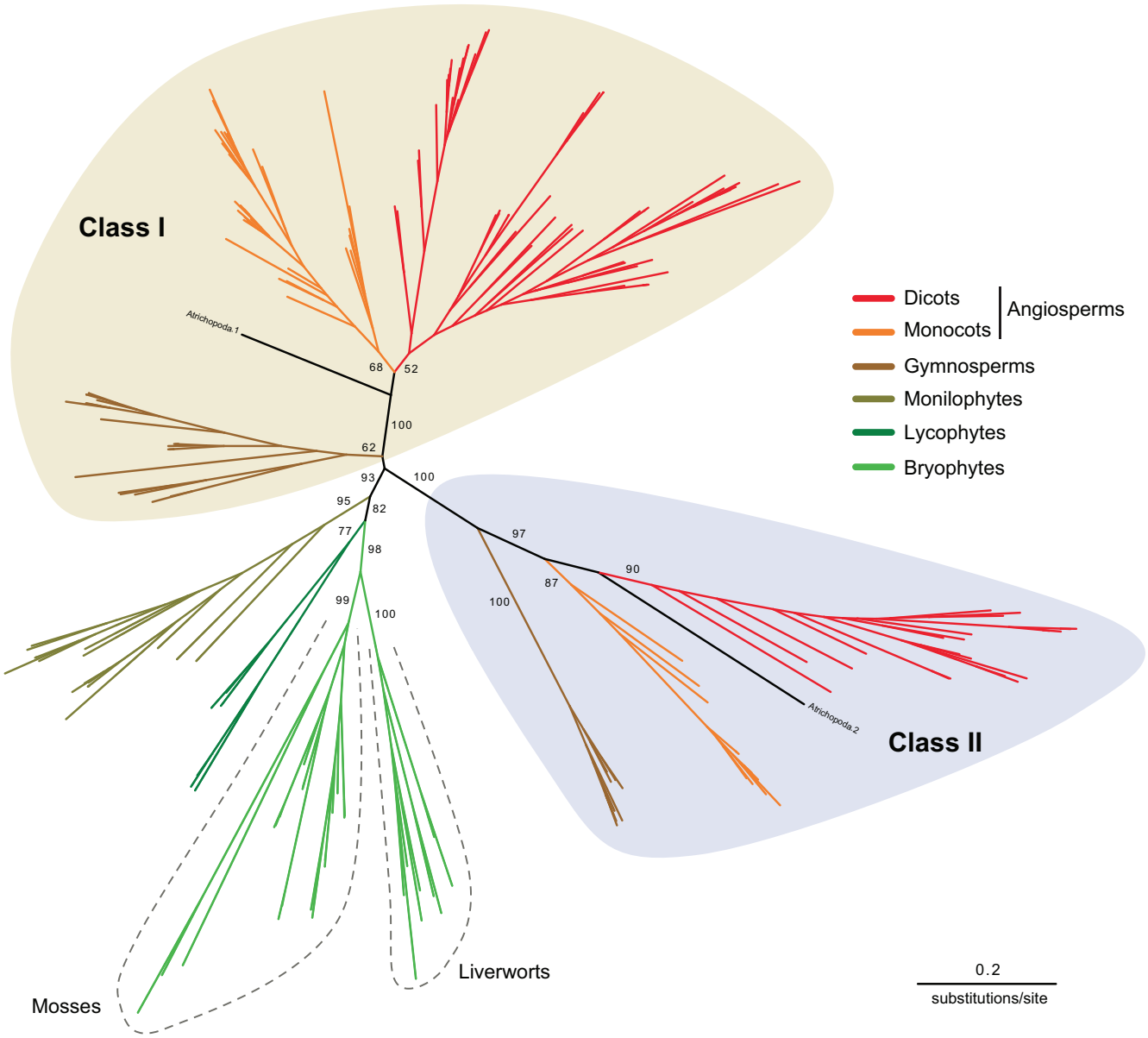

FIG. 1. Diversification of the *CYP73* family in seed plants. Unrooted maximum likelihood (ML) tree depicting the phylogenetic relationships between 204 *CYP73* coding sequences. The tree with the highest log likelihood (-87772.594699) is shown. The analysis involved 204 sequences and 1,218 nucleotide sites. Bootstrap values from 100 iterations are indicated on main branches. *A. trichopoda: Amborella trichopoda* (basalmost angiosperm). Tree is drawn to scale.

featuring true orthologs essential for lignin biosynthesis in vascular plants (Schilmiller et al. 2009), has a representative in all seed plants and corresponds to class I enzymes (fig. 2). A molecular evolution approach was taken to evaluate the selection prevailing during history of the *CYP73* family of genes. The CodeML software from the Phylogenetic Analysis by Maximum Likelihood (PAML) package (Yang 2007) was used to calculate the ratio of nonsynonymous (d*N*) to synonymous (d*S*) rates of substitution (d*N*/d*S* = ω). Under the M0 model that assumes a single ω ratio for the whole tree, the *CYP73* gene family was found with a global ω of 0.051 (supplementary note S1, Supplementary Material online), which is indicative of high purifying selection. This observation was further supported by the Nearly Neutral M1a site model that allows the ω ratio to vary among sites following two categories ($\omega 0 < 1$ and $\omega 1 = 1$), and which showed that 96.6% of *CYP73* codons have an ω of 0.048 (supplementary note S1, Supplementary Material online).

We next asked whether some gene lineages evolved under a different selection pressure. A branch model analysis was first tested, assigning ten different ω ratios to ten different clades (i.e., mosses, liverworts, lycophytes, monilophytes, gymnosperm class I, monocot class I, dicot class I, gymnosperm class II, monocot class II, and dicot class II). This model proved significantly better than the M0 model (*P* value = 3.28E-41), the 3 ratio model (i.e., non-seed plants, class I, class II; *P* value = 2.50E-20) and the 7 ratio model (i.e., nonseed plants, gymnosperm class I, monocot class I, dicot class I, gymnosperm class II, monocot class II, and dicot class II; *P* value = 1.85E-05) (supplementary note S1, Supplementary

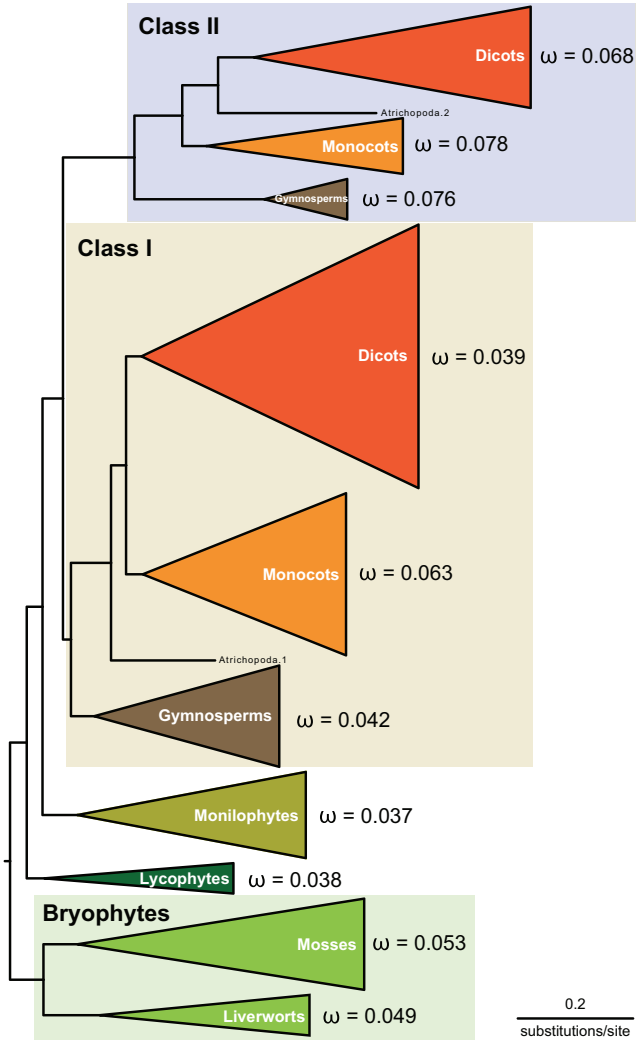

Fig. 2. Molecular evolution of the *CYP73* family. The dN/dS (ω) ratios of ten different clades were simultaneously computed with the CodeML branch model and mapped onto the *CYP73* ML tree. To orient the evolutionary scenario, the tree was rooted on the earliest diverging group of land plants (i.e., bryophytes). The tree is drawn to scale.

Material online). All ten groups were found to evolve under overall strong purifying selection (fig. 2), although *CYP73* duplication in seed plants appeared to result in a slight relaxation of the purifying selection in the class II genes, especially in gymnosperms and dicots (fig. 2). This relaxation is less obvious for monocots for which purifying selection exerted on class I genes is already looser than for gymnosperms and dicots, possibly linked to an additional gene duplication. Also noteworthy is a comparatively relaxed negative selection in mosses compared with vascular lower plants (fig. 2). Relaxation of purifying selection in class II clades and mosses was further supported by results of Clade model that allows one site category (out of three) to vary between defined branches of the tree (supplementary note S1, Supplmentary Material online).

To check whether some residues evolved under positive selection in the class II sequences, we performed selection M2a branch-site model analyses, allowing ω ratio to vary among sites in a defined clade (foreground) following three categories ($0 \leq \omega 0 \leq 1$, $\omega 1 = 1$ and $\omega 2 \geq 1$). In agreement with our prior site model analysis, we did not detect any sites under positive selection (supplementary note S1, Supplmentary Material online). However, positive selection episodes can be discrete, occurring transiently during gene evolution. Therefore, we implemented a Fitmodel analysis that allows a change of selection regime during evolution (Guindon et al. 2004). Even though the M2a + S1 model was significantly superior to the M2a model without selection regime switch (*P* value = 0.00E + 0), no site that had evolved under positive selection during CYP73 evolution was uncovered (supplementary note S1, Supplmentary Material online). Instead, 16.25% of the sites were shown to have evolved under neutral evolution, 16.25% had an ω value of 0.16 and the remaining (67.5%) had a very low ω value of 0.0032 (supplementary note S1, Supplmentary Material online).

Both Class I and Class II CYP73 Enzymes Maintain High Cinnamate Hydroxylase Activity

The next step was to determine whether the class II *CYP73* gene duplication led to functional innovation. To answer this question, we tested the activity of the two classes of proteins in angiosperms, using the protein supplement of *Brachypodium distachyon*, a model temperate grass (International Brachypodium Initiative 2010; Brkljacic et al. 2011). Three *CYP73* paralogs are encoded by the *B. distachyon* genome, two belonging to the class I group (i.e., *CYP73A92* and *CYP73A93*) and one belonging to the class II (*CYP73A94*). The CYP73A94 protein shares 62.7% and 65.7% of identity with CYP73A92 and CYP73A93 (sharing 76.3% identity with each other), respectively. The recombinant proteins were produced in yeast (*Saccharomyces cerevisiae*) (fig. 3*A*) and tested for binding and catalytic properties with *t*-cinnamic acid. All three enzymes showed high and very similar dissociation constants (in the μM range) for the binding of cinnamate (fig. 3*B* and *C*), and high catalytic efficiencies, around 100 min^{-1} μM^{-1}, for the conversion of cinnamate into *p*-coumarate (fig. 3*D*), which are among the best reported for membrane-bound P450 enzymes.

To ensure that all three *B. dystachyon* proteins also fill the same biological function, we expressed them in an *Arabidopsis thaliana C4H*-deficient mutant line (*cyp73a5-1*). Expression of the constructs was driven by the *A. thaliana CYP73A5* promoter ($AtC4H_{pro}$) (Bell-Lelong et al. 1997). The *A. thaliana C4H*-deficient mutants show a severe phenotype of stunted growth (fig. 3*E*). Introduction of all three *B. distachyon CYP73* genes in the mutant line successfully restored normal plant growth, comparable to the wild type (fig. 3*E*). These results demonstrate that the two classes of CYP73 enzymes can perform the same canonical function, cinnamate 4-hydroxylation, in the relevant plant tissues so as to support plant growth. However, they do not exclude the possibility of a divergent moonlighting function or of a specialization in specific plant tissues or subcellular compartment.

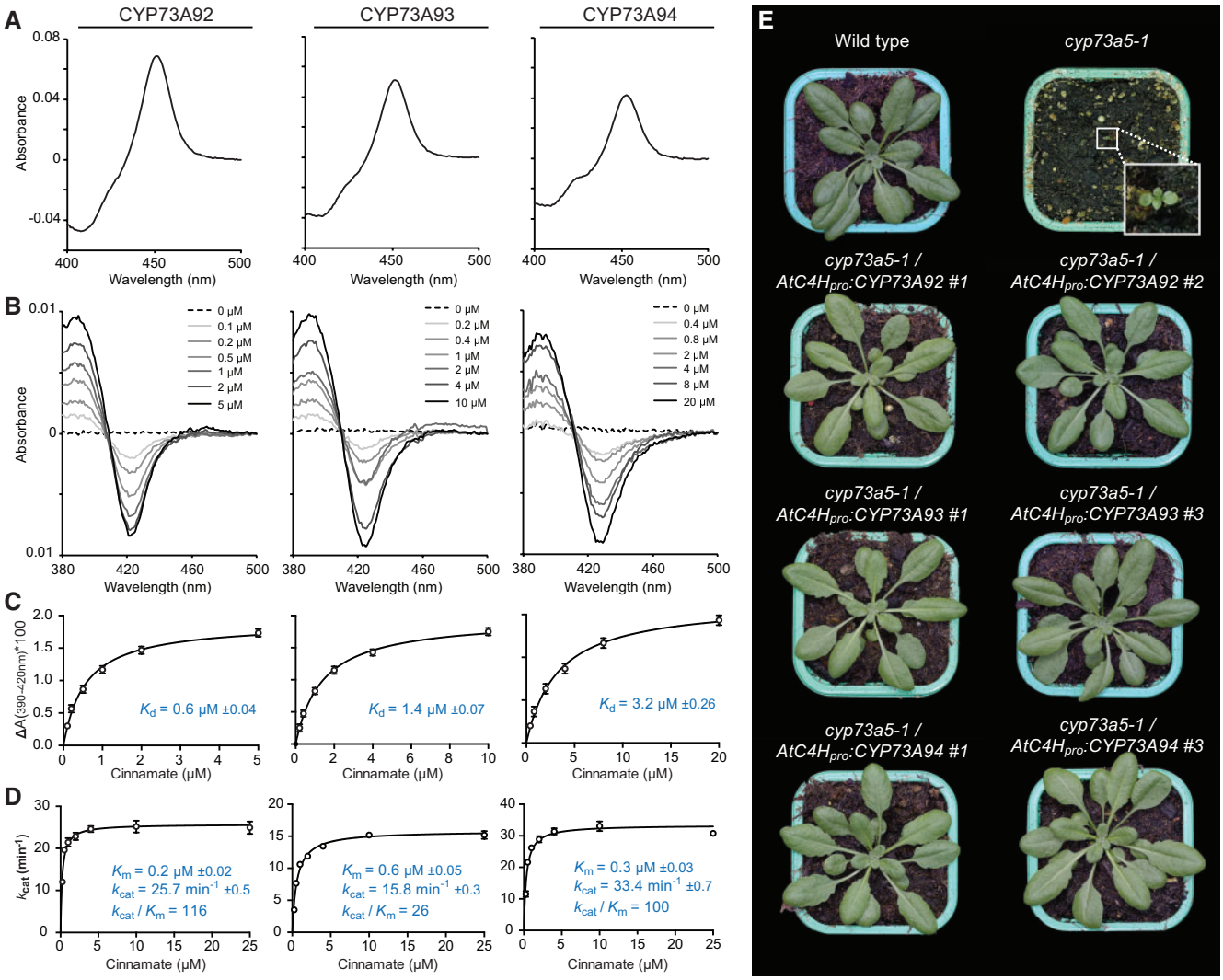

Fig. 3. *Brachypodium distachyon* CYP73 enzymes share common function in vitro and in vivo. (*A*–*D*) Evaluation of enzyme expression and catalytic activities using microsomal preparations of yeast expressing the CYP73 proteins. (*A*) Carbon monoxide-induced UV-visible difference spectra of dithionite-reduced *B. distachyon* CYP73s (recorded using 20-fold-diluted yeast microsomal preparations). (*B*) Representative cinnamate-induced type I difference spectra. Spectra were recorded with 150 nM native (oxidized) P450 enzyme, adding an increasing concentration of cinnamate in the assay. (*C*) Binding saturation curves based on the absorbance difference between 390 and 420 nm plotted against cinnamate concentration. Dissociation constant (K_d) was deduced from saturation curves according to Michaelis–Menten nonlinear regression. Results are the mean ± SD of 3–4 independent determinations. (*D*) Saturation curves of cinnamate 4-hydoxylase activity. Turnover (k_{cat}) and affinity (K_m) constants were deduced from saturation curves based on Michaelis–Menten nonlinear regression. Results are the mean ± SD of three independent enzyme assays. Catalytic efficiencies are expressed in min^{-1} μM^{-1}. (*E*) *A. thaliana C4H* promoter-driven expression of the three *B. distachyon CYP73* genes in the *cyp73a5-1* mutant restored a wild-type growth.

Class II CYP73 Proteins Share Unusual Characteristics in Their Primary Structure

We then searched for a hint to functional specialization that may explain conservation of the paralogs in the CYP73 protein alignment. This search pointed to distinctive features in the primary structure of the two classes of proteins. The most obvious was the unusual length of the *N*-terminal membrane-anchoring segment of class II compared with class I (fig. 4*A*; supplementary fig. S4, Supplementary Material online). A systematic comparison of the lengths of the *N*-terminal anchors of all CYP73 proteins supports the idea that it first decreased in the course of evolution, reaching a median of 42 amino acids (aa) in monilophytes (fig. 4*B*). Then, after *CYP73* duplication in seed plant ancestor, anchor length either kept decreasing in class I (median: 33 aa) or increased in class II (median: 58 aa), the class II CYP73s long *N*-terminus being reminiscent of that present in proteins from early diverging phyla such as bryophytes and lycophytes (fig. 4*B*).

The *N*-terminal membrane spanning segments from class I and class II enzymes also largely differ in their amino acid composition (fig. 4*C*). On average, the class II *N*-terminus is enriched in polar uncharged amino acids, such as serine and threonine, as compared with nonseed plants, but especially to the shortened membrane spanning segment of the class I proteins (fig. 4*C*). Class I proteins in addition feature the typical cluster of positively charged amino acids preceding the

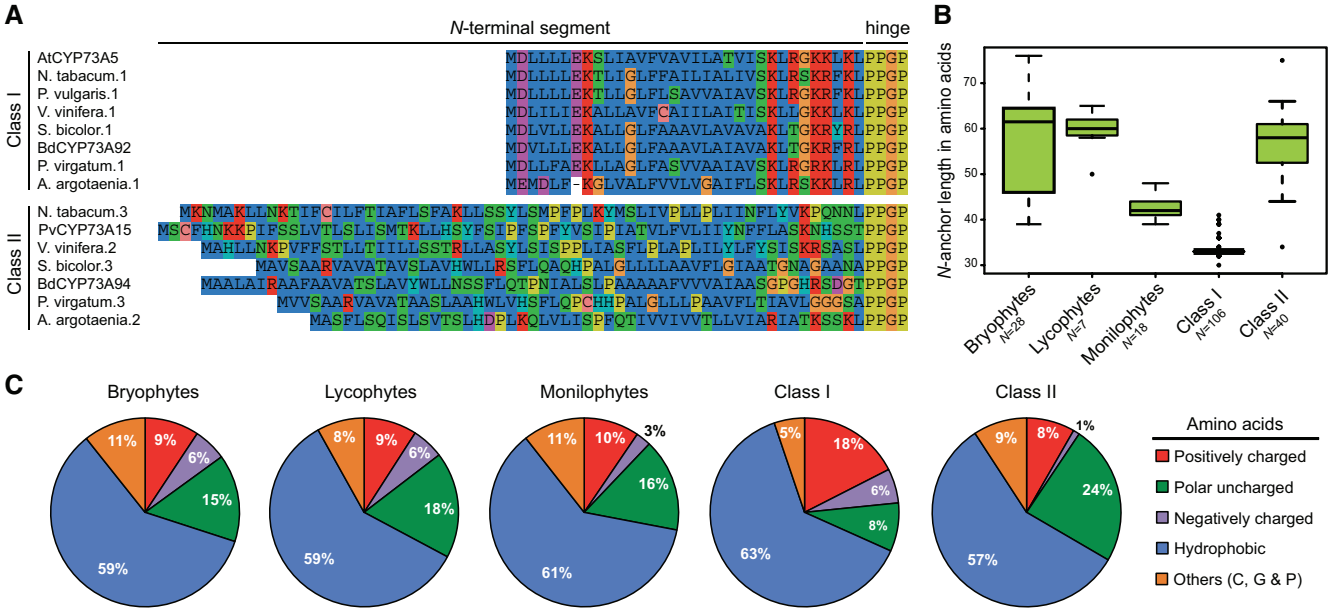

FIG. 4. Class I and Class II CYP73 proteins differ at their *N*-terminal membrane spanning segment. (*A*) Stack of *N*-terminal sequences of representative members of the seed plant CYP73s. At, *A. thaliana*; Pv, *P. vulgaris*; Bd, *B. distachyon*. (*B*) Box plots showing *N*-ter anchor length distribution among different gene lineages. *N*-ter anchor was defined as the protein segment spanning the first methionine amino acid to the first proline of the hinge (not included). (*C*) Amino acid composition of the *N*-ter anchor from different gene lineages. Relative abundance of positively charged (Arg, His, and Lys), polar uncharged (Ser, Thr, Asn, and Gln), negatively charged (Asp, Glu), hydrophobic (Ala, Val, Ile, Leu, Met, Phe, Tyr, and Trp) amino acids was determined. (Cys, Gly, and Pro) appear as Others.

proline rich hinge that makes the transition between the membrane spanning segment and the bulky protein domain at the surface of the membrane (fig. 4*A*). This cluster, expected to polarize protein membrane topology, is absent in the class II CYP73s. Most strikingly, the length and composition of the class I N terminus appear very conserved, when compared with the variable class II *N*-terminal sequence (fig. 4*A*).

The Two Classes of CYP73s Are Targeted to the Endoplasmic Reticulum

Altogether, the differences in the class I and II protein transmembrane segment length and composition seemed indicative of adaptation and targeting to specific membrane environments. In addition, a serine/threonine-enriched *N*-terminal extension, such as found in class II proteins, is characteristic of chloroplast transit peptides (von Heijne et al. 1989).

A prediction of subcellular localization of class II proteins using different algorithms (Aramemnon; http://aramemnon.uni-koeln.de/) provided contradictory results. The subcellular localization of EGFP fusions of the three *B. distachyon* CYP73 proteins was, therefore, experimentally compared after transient expression in the leaves of *Nicotiana benthamiana*. Both class I and class II proteins were detected exclusively associated with the membranes of the endoplasmic reticulum (fig. 5*A*).

We then considered the possibility that this subcellular localization might be tissue-dependent. To test this possibility, the *AtC4H*$_{pro}$-driven *CYP73A94:EGFP* fusion was expressed in the *A. thaliana cyp73a5-1* mutant. As shown in figure 5*B*, the CYP73A94:EGFP fusion protein fully restored wild-type growth. The protein subcellular localization was investigated in roots, cotyledons and hypocotyls of the complemented line. Consistent with the transient expression experiments, we observed a typical endoplasmic reticulum (ER) localization of the EGFP signal in all three organs and did not detect additional signal in other compartments (fig. 5*B*; supplementary fig. S5, Supplementary Material online).

Class I and Class II CYP73s Have Different Membrane Topologies

As the *N*-terminal divergence did not seem to impact the protein subcellular localization, we set out to investigate whether it may influence the membrane topology of CYP73 proteins. First, the TOPCONS software (Bernsel et al. 2009) was used to predict consensus membrane topology of the three *B. distachyon* CYP73 paralogs (supplementary fig. S6, Supplementary Material online). Both class I CYP73 proteins (i.e., CYP73A92 and CYP73A93) were predicted to have a single transmembrane helix (TMH) (H1) at the *N*-terminus followed by a cytoplasmic P450 domain and a weakly hydrophobic, second predicted TMH located *C*-terminally, immediately after the heme-binding cysteine (fig. 6*A*; supplementary fig. S6, Supplementary Material online). Conversely, the class II CYP73A94 was predicted to have two *N*-terminal transmembrane helices (H1 and H2), and a third, weakly hydrophobic TMH near the C terminus (fig. 6*A*; supplementary fig. S6, Supplementary Material online). These features were consistently observed in other angiosperms class II proteins (supplementary fig. S6, Supplementary Material online), but not in bryophytes and lycophytes

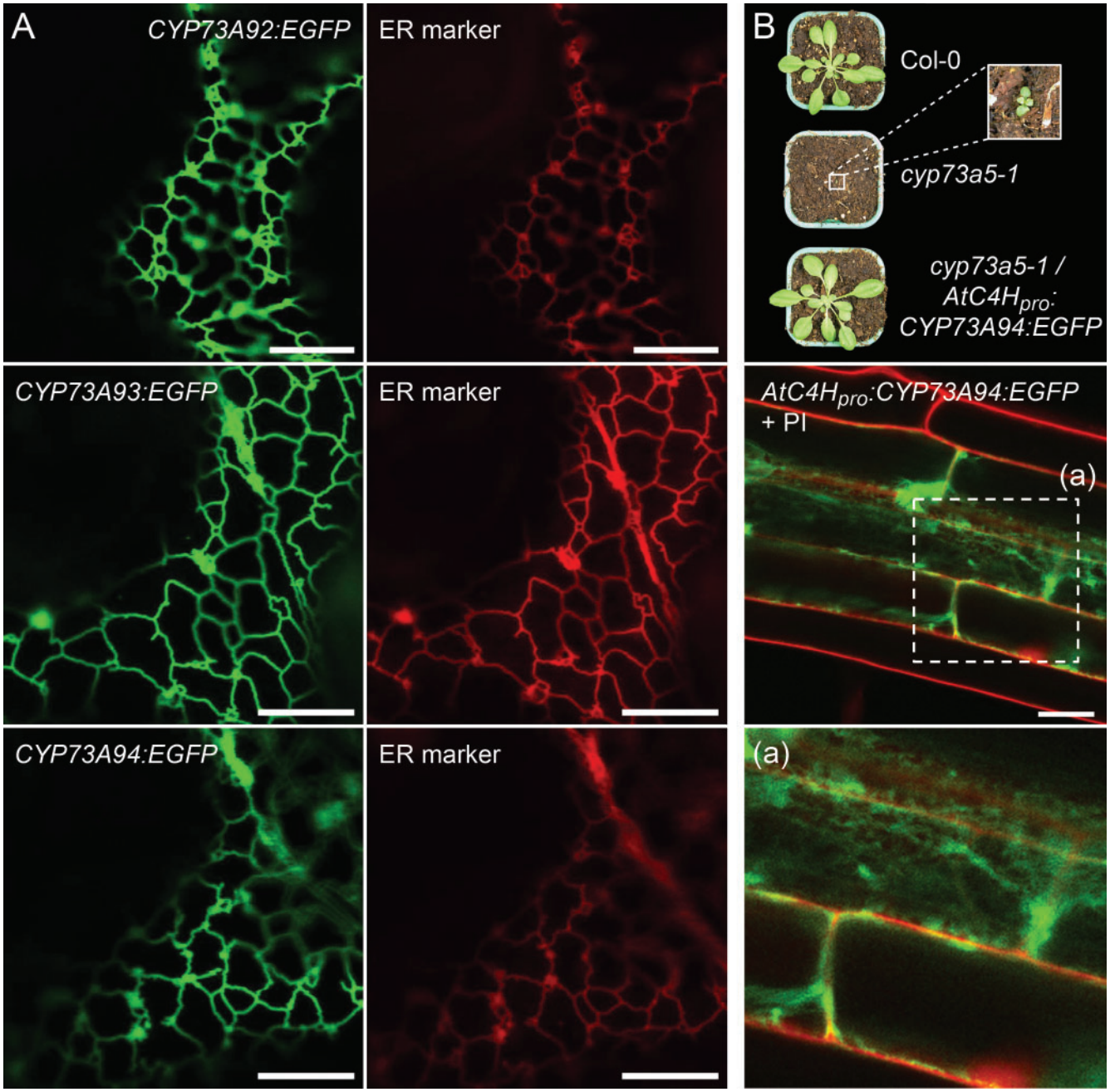

Fig. 5. Subcellular localization of the *Brachypodium* CYP73 proteins. (*A*) Typical confocal pictures taken 4 days after agro-infiltration of *N. benthamiana* leaves with EGFP fusion constructs. A construct containing the mRFP1 fluorophore *C*-terminally fused to the membrane-spanning domain of *A. thaliana* CYP51G1 (Bassard et al. 2012b) was coexpressed and used as ER marker. ER movement was restrained using 20 μM latrunculin B. Scale bars, 10 μm. (*B*) Typical confocal pictures of 5-day-old roots of the *cyp73a5-1* mutant complemented with the *AtC4H$_{pro}$:CYP73A94:EGFP* construct. Cell walls are counter-stained with propidium iodide (PI). Scale bar, 20 μm.

homologs that also harbor long *N*-terminal segments (fig. 4*B*; supplementary fig. S7, Supplementary Material online).

Homology models of the CYP73 proteins generated using as templates crystalized P450 structures indicate that the predicted *C*-terminal TMH, that includes the heme-binding cysteine residue, in fact belongs to the globular protein fold (supplementary fig. S8, Supplementary Material online). The X-ray crystal structure of the full-length CYP51 from *S. cerevisiae* (ScErg11p), including the membrane-spanning region, was recently reported (Monk et al. 2014) and further supports the conclusion of the CYP73 modeling. Its crystal structure demonstrates that the predicted *C*-terminal TMH is not located in the membrane, but corresponds to an apolar helix buried in the globular fold of the protein, far from the membrane interface (supplementary fig. S8, Supplementary Material online). In CYP73 proteins, like in yeast CYP51, this apolar helix, referred to as helix L in the CYP topology defined by Ravichandran et al. (1993), is thus unlikely to be inserted in the membrane.

Our further analyses of the CYP73 protein topology thus focused on the *N*-terminal region, using a well-established in

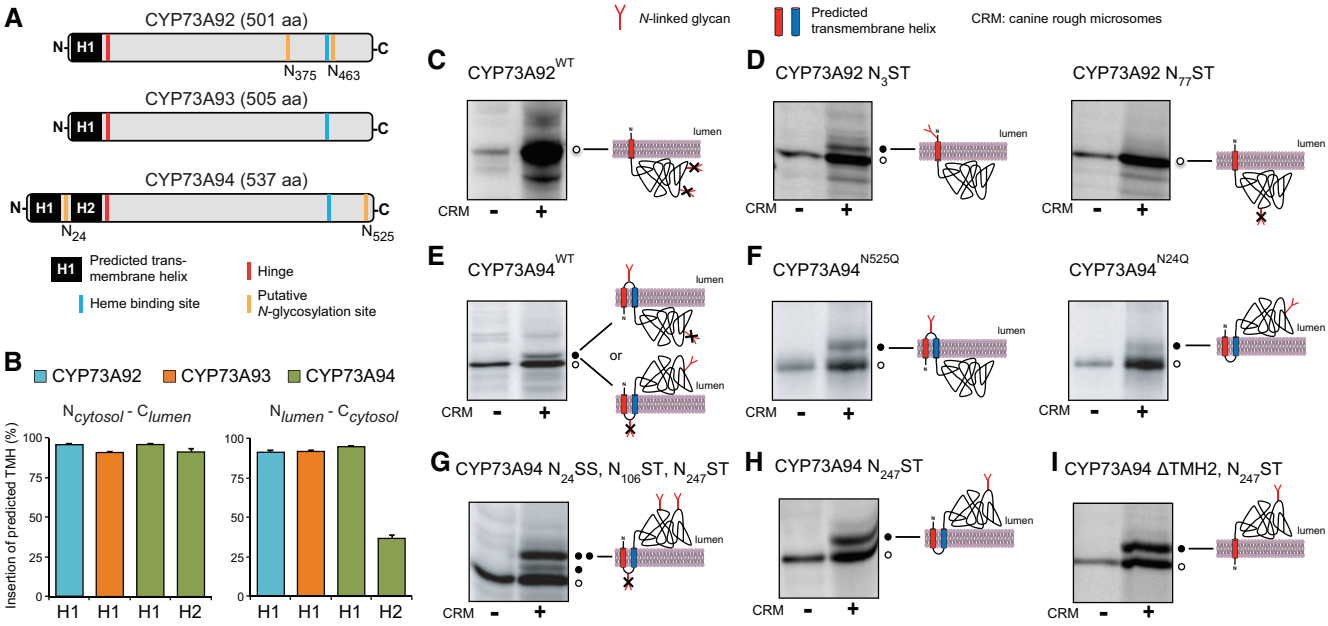

Fig. 6. Class II proteins have atypical membrane topology. (*A*) Cartoon depicting elements of the *B. distachyon* CYP73s primary structure predicted to influence the protein topology. (*B*) The efficiency of integration of the predicted transmembrane helices in the ER membrane of was determined using an in vitro system, where the predicted transmembrane helices were cloned into *E. coli* leader peptides (Hessa et al. 2005; Lundin et al. 2008). Error bars represent the SD of three independent experiments. (*C*–*F*) The membrane topology of the full-length CYP73s was studied using glycosylation sites as topology markers in an in vitro translation assays after addition of column-washed canine rough microsomes (CRM). Graphic on the right depicts the outcome of the assay. Open circles, nonglycosylated and/or untargeted protein; filled circles indicate the number of *N*-glycosylations. (*C*) In vitro translation of the native CYP73A92 protein. Note that a saturation pink color was part of the blot picture and was pseudo-colored in black. (*D*) Engineered glycosylation sites were introduced in CYP73A92 proteins (all natural glycosylation sites removed) at the *N*-terminus (N_3) or in the globular part of the protein (N_{77}), as indicated in the figure. (*E*) In vitro translation of the native CYP73A94 protein. (*F*) The native CYP73A94 protein bears two glycosylation sites, one (N_{24}) in the loop linking H1 and H2, and another (N_{525}) after the *C*-terminal H3. They were removed one at a time. (*G*) CYP73A94 bearing a single natural glycosylation site (N_{24}SS) was further engineered with introduction of two additional glycosylation sites (N_{106}ST and N_{247}ST). (H) An engineered glycosylation site (N_{247}) was introduced in the globular part of CYP73A94 protein (all natural glycosylation sites removed). (*I*) A natural *N*-glycosylation site-depleted version of CYP73A94 was further engineered with introduction of a glycosylation site located in the globular part of the protein (N_{247}ST) and by deletion of the second *N*-terminal transmembrane domain (ΔTMH2).

vitro system in which protein constructs are expressed in rabbit reticulocyte lysate in the presence of dog pancreas rough microsomes (Hessa et al. 2005; Lundin et al. 2008). The topology of a given construct is determined by inserting acceptor sites for *N*-glycosylation (that can only be glycosylated if translocated into the lumen of the rough microsomes) in strategic locations. We first tested whether the predicted transmembrane helices could actually insert into the ER membrane by themselves and whether they have an orientational preference, which could help determining the topology of the native protein. The predicted H1 and H2 helices (fig. 6*A*; supplementary table S5, Supplementary Material online), along with up to 20 amino acid residues from adjacent sequences, were individually inserted in the *Escherichia coli* leader peptidase (Lep) host protein as "H-segment" (supplementary fig. S9, Supplementary Material online), and tested in two opposite orientations (see Materials and Methods for detailed experimental strategy). These constructs were tested to determine whether the predicted TMHs have an orientational preference that could help to determine the topology of the native protein. The *N*-terminal H1 helix from both CYP73A92 and CYP73A93 inserted efficiently into the ER membrane, regardless of orientation (fig. 6*B*). The first helix (H1) of CYP73A94 inserted efficiently in both orientations, whereas the second (H2) inserted more efficiently when in the $N_{cytosol}$–C_{lumen} orientation (fig. 6*B*).

We also tested the membrane insertion of the H1–H2 segment from CYP73A94, using the Lep system. As expected from the results for the individual H1 and H2 helices, the H1–H2 segment inserted to >90% in a hairpin topology with both H1 and H2 spanning the membrane, regardless of whether its orientation was $N_{cytosol}$–$C_{cytosol}$ or N_{lumen}–C_{lumen} (supplementary fig. S10, Supplementary Material online).

Using the same in vitro translation system, we then investigated the topology of the full-length CYP73A92 and CYP73A94 proteins. When expressed in the presence of rough microsomes, wild type CYP73A92 remains nonglycosylated (fig. 6*C*). Since *N*-glycosylation occurs only in the lumen of the rough microsomes, this suggests that the catalytic domain, which has two putative glycosylation sites, faces the cytosol. This was confirmed by removing the two natural putative glycosylation sites and then adding either a short segment including a single glycolsylation site to the *N*-

terminus (CYP73A92 N_3ST), or introducing a single glycosylation site in the catalytic domain downstream of H1 (CYP73A92 N_{77}ST). The *N*-terminal glycosylation site was modified (albeit not very efficiently, probably because it is located rather close to H1), whereas the glycosylation site in the P450 domain was not (fig. 6*D*). We conclude that the H1 TMH anchors CYP73A92 in the ER membrane, with the catalytic domain facing the cytoplasm.

CYP73A94 has two putative *N*-glycosylation sites, one nested in the short loop between H1 and H2, and the other near the *C*-terminus (fig. 6*A*). When translated in the in vitro system, CYP73A94 only received a single glycan (fig. 6*E*). Interestingly, when either of the two putative glycosylation sites (N_{24}, N_{525}) was removed by asparagine (N) to glutamine (Q) substitutions, both mutated proteins were still mono-glycosylated. Thus, in a fraction of the molecules the N_{24} site located between H1 and H2 is glycosylated while the N_{525} site in the catalytic domain is not, whereas in another fraction the opposite is true. Given that the H1–H2 hairipin can insert efficiently in the ER membrane in both $N_{cytosol}$–$C_{cytosol}$ and N_{lumen}–C_{lumen} orientations (as shown above), the simplest interpretation is that CYP73A94 has two transmembrane helices (H1 and H2) and inserts into the ER with a dual topology, where in a fraction of the molecules the catalytic domain faces the cytosol, and in another fraction faces the lumen (fig. 6*F*). To provide further support for a dual topology, we removed the N_{525} site and added two new glycosylation sites in the catalytic domain (N_{106} and N_{247}). This construct was mainly di-glycosylated (fig. 6*G*). Another construct where only the N_{247} site was retained was mono-glycosylated (fig. 6*H*). When the H2 TMH was deleted from this construct, mono-glycosylated protein was still seen, again confirming that the H1 TMH can insert also with a $N_{cytosol}$–C_{lumen} orientation (fig. 6*I*).

Modeling of the Topology of CYP73s

The presence of two membrane-spanning segments at the *N*-terminus of the class II protein raises the questions of the geometry of the two apolar helices and of the protein orientation with regard to the membrane. To answer these questions, we first generated three-dimensional models of the transmembrane segments of *B. distachyon* CYP73 proteins folded as pure alpha-helical structures. The full protein experimental structure of yeast CYP51 (ScErg11p) recently reported (Monk et al. 2014) was used as a control for OPM (Orientations of Proteins in Membranes) (Lomize et al. 2006) database predictions. The membrane insertion/orientation parameters, including hydrophobic thickness, transfer energy and tilt angle determined by the PPM server (Lomize et al. 2006) revealed that the CYP73 predicted TMHs, including the two TMHs of CYP73A94, were fully compatible with membrane insertion (table 1). Conversely, the first amphipathic helix of the yeast CYP51 that lies on the inner side of the ER membrane is correctly predicted not to be a TMH.

Taking the predicted membrane-embedded residues and tilt angles into account, a model of the *B. distachyon* CYP73 membrane insertion topologies was generated. In agreement with our experimental data, this model proposes that the two class I CYP73s are inserted in the membrane via a single TMH, and the class II CYP73A94 via two TMH in opposite orientations, forming two antiparallel membrane-spanning domains arranged as an open hairpin (fig. 7). Charged and polar residues are found at the boundaries of the two trans-membrane helices, which suggests anchoring on the polar phospholipid heads on both membrane surfaces (fig. 7; supplementary fig. S11, Supplementary Material online).

Discussion

Our extensive mining of the most recent sequencing data is in line with the hypothesis that the CYP73 family of cytochrome P450 enzymes evolved upon plant colonization of land. It is present in bryophytes, but not in currently sequenced algal genome or transcriptome, and no clear ortholog can be spotted in more ancient phyla. Evolution of the C4H might thus constitute a determinant step for the colonization of land, essential for the efficient production of UV screens and biopolymer precursors. This hypothesis is further supported by the high conservation of the *CYP73* genes, present in all land plant genomes, and kept under strong purifying selection throughout evolution of land plants. This high negative selection is most likely required to maintain a constrained catalytic site structure, allowing high redox coupling efficiency. High coupling and catalytic efficiency are characteristic of C4H (Pierrel et al. 1994; Schalk et al. 1998) and ensure optimal use of the electrons from NAPDH for the formation of substrate without release of reactive oxygen. Additional and

Table 1. Comparison of the Membrane Insertion/Orientation Parameters of the *N*-Terminal Segments of *B. distachyon* CYP73 with Those of the Resolved Full-Length Yeast CYP51, Determined by the PPM Server.

N-Terminal Segments	Hydrophobic Thickness (Å)	$\Delta G_{transfer}$ (kcal/mol)	Tilt Angle (°)	Embedded Residues
ScErg11p (6-28)	**6.4 ± 3.2**	**−10.0**	**81 ± 0**	**22–23 + 25–28+ discrete residues 8, 12, 15, 18–19**
ScErg11p (6-55)	**28.2 ± 2.9**	**−27.1**	**37 ± 2**	**27–49**
CYP73A92 (1-33)	**30.0 ± 2.2**	**−22.3**	**17 ± 9**	**5–26**
CYP73A93 (1-33)	**28.4 ± 4.7**	**−23.6**	**26 ± 5**	**5–26**
CYP73A94 (1-26)	**28.6 ± 5**	**−20.7**	**39 ± 3**	**1–26**
CYP73A94 (25-61)	**30.6 ± 3.5**	**−26.5**	**39 ± 5**	**27–54**

NOTE.—The *N*-terminal sequences of the *B. distachyon* CYP73s have been modeled under Chimera as pure alpha-helical structures, from 1 to *n*. They are not experimental, neither obtained by homology modeling. They have been generated by constraining the polypeptide chain to adopt the ideal Phi/Psi dihedral angles of an α-helix. ScErg11p is the experimental structure of yeast CYP51 (#5EQB; itraconazole-bound) used as a control. The first helix of ScErg11p (6-28) is found to be amphiphilic, bound to inner leaflet of the ER membrane. The residues that are predicted to be embedded are partitioned in a discrete distribution, typical of an amphipathic helix. The second ScErg11p helix (27-49) is a transmembrane helix according to the crystal structure. By comparison, the two TMHs of CYP73A94 display features typical of fully embedded transmembrane segments.

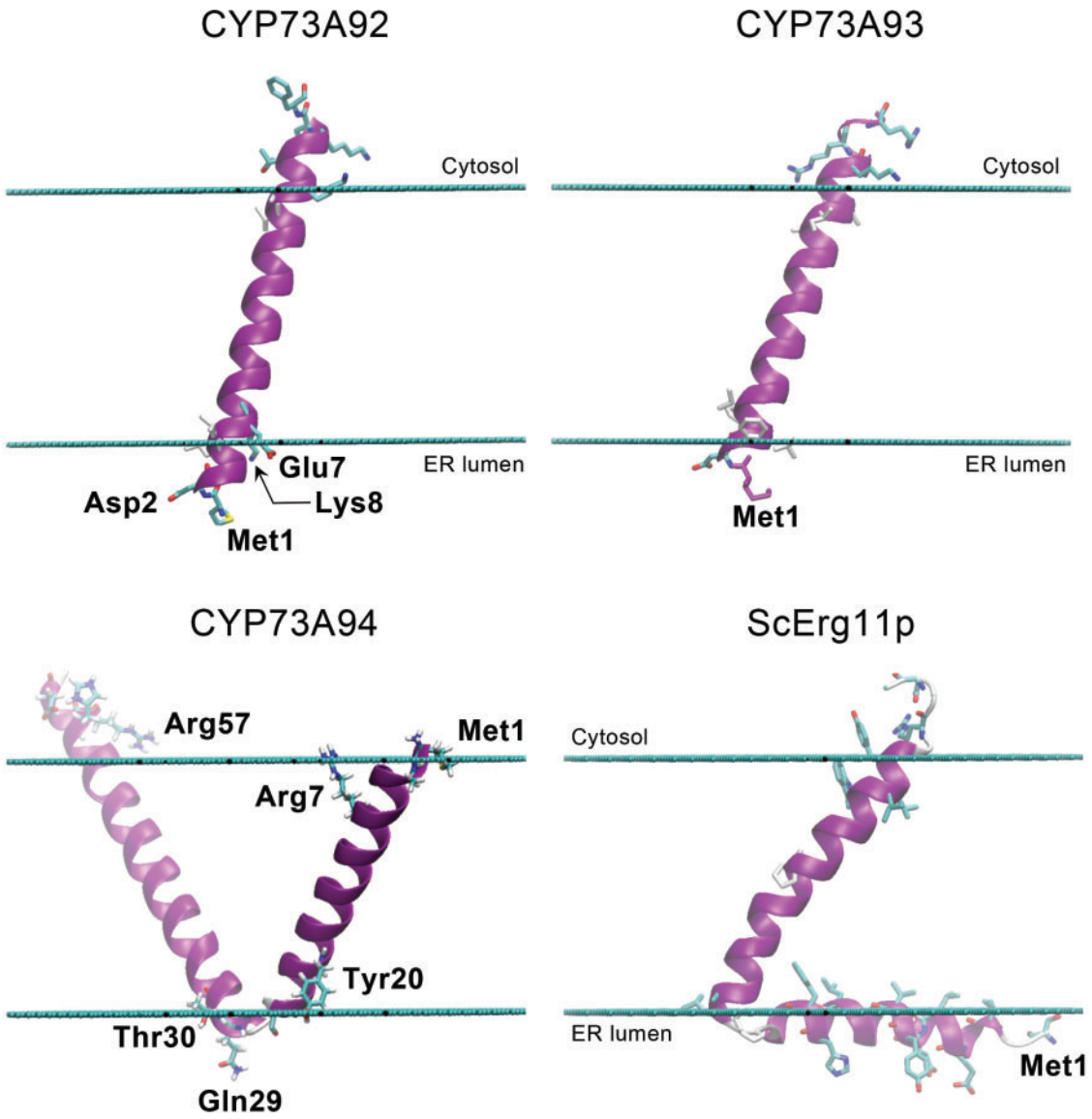

Fig. 7. Predicted membrane topology of the *B. distachyon* CYP73 proteins. Charged residues delimitating the interface with solvent are shown in atom type color mode. Some hydrophobic embedded residues of the helix at both edges are also displayed for illustrating the boundaries of hydrophobic embedded segments. The alignment used to determine the transmembrane helices, and a more detailed representation of the probable positioning of the *N*-terminal double trans-membrane helix of CYP73A94 are shown in supplementary figure S11, Supplementary Material online.

independent constraints on residues of the enzyme surface might result from the supramolecular organization of the pathway as functional units (Achnine et al. 2004; Bassard et al. 2012b). In support of this hypothesis, the strongest purifying selection applies on the clade that, in dicots, corresponds to the paralog most expressed in vascular tissues and involved in the synthesis of lignin (Sewalt et al. 1997; Lu et al. 2006; Millar et al. 2007; Schilmiller et al. 2009).

Single- or low-copy genes are often associated with essential functions (Liu et al. 2015, 2016). Correlated to this strong purifying selection is a very effective elimination of CYP73 duplicates. Although species-specific *CYP73* duplicates are observed, a single *CYP73* duplication became fixed throughout most of seed plant evolution. This duplication occurred in the common ancestor of gymnosperms and angiosperms. The loss of one of the duplicates (the class II paralog) is so far observed only in gymnosperms other than *Taxaceae*, and in *Brassicaceae*. This suggests that both duplicates hold specific functions, important enough for retention of each paralog. It also hints that the function of the class II protein has been assumed by another enzyme in *Brassicaceae*, possibly but not necessarily a class I CYP73.

Using as a proxy the *B. distachyon* paralogs, we determined the catalytic properties of class I and II proteins in vitro and in vivo, and demonstrated that both are very efficient cinnamate 4-hydroxylases. Striking differences are however observed in their primary structures, the main one at their *N*-terminus. The stretch extending from the *N*-terminus to the hinge connecting the catalytic domain is the membrane-anchoring segment of the protein. Comparison of this region

in class I and II CYP73s from all plant species hints to a clear specialization of class I that evolved to a fixed length and charge distribution at the *N*-terminus. We show that this constrained primary structure translates into a single membrane-spanning helix targeting the protein to the ER with a N_{lumen}–$C_{cytosol}$ orientation. A negative charge at the *N*-terminus and a typical stop-transfer cluster of positively charged residues at the membrane cytosolic surface most likely contribute to this protein orientation, as previously suggested for mammalian enzymes (Szczesna-Skorupa et al. 1988; von Heijne et al. 1989) and confirmed by TMH topology modeling of *B. distachyon* CYP73A92 and CYP73A93 (fig. 7). Conservation of the *N*-terminal TMH in class I CYP73 enzymes is higher than observed for orthologs in other P450 families. It might thus be required for retention into specific membrane microdomains or for interaction with partner proteins forming the lignin metabolon.

Conversely, much less conservation is observed in the *N*-terminus of class II proteins, about double in length, and variable in length and composition. The class II *N*-terminal sequences nevertheless maintain some conserved features: a large proportion of polar amino acid residues in its *N*-terminal half, a central cluster of prolines, and two TOPCONS-predicted TMHs. Still using as a proxy the *B. distachyon* enzyme, we demonstrate that this elongated *N*-terminus does not lead to a different subcellular localization, but instead appears to lead to an altered membrane topology. The protein is likely to be anchored in the ER membrane via two *N*-terminal transmembrane helices forming an open hairpin stabilized by ionic interactions with phospholipids. In addition, our results suggest a dual topology with two locations of the catalytic domain, either in the cytosol or in the ER lumen. This is an unprecedented case of versatile P450 topology, whether in plants or other organisms, and reveals a new and overlooked level of complexity in the plant phenolic pathway.

Then what is the selective advantage for the emergence and conservation of a second more versatile C4H? The functions of the phenolic derivatives are diverse, and all require involvement of a C4H activity. The open hairpin conformation of the membrane-spanning domain of the class II protein is likely to alter its mobility, lipid domain preferences, and to impact the formation of homomers or heteromers with the other enzymes of the pathway. We show that this modification does not preclude involvement in the biosynthesis of monolignols in *A. thaliana*, but it may favor feeding of nonlignin pathways such as those leading to flavonoids, stilbenes or coumarins. It might also favor interaction with dedicated *p*-coumaroyl ligases and divert the metabolic flux into specific phenolic pathways. Internalization of the class II protein in the ER could in particular support the production of secreted compounds or allow the enzyme itself to enter the secretory pathway. Some mammalian P450 enzymes with a single TMH have been reported to enter the secretory pathway and to be targeted to the plasma membrane (Neve and Ingelman-Sundberg 2008), some of them facing the extracellular space. The latter orientation implies at some point a topological inversion, within the ER or later during transport to the cell surface. For such export, lumen-oriented P450s, like class II CYP73s, would be expected to be favored. In our subcellular localization of GFP fused protein, due to close association of ER with the cellular membrane, the presence of enzyme on the plasma membrane was impossible to assert and thus cannot be excluded. However, in the tissues investigated, no enzyme was found associated with the Golgi apparatus or other endomembranes besides ER.

Another intriguing question raised by the dual topology of class II CYP73s is: what is the electron donor to the lumen-oriented protein? Cross-membrane electron transfer has been suggested to occur with a mammalian P450 enzyme (Loeper et al. 1998). However, a duplicate of the P450 oxidoreductase (POR), elongated at the *N*-terminus, is present in vascular plant genomes, with as prototype *A. thaliana* ATR2 (http://www.p450.kvl.dk/At_rel/CPRSeqs.html#AtATR2; last accessed May 22, 2017). The serine-enrichment of the *N*-terminus of this elongated POR is similar to that observed for class II CYP73s. The subcellular localization of the ATR2 ortholog from hybrid poplar has been investigated (Ro et al. 2002), with as conclusion a strict ER localization. This elongated POR is thus a plausible redox partner for class II CYP73s. It must also be pointed out that an elongated *N*-terminus is not the only conserved feature of class II CYP73s. A four to five amino-acid insertion, located 13 residues upstream of the conserved heme-binding cysteine, and including positively charged residues, is present in all class II proteins (supplementary fig. S4, Supplementary Material online). Homology modeling locates this insertion in a loop on the heme proximal surface of the protein (supplementary fig. S12, Supplementary Material online), a region described as a contact with the electron donor, and associated with electron transfer (Im and Waskell 2011). The presence of this insertion suggests the possibility of an alternative electron donor for the class II proteins.

A preliminary *in silico* investigation indicates a differential but overlapping expression of class I and II *CYP73* genes in the different organs of *B. distachyon*. Concerted expression and metabolic analyses at the tissue level are thus required to further describe the functional specialization of the CYP73 paralogs in angiosperms.

Materials and Methods

Phylogenetic Analyses

Transcriptome and Genome Sequences Mining

Brachypodium distachyon CYP73A93 homologs were searched by BLASTp (*e*-value threshold $<10^{-100}$) in algae and land plant whole-genome sequences available through Phytozome (http://phytozome.jgi.doe.gov; last accessed May 22, 2017). Genome of the charophyte algae *Klebsormidium nitens* was mined on its webpage (http://www.plantmorphogenesis.bio.titech.ac.jp/~algae_genome_project/klebsormidium/; last accessed May 22, 2017). *Tarenaya hassleriana*, *Aethionema arabicum*, and *Lotus japonicus* genomes were mined through the CoGe database (https://genomevolution.org/CoGe/; last accessed May 22, 2017) by tBLASTn (*e*-value threshold $<10^{-100}$). Search for CYP73 homologs was

extended to transcriptome data generated by the One Thousand Plants Project (1kP; www.onekp.com; last accessed May 22, 2017) by tBLASTn (*e*-value threshold $<10^{-100}$). Additional CYP73 homologs were also retrieved from NCBI nonredundant protein database (nr) via BLASTp search (*e*-value threshold $<10^{-100}$). Positive hits were BLAST queried against the *Brachypodium distachyon* genome and discarded when best hits were not CYP73.

BLAST search in algae genomes and transcriptomes typically yielded hits below the *e*-value threshold, with protein identity far below the family defining criterion (i.e., >40%). In order to detect any remote homology, the hidden Markov model-based software pHMMER v3.1b2 (http://hmmer.org/; last accessed May 22, 2017) was implemented. Best hits were then back searched in the thoroughly annotated *Arabidopsis thaliana* genome (supplementary table S3, Supplementary Material online).

All cytochromes P450 from *Chlamydomonas reinhardtii* and *Klebsormidium nitens* were retrieved using the pfam motif PF00067 (http://pfam.xfam.org/family/PF00067; last accessed May 22, 2017).

Sequences Alignment

The CYP73 data set contains 204 ORF derived from 118 land plant species (supplementary table S4, Supplementary Material online). Multiple sequences alignment (MSA) was built using MUSCLE algorithm (Edgar 2004) based on amino acids sequences, resulting in the "CYP73_full_MSA" (supplementary data set S1, Supplementary Material online). Alignment was carefully inspected and manually curated when necessary. Odd gaps present in protein sequences retrieved from whole-genome sequences were individually checked for annotation problems and altered when relevant. Ambiguous regions of the alignment were masked using Gblocks with default settings, except that shorter blocks were allowed (Castresana 2000), yielding the "CYP73_Gblocks_ MSA" (supplementary data set S2, Supplementary Material online).

Alignment of representative CYP73 and CYP98 proteins with all cytochromes P450 from *C. reinhardtii* and *K. nitens* was performed with MUSCLE. Ambiguous sites were removed with Gblocks using the least stringent options, generating the "Chlamydomonas_Klebsormidium_CYP73_CYP98_Gblocks_MSA" (supplementary data set S3, Supplementary Material online).

Phylogeny Reconstruction

Phylogeny of the CYP73 family was inferred from its nucleotide "CYP73_Gblocks_MSA." Maximum likelihood tree was built using the GTR + G+I (general time reversible) model and the PhyML 3.0 software (Guindon et al. 2010). Initial tree(s) for the heuristic search were obtained automatically by applying the BioNJ algorithm, and then selecting the topology with superior log likelihood value. Best of nearest neighbor interchange (NNI) and subtree pruning and regrafting (SPR) methods was used for improving tree.

Phylogenetic relationships of representative CYP73 and CYP98 sequences with all cytochromes P450 from *C. reinhardtii* and *K. nitens* were inferred from the "Chlamydomonas_Klebsormidium_CYP73_CYP98_Gblocks_MSA" by using PhyML 3.0 based on the LG + G+I + F′ model (Le and Gascuel 2008). Initial tree(s) for the heuristic search were obtained automatically by applying the BioNJ algorithm, and then selecting the topology with superior log likelihood value. The nearest neighbor interchange (NNI) method was used for improving tree.

We tested tree branches support by running 100 nonparametric standard bootstrap replications.

Molecular Evolution (*d*N/*d*S) Analysis.

PAML v4.8 (http://abacus.gene.ucl.ac.uk/software/paml.html; last accessed May 22, 2017) and Fitmodel v0.5.3 (https://github.com/stephaneguindon/fitmodel; last accessed May 22, 2017) were run on the "CYP73_full_MSA" described above and the phylogenetic tree visible in figure 1. We compared nested models by performing likelihood-ratio tests (LRT) in order to assess the statistical significance of the difference (Anisimova et al. 2001). All results are visible in supplementary note S1, Supplementary Material online.

Recombinant Proteins Production in Yeast

Coding sequences (CDS) of *Brachypodium CYP73A92*, *CYP73A93* and *CYP73A94* genes were optimized for yeast expression (sequences in supplementary note S2, Supplementary Material online) and synthesized by GeneCust Europe (Dudelange, Luxembourg) with addition of *Bam*HI and *Kpn*I restriction sites at the 5′ and 3′ termini, respectively. CDS of *Physcomitrella patens CYP73A48* was PCR-amplified from protonema cDNA (Grandsen strain) using primers containing *Bam*HI and *Kpn*I restriction sites and TA-cloned into the pGEM-T Easy vector, resulting in the pHR0153 vector. *Bam*HI/*Kpn*I sites were used for subcloning CDS into the yeast expression plasmid pYeDP60, resulting in the pHR0056, pHR0057, pHR0058, and pHR0164 expression plasmids. The *Saccharomyces cerevisiae* WAT11 strain was transformed with expression plasmids and selected as described before (Liu et al. 2016). Recombinant protein production and microsomes preparation procedures were performed according to (Liu et al. 2016). Microsomal membrane preparations containing CYP73 recombinant proteins were stored at −20 °C until processing.

Enzymes Kinetics

Standard assay for cinnamate 4-hydroxylase (C4H) was performed in a 100 µl reaction containing 50 mM potassium phosphate buffer (pH 7.4), 0.1–0.25 pmoles of yeast recombinant CYP73, 0.25–25 µM *trans*-cinnamic acid and 500 µM NADPH. The reaction was initiated by addition of NADPH, incubated at 28 °C in the dark for 3 min and terminated with 10 µl of 50% acetic acid. Substrate conversion never exceeded 10%. Microsomal fraction of enzyme assays was spun down by a 5 min centrifugation at 15,000 × g and 4 °C. Supernatant was recovered and *p*-coumaric acid (*p*CA) was analyzed by reverse-phase HPLC (Alliance 2695, Waters) with photo-

diode array detection at 310 nm (Photodiode 2996, Waters). To this end, 75 µl of sample were injected onto a Kinetex® core-shell C18 5 µm 4.6 × 150 mm column (Phenomenex) operated at 37 °C. Separation was carried out with 0.1% formic acid in water (A) and 0.1% formic acid in acetonitrile (B) at a 1 ml min^{-1} flow rate. Elution program was as follows: 5–100% B in 16 min (concave curve 8), 100% B for 1 min. Quantitative determination of *p*-coumaric acid was performed according to external calibration curve of authentic compound. Enzyme kinetics was determined with the GraphPad software and Michaelis–Menten nonlinear regression. Binding constant determination was performed by recording absorption difference spectra between the 390 and 420 nm wavelengths after addition of ligand to the medium.

Cloning of *CYP73* Coding Sequences and Construction of Fluorescent Fusion

Coding sequences (CDS) from *CYP73A92*, *CYP73A93*, and *CYP73A94* were PCR-amplified from *Brachypodium distachyon* (BD21 accession) cDNA using Gateway-compatible primers. PCR fragments were cloned into pDONR207 by BP recombination, generating the pHR0082 (*CYP73A92*), pHR0083 (*CYP73A93*), and pHR0084 (*CYP73A94*) pENTRY vectors. STOP codon-free CDS from *CYP73A92*, *CYP73A93*, and *CYP73A94* were likewise PCR-amplified using USER-compatible primers. Amplicons were cloned into the pCAMBIA2300uGFP plasmid (Bassard et al. 2012a) for 35S promoter-driven translational fusion with EGFP, resulting in the pHR0001, pHR0003, and pHR0005 plasmids, respectively. *CYP73A94:EGFP* coding sequence was reamplified using Gateway-compatible primers and cloned into the pDONR207 by BP recombination, generating the pHR0046 (*CYP73A94:EGFP*) pENTRY vector. *CYP73s* and *CYP73A94:EGFP* coding sequences were subsequently transferred by LR recombination into the pCC01061 (kanamycin resistance in plant) and pCC0996 (BASTA resistance in plant) vectors, respectively, which contain a 2977bp promoter fragment from *A. thaliana CYP73A5* gene (Weng et al. 2011). Resulting plant expression plasmids were named pHR0090 ($AtC4H_{pro}$:*CYP73A92*), pHR0091 ($AtC4H_{pro}$:*CYP73A93*), pHR0092 ($AtC4H_{pro}$:*CYP73A94*), and pHR0264 ($AtC4H_{pro}$:*CYP73A94:EGFP*). Primers used for molecular work are listed in supplementary table S6, Supplementary Material online.

Protein Membrane Topology Determination

Integration Assay of Predicted Transmembrane Helices

The membrane integration of predicted transmembrane helices (TMH) from CYP73A92, CYP73A93, and CYP73A94 were tested in microsomal membranes using a well-established translation based system (Hessa et al. 2005; Lundin et al. 2008). Protein chimeras were constructed by placing CYP73-derived segments within *Escherichia coli* leader peptidase (Lep) as H-segments (supplementary fig. S9, Supplementary Material online). To allow Lep to "host" protein segments from the three different CYP73As, *Spe*I and *Kpn*I restriction recognition sites had been introduced in the sequence encoding either the middle of the P2-domain or to replace Lep TMH2 (supplementary fig. S9, Supplementary Material online). In the latter construct, the *N*-terminal tail of Lep had been lengthened to allow efficient glycosylation of the upstream glycosylation site (Nilsson and von Heijne 1993). Native glycosylation sites NX[S/T] (Silberstein and Gilmore 1996) were eliminated by substitution to QX[S/T] (where X can be any amino acid except proline (Shakin-Eshleman et al. 1996) and new glycosylation acceptor sites were engineered into Lep as shown in supplementary figures S9 and S10, Supplementary Material online. Lep itself was introduced into a pGEM-1 vector (Promega Biotech AB, Madison, WI) allowing expression under the SP6-promoter and the 5′ of the initiator codon was changed to a Kozak consensus sequence to enhance translation efficiency (Kozak 1989). Double-stranded oligonucleotides (MWG Biotech AG, Ebersberg, Germany) of the predicted transmembrane regions from CYP73As, with up to 20 amino acids at both *N* and *C* terminus (supplementary table S5, Supplementary Material online), were introduced as their yeast optimized version into the *lepB* gene as *Spe*I-*Kpn*I PCR-amplified fragments.

Topology Assay of Individual CYPs

The genes for full-length CYP73A92 and CYP73A94 were cloned as their yeast optimized version under the SP6-promoter in the pGEM-1 vector using *Xba*I and *Sma*I restriction sites. All removal of natural glycosylation sites or introduction of new-to-nature glycosylation sites was using the site-directed mutagenesis method.

In vitro Translation

All constructs were transcribed and translated in the TNT SP6 Quick Coupled System (Promega Biotech AB, Madison, WI). 100 ng DNA was mixed with a master mix consisting of 0.5 µl ^{35}S-methionine (5 µCi) (PerkinElmer, Boston, MA). In positive samples, 0.5 µl column-washed canine pancreas rough microsomes (tRNAprobes, College Station, TX) were added. All samples were incubated at 30 °C for 90 min.

SDS-PAGE and Analysis

Translated proteins were separated by SDS-PAGE and visualized with a Fuji FLA-3000 phosphoimaginer (Fujifilm, Tokyo, Japan) with the Image Reader V1.8J/Image Gauge v3.45 software (Fujifilm). The MultiGauge software was used to create one-dimensional intensity profile for each lane on the gels. These profiles were then analyzed using the multi-Gaussian fit program from the Qtiplot software package (http://www.qtiplot.com/; last accessed May 22, 2017), and the peak areas of the glycosylated protein bands in the profile were calculated.

cyp73a5-1 *Loss-of-Function Mutant Complementation*

The GABI-Kat line GK_753B06 (Col-0 background) (Rosso et al. 2003), which contains a T-DNA insertion in the first exon of *CYP73A5*, was obtained from the European Arabidopsis Stock Centre and named *cyp73a5-1*. Heterozygous *cyp73a5-1* plants were transformed with the above-described pHR0090, pHR0091, pHR0092, and

pHR0264 vectors by the floral dip method (Clough and Bent 1998) using *Agrobacterium tumefaciens* c58c1 strain. Transformants were selected based on their resistance to either kanamycin (pHR0090-0092) or BASTA (pHR0264). T1 plants homozygous for the *cyp73a5-1* allele were identified by PCR (primers in supplementary table S6, Supplementary Material online) and further confirmed by the full resistance of T2 progeny to sulfadiazine.

Protein Subcellular Localization

For transient overexpression of EGFP fusion proteins, pHR0001, pHR0003, and pHR0005 vectors were introduced into 4-week-old *Nicotiana benthamiana* leaves by agroinfiltration as described previously (Bassard et al. 2012b). Leaf disks were excised 4 days postinfiltration for confocal observation. A previously described mRFP1-ER marker construct was concurrently introduced into *N. benthamiana* leaves (Bassard et al. 2012b). For stable expression of CYP73A94:EGFP fusion protein, pHR0264-complemented *cyp73a5-1* plants were sown in plates on half-strength MS medium solidified with 0.8% agar. *CYP73A5* promoter activity was enhanced by supplementing medium with 2% sucrose (Solfanelli et al. 2006). After 2–3 days of stratification, plates were moved to a 16 h-light growth chamber (22 °C/18 °C day/night temperatures). Five-day-old plants were used for experiments. Live imaging was performed using a Zeiss LSM780 laser scanning confocal microscope. Single section images were recorded using a Plan-Apochromat 63×/1.40 Oil DIC M27 objective lens. For *N. benthamiana* epidermal cells live imaging, EGFP and mRFP1 fluorescence were sequentially recorded using the excitation/emission wavelengths 488/495–555 nm and 561/565–615 nm, respectively. For *A. thaliana* live imaging, EGFP and propidium iodide signals were simultaneously recorded using the excitation/emission wavelengths 488/495–555 nm and 488/595–720 nm, respectively.

Structural Bioinformatics Tools

Three-Dimensional Models of N-Terminal Domains

Three dimensional models of isolated N-terminal segments of *Brachypodium distachyon* CYP73A92, CYP73A93, and CYP73A94 were rebuilt under constrained α-helical structures using Chimera molecular modeling suite developed by the UCSF Resource for Biocomputing, Visualization, and Informatics (Pettersen et al. 2004). The modeled N-terminal segments selected for transmembrane positioning calculations were the followings: CYP73A92 (1-33), CYP73A93 (1-33), CYP73A94 (1-26), and CYP73A94 (25-61). Under Chimera, the selected side-chain orientations were based on Dunbrack rotamer library as a first approach. In some cases, in CYP73A92 and CYP73A93 peptides, the orientation of some residues has been modified by selecting the second or third high-probability rotamer of the side-chain in the rotamer library. This allowed optimizing the α-helical structure for membrane spanning. The selection of alternative sidechain rotamers improved the interaction of the model helical peptide with the polar groups of the lipids, as revealed by the gain in the $\Delta G_{transfer}$ calculated by PPM web server, which could be lowered by up to 4 kcal/mol.

Spatial Arrangements Calculations by OPM Database

The PDB coordinates of model CYP73 N-terminal segments were submitted to PPM server (http://opm.phar.umich.edu/server.php; last accessed May 22, 2017) for calculating their positioning in a lipid bilayer of adjustable thickness by minimizing the ΔG transfer energy from water to the membrane. As a control, experimental N-terminal helical segments of ScErg11p (CYP51), the first crystal structure for any full-length cytochrome P450 enzyme showing resolution of the membrane spanning helix available in PDB, were also submitted to PPM server. The two N-terminal helices (the amphipathic 6-28, and the TMH 27-52) from two different crystal structures (4LXJ, lanosterol, and 5EQB, itraconazole cocrystals) were used. Structural prediction of transmembrane spanning segments was done by comparing the geometrical (hydrophobic thickness, rotational and translational positions with tilt angles and number of residues spanning), and energetic features ($\Delta G_{transfer}$ in kcal/mol) computed by PPM server. An amphipathic helix can be detected by smaller $\Delta G_{transfer}$ values (10 kcal/mol or less), or short hydrophobic thickness (lower than 8–10 Å), with a list of noncontiguous embedded residues.

Homology Modeling of Oxygenase and Full-Length of BdCYP73 Paralogs

3D models of *B. distachyon* CYP73A92, CYP73A93 and CYP73A94 were first generated deprived of their membrane spanning domain using Modeller9v12 with a procedure similar to that described in (Liu et al. 2016), with mammalian CYP2 PDB templates three-dimensionally aligned prior to multiple sequence alignment with CYP73 targets. Multiple alignments were performed with MAFFT-L-INS-I algorithm. The truncated 3D models were assessed using the same scoring functions detailed in (Liu et al. 2016).

For positioning the 3D-models of *B. distachyon* CYP73 paralogs with respect to the membrane surface, approximate full-length structures were also generated using Modeller by structural alignment of the rebuilt oxygenase domains with two PDB templates of full-length ScCYP51 (PDB codes 4LXJ and 5EQB, lanosterol- and itraconazole-bound, respectively) (Monk et al. 2014). Modeller was applied to reconstruct the N-terminal helical segment of CYP73A92, CYP73A93, and the first transmembrane helix of CYP73A94 attached to the oxygenase domain. These hybrid homology models were essentially used to confirm the OPM parameters of orientation and embedment, and to help the anchoring of oxygenase domains modeled at the membrane interface.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Author Contributions

H.R., D.W., M.D.M, G.v.H, I.M.N., and F.A. conceived the project. H.R. planned and performed most of the experiments, except for in vitro protein expression (M.D.M. and P.L.) and homology modeling (F.A. and G.J.). D.N. checked for the

presence of algal CYP73 in his full P450 database. D.W., F.A., G.v.H. and I.M.N. obtained funding. D.W. and H.R. wrote the initial draft, which was subsequently edited by all authors. Authors declare no competing interests.

Acknowledgments

Authors are thankful to Prof. Clint Chapple (Purdue University, USA) for providing the pCC0996 and pCC1061 vectors. We are grateful to Etienne Grienenberger for the critical reading of the manuscript and to Stéphane Guindon (University of Auckland) for his help with the Fitmodel software. This project received support from the Agence Nationale pour la Recherche to the PHENOWALL ANR-10-BLAN-1528 project, and from the Swedish Cancer Foundation (130624) to I.N. and G.v.H., from the Swedish Foundation for International Cooperation in Research and Higher Education (STINT) (210/083(12) and KU 2003-4674) to I.N., and from the Swedish Foundation for Strategic Research (A302:200) to I.N. and G.v.H.

References

Achnine L, Blancaflor EB, Rasmussen S, Dixon RA. 2004. Colocalization of L-phenylalanine ammonia-lyase and cinnamate 4-hydroxylase for metabolic channeling in phenylpropanoid biosynthesis. *Plant Cell* 16:3098–3109.

Anisimova M, Bielawski JP, Yang Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol.* 18:1585–1592.

Bak S, Beisson F, Bishop G, Hamberger B, Höfer R, Paquette S, Werck-Reichhart D. 2011. Cytochromes P450. *Arabidopsis Book* 9:e0144.

Barros J, Serrani-Yarce JC, Chen F, Baxter D, Venables BJ, Dixon RA. 2016. Role of bifunctional ammonia-lyase in grass cell wall biosynthesis. *Nat Plants* 2:16050.

Bassard JE, Mutterer J, Duval F, Werck-Reichhart D. 2012a. A novel method for monitoring the localization of cytochromes P450 and other endoplasmic reticulum membrane associated proteins: a tool for investigating the formation of metabolons. *FEBS J.* 279:1576–1583.

Bassard JE, Richert L, Geerinck J, Renault H, Duval F, Ullmann P, Schmitt M, Meyer E, Mutterer J, Boerjan W, et al. 2012b. Protein–protein and protein-membrane associations in the lignin pathway. *Plant Cell* 24:4465–4482.

Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S. 2010. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A.* 107:18724–18728.

Bell-Lelong DA, Cusumano JC, Meyer K, Chapple C. 1997. Cinnamate-4-hydroxylase expression in Arabidopsis. Regulation in response to development and the environment. *Plant Physiol.* 113:729–738.

Bernsel A, Viklund H, Hennerdal A, Elofsson A. 2009. TOPCONS: consensus prediction of membrane protein topology. *Nucleic Acids Res.* 37:465–468.

Boerjan W, Ralph J, Baucher M. 2003. Lignin biosynthesis. *Annu Rev Plant Biol.* 54:519–546.

Bowman JL. 2013. Walkabout on the long branches of plant evolution. *Curr Opin Plant Biol.* 16:70–77.

Brkljacic J, Grotewold E, Scholl R, Mockler T, Garvin DF, Vain P, Brutnell T, Sibout R, Bevan M, Budak H, et al. 2011. *Brachypodium* as a model for the grasses: today and the future. *Plant Physiol.* 157:3–13.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17:540–552.

Chang C, Bowman JL, Meyerowitz EM. 2016. Field guide to plant model systems. *Cell* 167:325–339.

Chen HC, Li Q, Shuford CM, Liu J, Muddiman DC, Sederoff RR, Chiang VL. 2011. Membrane protein complexes catalyze both 4- and 3-hydroxylation of cinnamic acid derivatives in monolignol biosynthesis. *Proc Natl Acad Sci U S A.* 108:21253–21258.

Clough SJ, Bent AF. 1998. Floral dip: a simplified method for Agrobacterium-mediated transformation of *Arabidopsis thaliana*. *Plant J.* 16:735–743.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

Ehlting J, Hamberger B, Million-Rousseau R, Werck-Reichhart D. 2006. Cytochromes P450 in phenolic metabolism. *Phytochem Rev.* 5:239–270.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 59:307–321.

Guindon S, Rodrigo AG, Dyer KA, Huelsenbeck JP. 2004. Modeling the site-specific variation of selection patterns along lineages. *Proc Natl Acad Sci U S A.* 101:12957–12962.

Hessa T, Kim H, Bihlmaier K, Lundin C, Boekel J, Andersson H, Nilsson I, White SH, von Heijne G. 2005. Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* 433:377–381.

Im SC, Waskell L. 2011. The interaction of microsomal cytochrome P450 2B4 with its redox partners, cytochrome P450 reductase and cytochrome b(5). *Arch Biochem Biophys* 507:144–153.

International Brachypodium Initiative 2010. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* 463:763–768.

Kenrick P, Crane PR. 1997. The origin and early evolution of plants on land. *Nature* 389:33–39.

Kozak M. 1989. Context effects and inefficient initiation at non-AUG codons in eucaryotic cell-free translation systems. *Mol Cell Biol.* 9:5073–5080.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. *Mol Biol Evol.* 25:1307–1320.

Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, De Smet R. 2016. Gene duplicability of core genes is highly consistent across all angiosperms. *Plant Cell* 28:326–344.

Liu Z, Boachon B, Lugan R, Tavares R, Erhardt M, Mutterer J, Demais V, Pateyron S, Brunaud V, Ohnishi T, et al. 2015. A conserved cytochrome P450 evolved in seed plants regulates flower maturation. *Mol Plant* 8:1751–1765.

Liu Z, Tavares R, Forsythe ES, Andre F, Lugan R, Jonasson G, Boutet-Mercey S, Tohge T, Beilstein MA, Werck-Reichhart D, et al. 2016. Evolutionary interplay between sister cytochrome P450 genes shapes plasticity in plant metabolism. *Nat Commun.* 7:13026.

Loeper J, Louerat-Oriou B, Duport C, Pompon D. 1998. Yeast expressed cytochrome P450 2D6 (CYP2D6) exposed on the external face of plasma membrane is functionally competent. *Mol Pharmacol.* 54:8–13.

Lomize MA, Lomize AL, Pogozheva ID, Mosberg HI. 2006. OPM: orientations of proteins in membranes database. *Bioinformatics* 22:623–625.

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

Lundin C, Kim H, Nilsson I, White SH, von Heijne G. 2008. Molecular code for protein insertion in the endoplasmic reticulum membrane is similar for N(in)-C(out) and N(out)-C(in) transmembrane helices. *Proc Natl Acad Sci U S A.* 105:15702–15707.

Millar DJ, Long M, Donovan G, Fraser PD, Boudet AM, Danoun S, Bramley PM, Bolwell GP. 2007. Introduction of sense constructs of cinnamate 4-hydroxylase (CYP73A24) in transgenic tomato plants shows opposite effects on flux into stem lignin and fruit flavonoids. *Phytochemistry* 68:1497–1509.

Monk BC, Tomasiak TM, Keniya MV, Huschmann FU, Tyndall JD, O'Connell JD 3rd, Cannon RD, McDonald JG, Rodriguez A, et al. 2014. Architecture of a single membrane spanning cytochrome P450 suggests constraints that orient the catalytic domain relative to a bilayer. *Proc Natl Acad Sci U S A.* 111:3865–3870.

Nedelkina S, Jupe SC, Blee KA, Schalk M, Werck-Reichhart D, Bolwell GP. 1999. Novel characteristics and regulation of a divergent cinnamate 4-hydroxylase (CYP73A15) from French bean: engineering expression in yeast. *Plant Mol Biol.* 39:1079–1090.

Neve EP, Ingelman-Sundberg M. 2008. Intracellular transport and localization of microsomal cytochrome P450. *Anal Bioanal Chem.* 392:1075–1084.

Nilsson IM, von Heijne G. 1993. Determination of the distance between the oligosaccharyltransferase active site and the endoplasmic reticulum membrane. *J Biol Chem.* 268:5798–5801.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera: a visualization system for exploratory research and analysis. *J Comput Chem.* 25:1605–1612.

Pierrel MA, Batard Y, Kazmaier M, Mignotte-Vieux C, Durst F, Werck-Reichhart D. 1994. Catalytic properties of the plant cytochrome P450 CYP73 expressed in yeast. Substrate specificity of a cinnamate hydroxylase. *Eur J Biochem.* 224:835–844.

Ravichandran KG, Boddupalli SS, Hasermann CA, Peterson JA, Deisenhofer J. 1993. Crystal structure of hemoprotein domain of P450BM-3, a prototype for microsomal P450's. *Science* 261:731–736.

Renault H, Alber A, Horst NA, Basilio Lopes A, Fich EA, Kriegshauser L, Wiedemann G, Ullmann P, Herrgott L, Erhardt M, et al. 2017. A phenol-enriched cuticle is ancestral to lignin evolution in land plants. *Nat Commun.* 8:14713.

Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud PF, Lindquist EA, Kamisugi Y, et al. 2008. The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science* 319:64–69.

Ro DK, Ehlting J, Douglas CJ. 2002. Cloning, functional expression, and subcellular localization of multiple NADPH-cytochrome P450 reductases from hybrid poplar. *Plant Physiol.* 130:1837–1851.

Rosso MG, Li Y, Strizhov N, Reiss B, Dekker K, Weisshaar B. 2003. An *Arabidopsis thaliana* T-DNA mutagenized population (GABI-Kat) for flanking sequence tag-based reverse genetics. *Plant Mol Biol.* 53:247–259.

Schalk M, Cabello-Hurtado F, Pierrel MA, Atanossova R, Saindrenan P, Werck-Reichhart D. 1998. Piperonylic acid, a selective, mechanism-based inactivator of the trans-cinnamate 4-hydroxylase: a new tool to control the flux of metabolites in the phenylpropanoid pathway. *Plant Physiol.* 118:209–218.

Schilmiller AL, Stout J, Weng JK, Humphreys J, Ruegger MO, Chapple C. 2009. Mutations in the cinnamate 4-hydroxylase gene impact metabolism, growth and development in Arabidopsis. *Plant J.* 60:771–782.

Sewalt V, Ni W, Blount JW, Jung HG, Masoud SA, Howles PA, Lamb C, Dixon RA. 1997. Reduced lignin content and altered lignin composition in transgenic tobacco down-regulated in expression of L-phenylalanine ammonia-lyase or cinnamate 4-hydroxylase. *Plant Physiol.* 115:41–50.

Shakin-Eshleman SH, Spitalnik SL, Kasturi L. 1996. The amino acid at the X position of an Asn-X-Ser sequon is an important determinant of N-linked core-glycosylation efficiency. *J Biol Chem.* 271:6363–6366.

Silberstein S, Gilmore R. 1996. Biochemistry, molecular biology, and genetics of the oligosaccharyltransferase. *FASEB J.* 10:849–858.

Solfanelli C, Poggi A, Loreti E, Alpi A, Perata P. 2006. Sucrose-specific induction of the anthocyanin biosynthetic pathway in Arabidopsis. *Plant Physiol.* 140:637–646.

Szczesna-Skorupa E, Browne N, Mead D, Kemper B. 1988. Positive charges at the NH2 terminus convert the membrane-anchor signal peptide of cytochrome P-450 to a secretory signal peptide. *Proc Natl Acad Sci U S A.* 85:738–742.

Teutsch HG, Hasenfratz MP, Lesot A, Stoltz C, Garnier JM, Jeltsch JM, Durst F, Werck-Reichhart D. 1993. Isolation and sequence of a cDNA encoding the Jerusalem artichoke cinnamate 4-hydroxylase, a major plant cytochrome P450 involved in the general phenylpropanoid pathway. *Proc Natl Acad Sci U S A.* 90:4102–4106.

von Heijne G, Steppuhn J, Herrmann RG. 1989. Domain structure of mitochondrial and chloroplast targeting peptides. *Eur J Biochem.* 180:535–545.

Weng JK, Akiyama T, Ralph J, Chapple C. 2011. Independent recruitment of an O-methyltransferase for syringyl lignin biosynthesis in *Selaginella moellendorffii*. *Plant Cell* 23:2708–2724.

Weng JK, Chapple C. 2010. The origin and evolution of lignin biosynthesis. *New Phytol.* 187:273–285.

Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc Natl Acad Sci U S A.* 111:E4859–48E4868.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.