PNAS RESEARCH ARTICLE | PLANT BIOLOGY

Elucidation of the 1-phenethylisoquinoline pathway from an endemic conifer *Cephalotaxus hainanensis*

Fei Qiao[a,1], Yuedong He[a,b,1], Yuhao Zhang[c,1], Xuefei Jiang[d,e], Hanqing Cong[a], Zhiming Wang[d,e], Huapeng Sun[a,2], Yibei Xiao[f,g,2], Yucheng Zhao[c,2], and Peter Nick[h]

Edited by Asaph Aharoni, Weizmann Institute of Science, Rehovot, Israel; received May 31, 2022; accepted November 15, 2022 by Editorial Board Member Natasha V. Raikhel

Cephalotaxines harbor great medical potential, but their natural source, the endemic conifer *Cephalotaxus* is highly endangered, creating a conflict between biotechnological valorization and preservation of biodiversity. Here, we construct the whole biosynthetic pathway to the 1-phenethylisoquinoline scaffold, as first committed compound for phenylethylisoquinoline alkaloids (PIAs), combining metabolic modeling, and transcriptome mining of *Cephalotaxus hainanensis* to infer the biosynthesis for PIA precursor. We identify a novel protein, *Ch*PSS, driving the Pictet–Spengler condensation and show that this enzyme represents the branching point where PIA biosynthesis diverges from the concurrent benzylisoquinoline-alkaloids pathway. We also pinpoint *Ch*DBR as crucial step to form 4-hydroxydihydrocinnamaldehyde diverging from lignin biosynthesis. The elucidation of the early PIA pathway represents an important step toward microbe-based production of these pharmaceutically important alkaloids resolving the conflict between biotechnology and preservation of biodiversity.

C. hainanensis | phenylethylisoquinoline alkaloids | 1-phenethylisoquinoline | *Ch*PSS

Significance

1-phenethylisoquinoline is the common scaffold of the phenylethylisoquinoline alkaloids (PIAs) that are relevant as anti-tumor compounds. The endemic conifer *Cephalotaxus hainanensis* produces the PIA homoharringtonine, highly potent against leukemia. The PIA scaffold is formed by a Pictet–Spengler condensation from dopamine and 4-HDCA, but the respective enzyme has remained elusive. Using a novel strategy, we identify this enzyme, construct the entire biosynthetic pathway leading to the precursors, and validate the implications of the identified enzymes by measuring their metabolic activities in vitro.

Phenylethylisoquinoline alkaloids (PIAs) are crucial secondary metabolites with high pharmacological potential, especially against cancer, inflammation, and cardio-cerebrovascular disease. Their medicinal impact is illustrated by well-known compounds, e.g., colchicine. The PIA homoharringtonine (HHT, *SI Appendix*, Fig. S1) has been approved by the US Food and Drug Administration for the therapy of chronic myeloid leukemia (1, 2). The great potential of PIAs is constrained by their sporadic occurrence to the conifer genus *Cephalotaxus* (3, 4), the monocot family *Colchicaceae*, and the dicot genus *Phelline* (5). For instance, HHT is uniquely found in *Cephalotaxus* which is endemic to East Asia (5). Although HHT has been generated by chemical synthesis (6), extraction from plants has remained the only commercial source of HHT. Since several species of this genus, for instance *Cephalotaxus hainanensis*, have been shifted to the verge of extinction, sustainable alternatives, safeguarding biodiversity, are urgently needed (3). However, semisynthetic strategies, such as metabolic engineering of microbes, are hampered by the fact that the biosynthesis mechanism of these PIAs has remained largely elusive.

A nearly complete reconstruction of colchicine biosynthesis has been achieved recently, which can be used as template for PIA biosynthesis in general, for all the PIAs share a common precursor, the 1-phenethylisoquinoline scaffold **15** (Fig. 1 and *SI Appendix*, Figs. S2 and S3) (6). Starting from this biogenetic intermediate **15**, nine newly identified enzymes can generate *N*-Formyldemecolcine, a precursor of colchicine (7). Moreover, a CYP71D12 protein, an α/β hydrolase protein, and a *N*-acetyltransferase superfamily protein were shown to be involved in colchicine biosynthesis (8). However, the pathway leading to this crucial precursor **15** has remained elusive, although a plausible upstream pathway for this PIA was proposed and reconstructed in *Nicotiana benthamiana* using nine enzymes from *Gloriosa superba*, *Coptis japonica,* and *Beta vulgaris* (7).

Isotope tracing experiments suggest that the PIA scaffold **15** derives from both, phenylalanine and tyrosine, which means that a part of the pathway is shared with that of benzylisoquinoline alkaloids (BIAs) (9–11). The two pathways diverge from the precursor for the PIA and BIA scaffold, respectively (Fig. 1 and *SI Appendix*, Fig. S4). In case of BIAs, (*S*)-norcoclaurine synthase (NCS) joins dopamine with 4-hydroxyphenylacetylaldehyde (4-HPAA) by a Pictet–Spengler condensation, while for PIAs, dopamine is fused with 4-HDCA. Considering the structural similarity of 4-HPAA and 4-HDCA, we assumed that their biosynthesis might be similar, and the fusion into the 1-phenethylisoquinoline scaffold should also be achieved by a similar enzymatic Pictet–Spengler condensation driven by an NCS-like protein. This notion is supported by the fact that PIAs can also be produced in *N. benthamiana* upon expression of a NCS from *C. japonica* (7). However, it has remained unclear, how phenylalanine is converted into 4-HDCA, and

Author contributions: F.Q., H.S., Y.X., and Y. Zhao designed research; F.Q., Y.H., Y. Zhang, Z.W., and Y. Zhao performed research; Y.H., Y. Zhang, X.J., H.C., Z.W., and Y. Zhao analyzed data; and F.Q., H.S., Y.X., Y. Zhao, and P.N. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. A.A. is a guest editor invited by the Editorial Board.

[1]F.Q., Y.H., and Y.Z. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: huapeng_sun@catas.cn, yibei.xiao@cpu.edu.cn, or zhaoyucheng1986@126.com.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2209339120/-/DCSupplemental.

Published December 28, 2022.

which enzyme is responsible for the condensation in plants capable of PIA biosynthesis.

Results and Discussion

A Novel Pr10 Enzyme was Screened Out and Identified. The attempt to find homologs of the *Cj*NCS in the transcriptomes of *G. superba* (7) or *C. hainanensis* (NCBI accession no. SRX12392777) did not lead to any candidate, indicating that, in those species, the Pictet–Spengler condensation might be driven by enzymes that derive from a convergent evolution. This assumption is supported by the fact that strictosidine synthase, that catalyzes the Pictet–Spengler condensation in the tetrahydroindole alkaloid pathway by joining tryptamine with secologanin, lacks homology with NCS (12, 13). Since the gymnosperm *C. hainanensis* and the dicot *C. japonica* have diverged almost 400 Mya (14), and this condensation step occurs sporadically, dispersed over several, unrelated taxa, homology is not to be expected, anyway. We, therefore, changed the strategy, focusing on crucial features of three-dimensional structure rather than on overall sequence homology in the first place. The substrate of the putative Pictet–Spengler enzyme in *C. hainanensis* (4-HDCA) differs from 4-HPAA, the substrate of NCS, only by a reduction of one carbon in the side chain of the phenolic ring. Since NCS belongs to the pathogenesis-related 10/Bet v1 proteins, we inferred that the unknown enzyme might be a member of the same protein family (15). In fact, we could identify 32 candidates from our recently constructed *C. hainanensis* transcriptome (*SI Appendix*, Figs. S5 and S6) that qualified as Pr10-like proteins, among those we found 22 members of the Pr10/Bet v1 family (*SI Appendix*, Table S1).

To filter out the most relevant candidate for the Pictet–Spengler reaction, we made use of a working model that had been developed for the NCS from *Thalictrum flavum* (a member of the Ranunculaceae) based on quantum chemical calculations, crystal structures inferred from X-ray diffraction, as well as in vitro assays using recombinant enzyme tailored by site-specific mutagenesis (10, 16, 17). These studies identified a glycine-rich loop, which is conserved in all identified NCSs, connecting the β2 and β3 sheets, and linked with enzymatic activity (18). Furthermore, a highly conserved motif in β-sheet 4 was predicted to be crucial for the activity (*SI Appendix*, Fig. S7) (19). Among the Pr10/Bet v1 candidates, only six were found to exhibit this glycine-rich loop motif (*SI Appendix*, Fig. S8). Thus, these six proteins represented the most likely candidates for the putative Pictet–Spengler enzyme responsible for the formation of the PIA backbone and were scrutinized further by heterologous expression in *Escherichia coli*. After feeding the substrates, dopamine and 4-HDCA, only one of these candidates converted the substrates into a product, which displayed a $[M+H]^+$ ion peak at a m/z 286.14319 (Fig. 2 *A* and *B*) diagnostic for the phenethylisoquinoline backbone. Although this enzyme shows only 17% identity with *Tf*NCS, it seems to mediate phenethylisoquinoline scaffold synthesis in *C. hainanensis* and was, therefore, named phenethylisoquinoline scaffold synthase (*Ch*PSS). The identification of intermediate **15** from the leaves of *C. hainanensis* supports a role for this intermediate in the biosynthesis of HHT (*SI Appendix*, Fig. S9). To further linked the recombinant enzyme activity with the actual activity in the plant itself, we extracted the crude protein of leaves of *C. hainanensis*, after adding the precursors **9** and **13**, and we detected the product **15** (*SI Appendix*, Fig. S10).

Exploring the Catalytic Mechanism of *Ch* PSS. As next step, we tried to get insight into the mechanism, by which *Ch*PSS is mediating the formation of the PIA scaffold. We were not successful in generating crystals from the recombinant protein, despite testing several thousands of conditions. Therefore, we modeled the 3D structure of *Ch*PSS using colabalphafold2 (*SI Appendix*, Fig. S11) (20, 21). The model obtained for *Ch*PSS predicted a strong overlap with that for *Tf*NCS, implying they may have a similar catalytic mechanism (11). Both proteins consist of three alpha helices and seven beta strands, enclosing a cavity. We decided, therefore, to use site-directed mutagenesis of potentially crucial amino acids to assess the effect on the in vitro activity of *Ch*PSS. A crucial tyrosine residue at position 108 in *Tf*NCS was conserved among all four known NCSs (*SI Appendix*, Fig. S7 and Fig. 2*C*) and had been shown to be essential for enzymatic function (15). While this tyrosine is missing in the corresponding position of *Ch*PSS β-sheet 4 (*SI Appendix*, Fig. S7 and Fig. 2*C*), we wondered whether a tyrosine in the neighborhood of the glycine-rich loop and also associated with a β-sheet might be functionally equivalent.

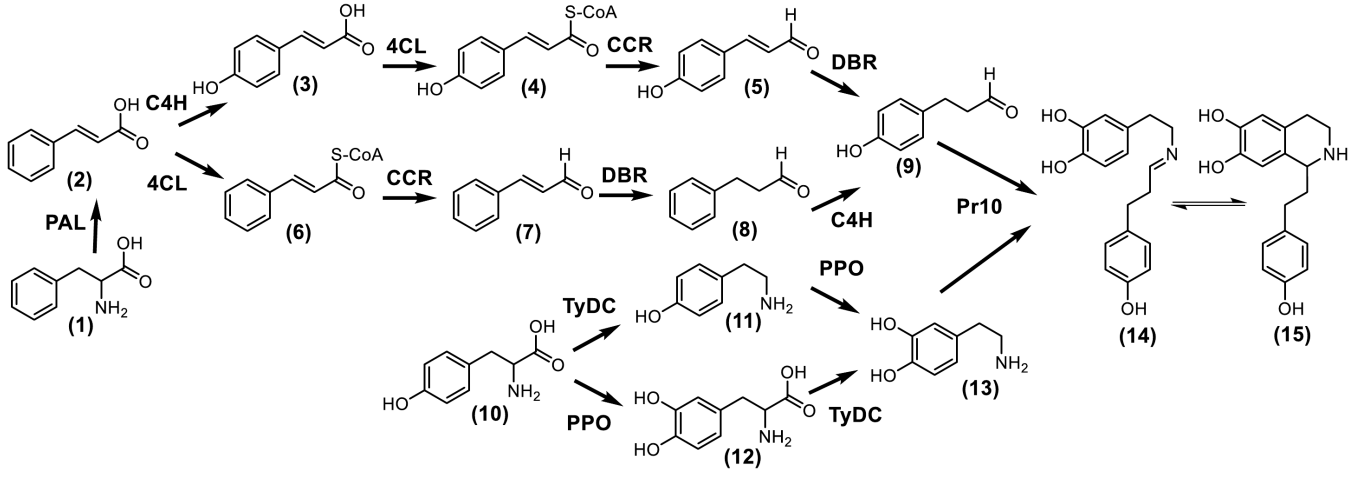

Fig. 1. Proposed upstream biosynthesis pathway of PIAs. Compounds: 1) L-phenylalanine; 2) cinnamic acid; 3) *p*-coumaric acid; 4) *p*-coumaroyl-CoA; 5) *p*-coumaroyl aldehyde; 6) cinnamoyl-CoA; 7) cinnamaldehyde; 8) phenylpropyl aldehyde; 9) 4-hydroxydihydrocinnamaldehyde (4-HDCA); 10) L-tyrosine; 11) tyramine; 12) L-DOPA; 13) dopamine; 14) 6,7-dihydroxy-1-(4-hydroxyphenylethyl)-1,2,3,4-tetrahydroisoquinoline intermediate; 15) 6,7-dihydroxy-1-(4-hydroxyphenylethyl)-1,2,3,4-tetrahydroisoquinoline (1-phenethylisoquinoline scaffold). Enzymes: PAL, phenylalanine ammonia-lyase; C4H, cinnamate 4-hydroxylase; 4CL, 4-coumarate CoA ligase; DBR, NADPH-dependent double-bond reductases; CCR, cinnamoyl-CoA reductase; TyDC/DODC, tyrosine/DOPA decarboxylase; PPO, polyphenoloxidase; Pr10, pathogenesis-related 10/Bet v1 proteins (*Ch*PSS).

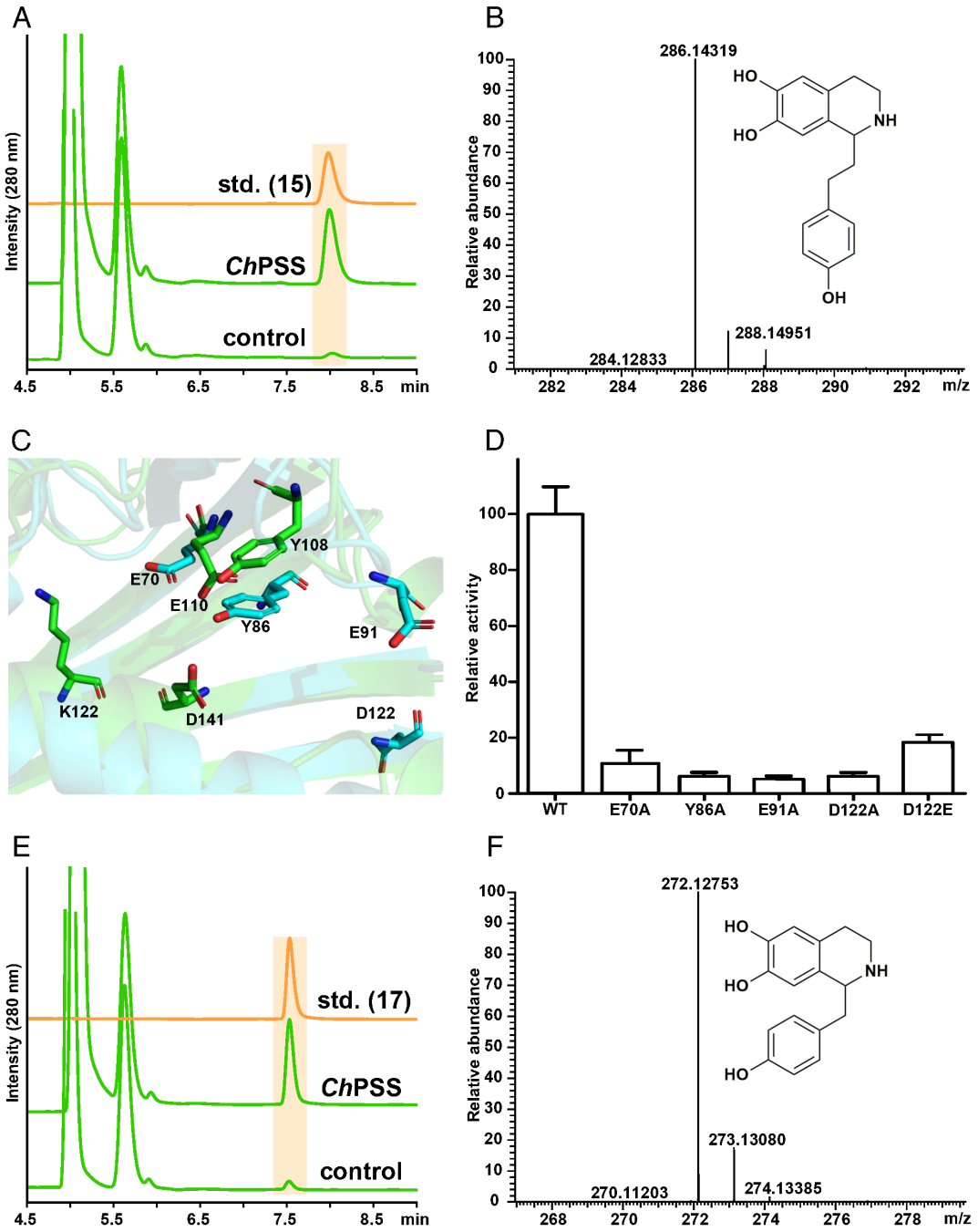

Fig. 2. Identification of *Ch*PSS. (*A*) The activity of *Ch*PSS was detected by feeding compounds **9** and **13** with 10 μg purified protein. A weak spontaneous reaction was detected in the control. (*B*) The reaction product of *Ch*PSS was confirmed by LC-MS (m/z 286.14319). (*C*) The superposed structures of *Ch*PSS (blue) with *Tf* NCS (green) and their key amino acid residues. (*D*) The relative activity of wild type and mutants of *Ch*PSS was detected by HPLC. The reaction system containing 5 mM dopamine, 2.5 mM 4-HDCA in 50 mM HEPES (5 mM ascorbate sodium, pH 7.0). The reaction product **15** was detected. The results were displayed as mean ± SD of three biological replicates, and the spontaneous activity was deducted from every measurement. (*E*) The activity of *Ch*PSS was detected by feeding compounds 4-HPAA (16) and **13**. A weak spontaneous reaction was also detected in the control. (*F*) The reaction product of *Ch*PSS was confirmed by LC-MS (m/z 272.127).

These criteria were met by Tyr86 of *Ch*PSS (Fig. 2*C*). When we mutated this tyrosine to alanine, the in vitro activity decreased significantly to 8% compared with the wild-type protein (Fig. 2*D* and *SI Appendix*, Figs. S12–S14). Likewise, a lysine (Lys122) plays a key role for the cyclization driven by *Tf* NCS. However, it was also missed in *Ch*PSS (Fig. 2*C*). In contrast, we identified the key amino acid residues corresponding to Glu110 and Asp141 of *Tf* NCS. When we mutated Glu70 in *Ch*PSS located in the above-mentioned conserved charge pattern in β-sheet 4 into alanine, a more substantial decrease of activity to 26% was achieved (Fig. 2*D* and *SI Appendix*, Figs. S12–S14) (13). Another Lewis acid residue, Asp122, also turned out to be crucial to maintain its activity, because mutant E91A nearly lost the activity, while mutating it to another Lewis acid glutamic acid (D122E), ~21% activity remained (Fig. 2*D* and *SI Appendix*, Figs. S12–S14). In summary, the implications from the model for *Ch*PSS based on the *Tf* NCS template were not totally met by the experimental results; however, some key residues with potential similar function with that of *Tf* NCS were identified, implying a similar catalytic process may be governed by these residues (*SI Appendix*, Fig. S15).

The observed disparity between the two enzymes has to be seen along with the lacking phylogenetic relationship between the two classes of enzymes (*SI Appendix*, Figs. S16 and S17). Thus, it seems that their function (Pictet–Spengler condensation) has been acquired by convergent evolution. While the low sequence identity does not rule out the possibility of a common ancestor for these proteins, a deeper insight into structural aspects of these quite divergent enzymes might help to pinpoint structural requirements of functionality that might act on the background of a poor of even missing overall similarity.

In the next step, we investigated the substrate specificity of recombinantly expressed *Ch*PSS. Interestingly, the enzyme also accepted 4-HPAA and dopamine to produce (*S*)-norcoclaurine reported by the diagnostic m/z peak at 272.12753 (Fig. 2 *E* and *F*). In addition, we were able to confirm the stereoselectivity of *Ch*PSS toward **15** by CD spectroscopy. These results are in good accordance with findings on *Cj*NCS, where a pure (*S*)-enantiomer was found (*SI Appendix*, Fig. S18). Thus, *Ch*PSS, in addition to its function in PIA biosynthesis, was able to act as NCS. While this broad substrate promiscuity was to be expected from the large catalytic cavity (Fig. 2*C* and *SI Appendix*, Fig. S11) (16, 22), this observation accentuates the question, how specificity is brought about. The fact that the enzymes from *G. superba* and *C. hainanensis* generate PIAs, while their counterparts from opium poppy and *C. japonica* produce BIAs, had remained elusive. When the catalytic cavities are permissive and not homologous, it might be substrate availability that delineates the biosynthesis of PIAs (deriving from dopamine and 4-HDCA) and BIAs (deriving from dopamine and 4-HPAA). Isotope trace analysis had shown that these substrates derive from phenylalanine and tyrosine, respectively (Fig. 1) (2, 5). However, so far, the details for 4-HPAA and 4-HDCA biosynthesis as well as for the origin of dopamine have remained unclear in both BIA and PIA accumulating plants (23).

Elucidation of the Biosynthetic Pathway for Dopamine. Decarboxylation of tyrosine by the aromatic amino acid decarboxylase (AAADC) and subsequent hydroxylation by the cytochrome P450 family protein CYP71AD6 seem to be crucial for the synthesis of BIAs, as well as for the production of betanin in *B. vulgaris* (23). For this reason, we searched for *Ch*AAADC candidates. We were able to identify one AAADC candidate (*Ch*TyDC1) and demonstrated that this enzyme was able to decarboxylate both tyrosine and L-DOPA as substrate (*SI Appendix*, Fig. S19). Our findings are consistent with the data from other plants, where AAADC was shown to accept both tyrosine (generating tyramine) and L-DOPA (generating dopamine) as well (*SI Appendix*, Fig. S20) (24). Since a CYP71AD family protein with a hydroxylase function, present in red beet (23), qualified as candidate for the subsequent hydroxylation, we searched the transcriptome of *C. hainanensis* for potential homologs but did not identify any member of the CYP71AD family. However, this kind of hydroxylase seems also to be absent from other plants that accumulate BIAs (8). Searching for alternatives, we wondered, whether a polyphenol oxidase (PPO) might bridge the gap, for it was reported that silencing of PPO in *Juglans regia* led to accumulation of tyramine, the direct precursor of dopamine (21). Since PPOs generally catalyze the oxidation of aromatic rings, which has also been proposed for colchicine biosynthesis (8), we tested the hypothesis that a PPO may exert a similar function in *C. hainanensis*. In fact, we identified two candidates from the *C. hainanensis* transcriptome, and one of them, *Ch*PPO1, was able to generate L-DOPA (12) after feeding tyrosine (Fig. 3*A*). The same protein was also able to form dopamine (13) after feeding tyramine (Fig. 3*B*). Thus, *Ch*PPO1 qualifies as the elusive aromatic hydroxylase (Fig. 3). Furthermore, when we co-expressed recombinantly, under the same bacterial promoter, *Ch*TyDC1 and *Ch*PPO1 (Fig. 3*C*), we achieved the complete biosynthesis of dopamine after feeding tyrosine (Fig. 3 *C* and *D*). Therefore, PPO might function as a hydroxylase in the biosynthesis of dopamine in our PIA-producing plant, which is in good accordance with the role suggested for this enzyme in colchicine biosynthesis (8).

Elucidation of the Biosynthetic Pathway for 4-HDCA. The next open issue was the formation of 4-HPAA and 4-HDCA. The two compounds differ mainly by a C-atom in the aliphatic chain of 4-HDCA (7). This structural difference might be crucial for channeling the metabolic flow between BIAs and PIAs (*SI Appendix*, Fig. S4), but it is not clear, how 4-HDCA is generated. To resolve this uncertainty in the pathway, we used information from colchicine biosynthesis and labeling experiments in cephalotaxine biosynthesis (4, 6) as template for two concurrent models of 4-HDCA from phenylalanine (Fig. 1 and *SI Appendix*, Fig. S21) that were also congruent with the early pathways proposed for *G. superba* and other PIA-producing plants (4, 7, 24). The difference between these concurrent schemes for 4-HDCA biosynthesis is the position of hydroxylation within this pathway. We, therefore, tested the possibility of a "hydroxylation-first" and a "hydroxylation-last" model. In the hydroxylation-first model, **2** is hydroxylated by C4H and then converted by 4CL and CCR to form **5**. Here, the sequence of reactions would be shared with the pathway leading to monolignols (25, 26), which would also be consistent with our in vitro enzymatic reaction (Fig. 4 *A* and *B* and *SI Appendix*, Figs. S22–S25). We found that *Ch*C4H1-3 behaved as to be expected from the hydroxylation-first model. More importantly, however, feeding compound **8** did not yield 4-HDCA for none of the four C4H members (Fig. 4*C*), falsifying a central implication predicted by the hydroxylation-last model. The reduction of the double bound in **5** to form **9** would represent the point where PIA biosynthesis diverges from lignin biosynthesis. For the PIA accumulator *G. superba*, an alkenal reductase has been described recently (7), and the search for double-bond reductases (DBR) in the transcriptome of *C. hainanensis* recovered five candidates. In fact, two of them, *Ch*DBR2 and *Ch*DBR3, were able to convert **5** to **9** in vitro (Fig. 4 *D* and *E*). In contrast, none of the five *Ch*DBRs was able to convert **7** to **8** (Fig. 4*F*), which imply the hydroxylation-last model is infeasible, even though that 4CL and CCR are able to accept **2** and **6** as substrates to form **7** (*SI Appendix*, Fig. S25).

In conclusion, using a combination of pathway modeling, mining the *Cephalotaxus* transcriptome, and experimental verification of recombinantly expressed candidates by precursor feeding in vitro, we could construct the pathway leading to the first committed compound in PIA biosynthesis. A novel member of the Pr10/Bet v1 family were identified as the key enzyme driving the Pictet–Spengler condensation to give rise to the 1-phenethylisoquinoline scaffold (Fig. 2). In addition, we could pinpoint DBR as crucial step, where the PIA precursor 4-HDCA diverges from the early lignin pathway. We proved that the hydroxylation of cinnamic acid (hydroxylation-first) as a necessary step to deliver the substrate for DBR to produce 4-HDCA (Fig. 4*D*). To what extent DBRs define, whether a given plant accumulates PIAs or their 4-HPAA derived counterparts, the BIAs, remains to be investigated. This work provides an important stepstone for the subsequent analysis and biotechnological application of cephalotaxine biosynthesis, which is expected to differ considerably from the alkaloid pathways that have already been constructed in other plants.

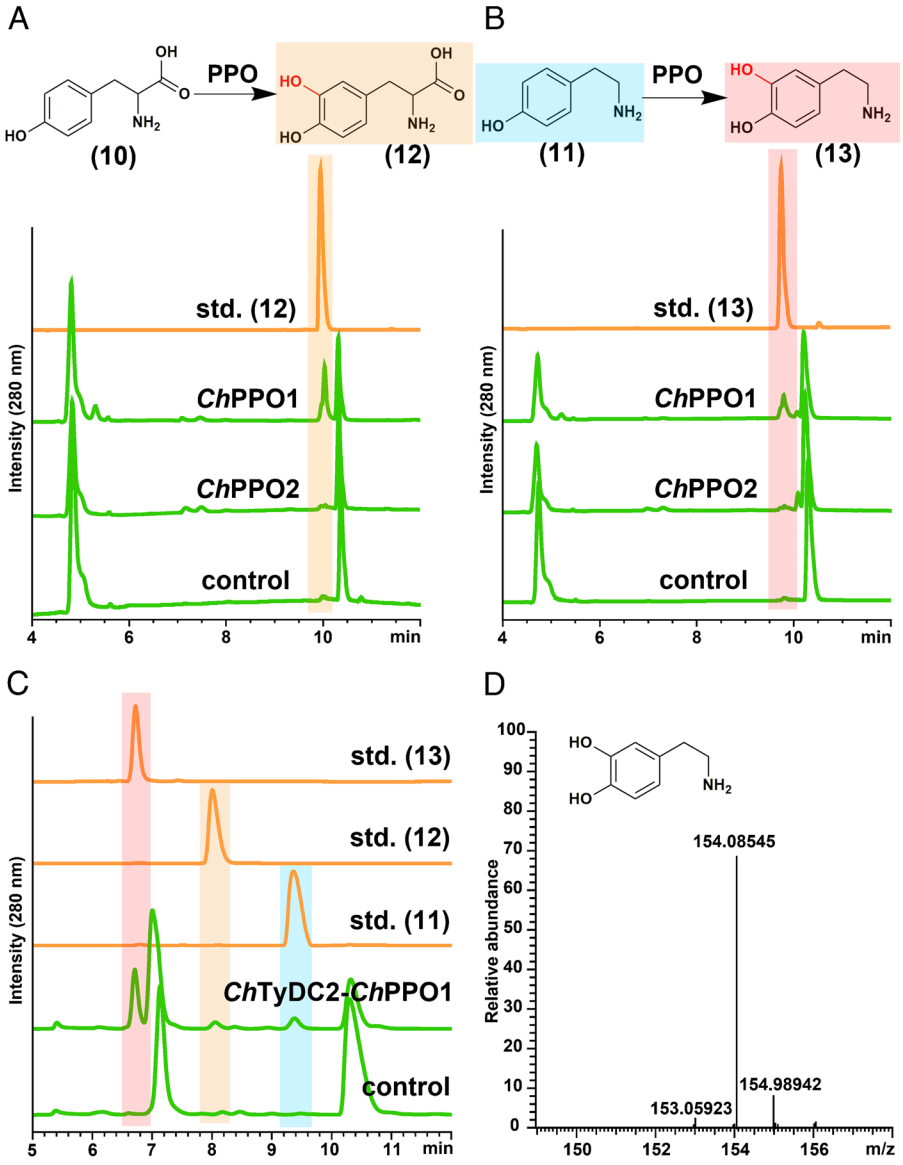

Fig. 3. Identification of *Ch*PPO. (*A*) The activity of *Ch*PPO was detected by feeding L-tyrosine **10** with 10 μg purified protein. (*B*) The activity of *Ch*PPO was detected by feeding tyramine **11**. Sample without PPO protein was used as control. (*C*) The activity of *Ch*PPO1 was further confirmed by using co-expression of *Ch*TyDC1 which was detected in HPLC. The intermediates **12** and **13** were detected. (*D*) The catalytic product of *Ch*TyDC1-PPO1, dopamine, was confirmed by LC-MS (m/z 154.08545).

Materials and Methods

Plant Materials and c-DNA Preparation. Plant materials of *C. hainanensis* were collected from Jianfengling in Ledong, Hainan, China (18°75′N, 108°85′E). Needle leaf, phloem, and root samples were harvested, then immediately frozen in liquid nitrogen and stored at −80 °C until use. Total RNA was extracted using the RNAprep Pure Kit (TIANGEN) according to the manufacturer's instructions. The quality and quantity of total RNA were measured by a NanoPhotometer NP50 (Implen). Then, PrimeScript™ RT Master Mix (Takara) was employed to perpetrate c-DNA template for gene clone and expression analysis.

Chemicals, Strains, and Enzymes. Compounds **1** to **13** were purchased from Sinoreagent (https://www.sinoreagent.com/), Bidepharm (https://www.bidepharm.com/), Aladdin (https://www.aladdin-e.com/), and Sigma-Aldrich (https://www.sigmaaldrich.cn/CN/zh). Compound **15** was synthesized by Wuxi Apptec (https://www.wuxiapptec.com/zh-cn), and the NMR and LC-MS data are listed in *SI Appendix*, Figs. S2 and S3. Compound **16** was synthesized according to the previously report (14). The compatible vectors pET28a, pYeDP60, and pCDFDuet-1 (Novagen) were used to express multiple genes. *E. coli* Top10 competent cells were used for plasmid amplification and isolation in the vectors construct process. *E. coli* BL21 (DE3) containing the corresponding expression vectors was used for all protein expression except *Ch*C4Hs were expressed by *Saccharomyces cerevisiae* WAT11. PrimeSTAR® Max DNA Polymerase (Takara) that used for PCR amplification, all other enzymes used for cloning were purchased from New England Biolabs (NEB). PCR products and Plasmid DNA were purified according to the procedure of E.Z.N.A.® Gel Extraction Kit and Plasmid mini Kit (Omega BIO-TEK).

SMRT Library Preparation, Sequencing, and Analysis. For Iso-Seq library construction, the c-DNA was prepared from five whole seedlings including leaves, roots, twigs, and barks by using of the SMARTer PCR c-DNA Synthesis Kit (Takara), and the full-length transcriptome was achieved by Iso-Seq method performed by Novagene Co., Ltd. Size fractionation and selection (non-fractionation and >4 kb) were performed using the BluePippin Size Selection System protocol as described by Pacific Biosciences (PN 100-092-800-03). Sequence data were processed using the SMRTlink 5.1 software, and circular consensus sequence (CCS) was generated from subread BAM files. A total of 18.17 Gb of clean data and 785,720,100 polymerase reads were obtained. A total of 962,965 CCS reads were generated, with a total of 16,598,265 subreads from 8 SMRT cells of non-normalized bins (0.5 to 2.5 kb, 2 to 3.5 kb, 3 to 6 kb, and 5 to 10 kb) and normalized bins (0.5 to 2.5 kb and 2 to 3.5 kb), including 581,486 (60%) full-length reads and 381,479 (40%)

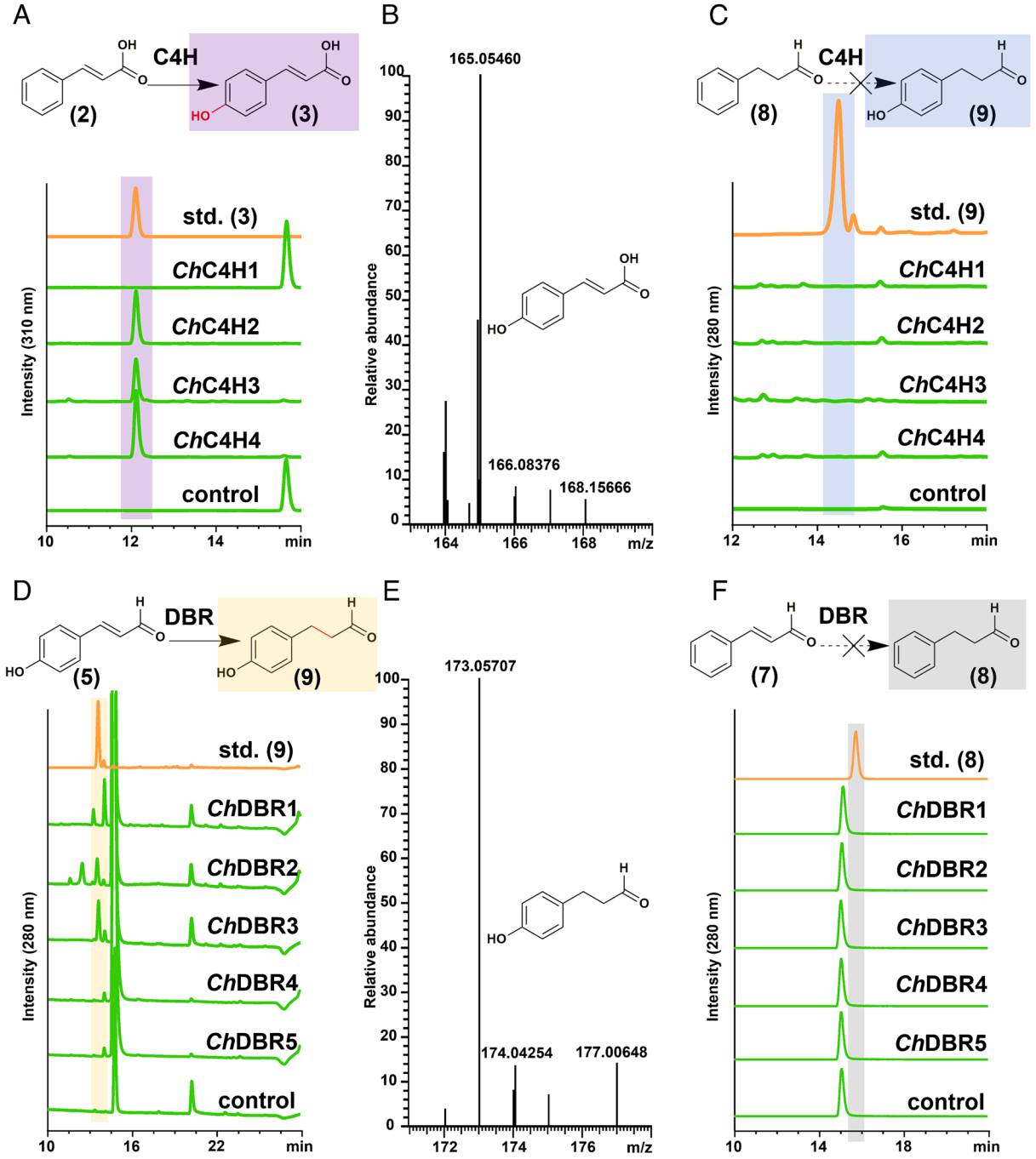

Fig. 4. Identification of *Ch*C4Hs and *Ch*DBRs. (*A*) The activity of *Ch*C4Hs was detected by HPLC. *Ch*C4H1-3 could catalyze **2** to form **3,** while *Ch*C4H4 has no function to **2**. (*B*) The catalytic product of **3** was further confirmed by LC-MS (m/z 165.05460). (*C*) The catalytic activity of *Ch*C4Hs was detected by HPLC after feeding compound **8**. These results indicated that the hydroxylation-last routine could not be realized. (*D*) The activity of *Ch*DBRs was detected by HPLC. *Ch*DBR2 and 3 could catalyze **5** to form **9,** while *Ch*DBR1, 4, and 5 have no function to **5**. (*E*) The catalytic product of **9** was further confirmed by LC-MS in $[M+Na]^+$ mode (m/z 173.05707). (*F*) The catalytic activity of *Ch*DBRs was detected by HPLC after feeding compound **7**. These results also indicated that the hydroxylation-last routine could not be realized.

non-full-length reads. The length of CCS reads ranged from 200 bp to 8,900 bp. Additional nucleotide errors in consensus reads were corrected using the Illumina RNA-seq data with the software LoRDEC. Any redundancy in corrected consensus reads was removed by CD-HIT to obtain final transcripts for the subsequent analysis (27, 28). Finally, a total of 282,151 transcripts were obtained. Gene function was annotated based on seven databases (NR, NT, Pfam, KOG/COG, Swiss-port, KO and GO, *SI Appendix*, Figs. S5 and S6) (29–34). BLAST (setting the e-value threshold to 10^{-10}), Diamond BLASTX (setting the e-value to threshold to 10^{-10}), and Hmmscan software were used in NT, NR/KOG/Swiss-Prot/KEGG, and Pfam database analysis, respectively.

Gene mining and Selection. Experiments with ^{14}C- and ^{3}H-labeled compounds show that both, tyrosine and phenylalanine, are the precursors of PIAs (4), and this also had been confirmed by recent work (7). However, the elusive 3′-hydroxylase on **10** or **11** renders the exact order of the upstream PIA biosynthetic pathway unclear. Even though *Bv*CYP76AD5 and *Cj*NCS had been suggested to complement the pathway, neither the true homologues, nor the order of the reactions is understood. Hence, we used a model for the upstream PIA pathway to focus on the hydroxylase and NCS-like protein (Fig. 1). Based on this metabolic-flux analysis, we can propose that eight enzymes should be sufficient to catalyze this part, which allowed to select candidate genes from our functionally annotated full-length

transcriptome dataset. As a result, in total 309 transcripts were selected, including 120 transcripts for PAL genes, 27 for C4H genes, 15 for 4CL genes, 62 for CCR genes, 28 for DBR genes, 20 for TyDC genes, five for PPO genes, and 32 for Pr10/Bet v1 genes. Because the SMRT sequencing platform allows for obtaining the full-length CDS with complete ORFs, we finally screened out 30 genes sequences with complete ORFs after removing repetitive and redundant transcripts (*SI Appendix*, Tables S1 and S2). The sequences used in this study were listed in *SI Appendix*, Table S6.

Cloning of Candidate Genes. All candidate genes were amplified from a c-DNA pooled from leaf, phloem and root of *C. hainanensis*. The purified amplicons were then inserted into the corresponding expression vectors using in-fusion cloning (Clontech). The genes of C4H were inserted into pYeDP60 vector between *BamH I* and *EcoR I* sites. All the other genes were inserted into pET28a vector between *Nde I* and *EcoR I* restriction sites, except for *Ch*PSS that was inserted between *Nco I* and *Xho I* sites. To obtain a co-expression construct combining TyDC and PPO, the DNA fragment containing the PPO1 coding sequence, T7 promoter region, and RBS was amplified with primers PPO-F-SacI and PPO-R-XhoI from pET28a-*Ch*PPO1 and inserted into the linearized pET28a-*Ch*TyDC2 to generate pET28a-TyDC2+PPO1. For co-expression of 4CL and CCR, *Ch*4CL2 and *Ch*CCR1 coding sequences were obtained using primers 4CL2-F-NcoI/4CL2-R-BamHI and CCR1-F-NdeI/CCR1-R-EcoRV, respectively, and cloned into pCDFDuet-1 between the *NcoI* / *BamH I* and *Nde I* / *EcoR V* sites, giving rise to the vector pCDFDuet-4CL2+CCR1. All primers used in this study are listed in *SI Appendix*, Table S3.

Heterologous Expression of C4H in Yeast. The reconstructed plasmid pYeDP60-C4H was transformed into the *S. cerevisiae* strain WAT11, and positive transformants were screened on solid plates of SC-U (SC dropout medium without uracil) containing 20 g/L glucose. Positive clones were incubated under shaking at 30 °C until the OD_{600} reached to 2. The cells were then spun down and the sediment washed several times with ddH_2O to remove residual glucose. Subsequently, the sediment was resuspended in SC-U medium containing 20 g/L galactose to induce expression of the target protein. As substrates, cinnamic acid or phenylpropyl aldehyde was added to a final concentration of 100 μM to the resuspended cells and then incubated at 28 °C for 12 h. After collection of the cells, they were resuspended in 2 mL methanol and then ultrasonicated twice for 30 min, collecting the supernatant for HPLC-MS analysis.

Heterologous Expression and Purification of Candidate Proteins in *E. coli*. All the inserts were transformed into *E. coli* BL21 (DE3) by the heat shock method and then selected on LB plates containing 50 mg/L kanamycin. After sequencing, positive single colonies were inoculated separately into LB medium containing 50 mg/L kanamycin for seed culture. Then, the overnight culture was inoculated into an appropriate volume of LB medium containing 50 mg/L kanamycin and grown at 37 °C until a OD_{600} of 0.6, before inducing with 0.5 mM isopropyl-β-D-thiogalactoside (IPTG) at 25 °C and 180 rpm. After incubation for additional 16 to 20 h, the culture was spun down at 5,000 rpm for 10 min at 4 °C. Then, the pellet was resuspended in lysis buffer containing 20 mM HEPES, 500 mM NaCl, 20 mM imidazole, and 10% (v/v) glycerol. The suspension was homogenized in by ultrasonication, and then the homogenate was collected at 25,000 rpm at 4 °C. The supernatant was purified on a Ni-NTA column, and purity and molecular weight of the fusion proteins were verified by SDS-PAGE and nanodrop 2,000 ultramicro-spectrophotometer (*SI Appendix*, Fig. S12). The *Ch*PSS proteins were further purified on AKTA™ pure protein purification system by using Hitrap Q and Superdex 200 Increase 10/300 GL columns (*SI Appendix*, Fig. S12). The purified protein was concentrated and dialyzed against storage buffer (20 mM HEPES, 200 mM NaCl, and 20% (v/v) glycerol) and then stored at −80 °C till analysis.

Structure Prediction. The model prediction of *Ch*PSS was conducted using colabAlphaFold2 which is released on the GitHub by DeepMind. We also adopt auto dock vina to dispose the interaction between the *Ch*PSS and two ligands (dopamine, 4-HDCA). Finally, we adopt open sourced PyMOL to show the track and the binding mode of the *Ch*PSS and two ligands.

In Vitro Characterization and Screening of Recombinant Candidate Genes. If not specified otherwise, reaction conditions and detection procedures were as specified in *SI Appendix*, Tables S4 and S5, respectively. All the reactions were conducted in 30 °C for 30 min, while, for *Ch*PSS, temperature was raised to 45 °C. Reactions were stopped by transfer of the reaction mix on ice. The conditions had been adjusted based on extensive preparatory studies exploring different reaction and detection conditions, with special focus on detection of L-DOPA and dopamine by HPLC. For co-expression active assays of pET28a-TyDC2+PPO1 and pCDFDuet- 4CL2+CCR1, after 8 h of IPTG induction, the culture was centrifuged, the supernatant was discarded, cell pellet was re-suspended in 100 mM tris base buffer at pH 7.5. Then, the corresponding substrates were added to the mixture in 30 °C for about 8 h. Kinetic assays were performed under the same conditions as routine activity assays (10 μg protein and reaction at 45 °C for 30 min) and kinetic parameters determined by varying substrate concentrations while maintaining other reactants at saturation (same concentrations as in routine activity assays). Kinetic constants were calculated by nonlinear regression analysis (Origin 8; OriginLab Corp).

For products identification, the samples were diluted by 500 μL acetonitrile and analyzed by UPLC-MS/MS after filtering through a 0.22 μm filter membrane. UPLC was performed at a flow rate of 0.3 mL/min in solvent A (98% [v/v] acetonitrile, 2% [v/v] water, and 0.2% formic acid) and solvent B (2% [v/v] acetonitrile, 98% [v/v] water, and 0.2% formic acid) using an InfinityLab Poroshell 120 EC-C18 column (2.1 × 100 mm, 2.7 μm particle size; Agilent Technologies). The gradient elution program is as follows: 5% solvent A, 0 to 1.5 min; 50% solvent A, 1.5 to 2 min; 95% solvent A, 2 to 2.5 min; 100% solvent A, 2.5 to 3 min; 95% solvent A, 3 to 4 min; 50% solvent A, 4 to 5 min; 95% solvent A, 5 to 6 min.

For collecting of all MS data, the Exactive™ Plus Orbitrap Mass Spectrometer (ThermoFisher Scientific™) equipped with an electrospray ionization (ESI) probe inlet was used. The signal was achieved in positive ion mode with ESI. Ions were generated and focused using an ESI voltage of 4.0 kV; sheath gas (nitrogen) flow rate, 40 arb; aux/sweep gas (nitrogen) flow rate, 15 arb; capillary temperature, 350 °C.

Data, Materials, and Software Availability. Full-length RNA Sequencing data have been deposited into the National Center for Biotechnology Information, NIH, Sequence Read Archive (Accession number: SRX12392777) (35). All data are available in the main text or the *SI Appendix*.

ACKNOWLEDGMENTS. We thank Prof. Daizhu Lv and Deqing Zhao for technical assistance with the UPLC-MS/MS analysis. This work was funded by National Science Foundation of China (3207036431570326), the Hainan Provincial Science Foundation (2019RC309), and the Central Public-Interest Scientific Institution Basal Research Fund of the Chinese Academy for Tropical Agricultural Sciences (1630032019038).

Author affiliations: [a]Key Laboratory of Crop Gene Resources and Germplasm Enhancement in Southern China, ministry of Agriculture; Tropical Crops Genetic Resources Institute, Chinese Academy of Tropical Agricultural Sciences, Haikou 571101, China; [b]College of Horticulture, Hunan Agricultural University, Changsha 410128, China; [c]Department of Resources Science of Traditional Chinese Medicines, School of Traditional Chinese Pharmacy, China Pharmaceutical University, Nanjing 210009, China; [d]Hainan Key Laboratory of Sustainable Utilization of Tropical Bioresources, College of Horticulture, Hainan University, Haikou 570228, China; [e]Key Laboratory for Quality Regulation of Tropical Horticultural Plants of Hainan Province, College of Horticulture, Hainan University, Haikou 570228, China; [f]Department of Pharmacology, School of Pharmacy, China Pharmaceutical University, Nanjing 210009, China; [g]State Key Laboratory of Natural Medicines, China Pharmaceutical University, Nanjing 210009, China; and [h]Molecular Cell Biology, Botanical Institute, Karlsruhe Institute of Technology, Karlsruhe D-76131, Germany

1. T. Jin *et al.*, Homoharringtonine-based induction regimens for patients with *de-novo* acute myeloid leukaemia: A multicentre, open-label, randomised, controlled phase 3 trial. *Lancet Oncol.* **14**, 599–608 (2013).
2. H. M. Kantarjian, S. O'Brien, J. Cortes, Homoharringtonine/omacetaxine mepesuccinate: The long and winding road to food and drug administration approval. *Clin. Lymphoma Myeloma Leuk.* **13**, 530–533 (2013).
3. Y. Yang, D. Luscombe, T. Katsuki, *Cephalotaxus harringtonii*. The IUCN Red List of Threatened Species 2013: e.T39589A2929537. https://dx.doi.org/10.2305/IUCN.UK.2013-1.RLTS.T39589A2929537.en. Accessed 11 September 2022.
4. Y. Yang,W. Liao,*Cephalotaxus hainanensis*. The IUCN Red List of Threatened Species 2013:e.T34065A2842288. https://dx.doi.org/10.2305/IUCN.UK.2013-1.RLTS.T34065A2842288.en. Accessed 11 September 2022.

5. T. Kametani, M. Koizumi, "Phenethylisoquinoline alkaloids" in *The Alkaloids: Chemistry and Physiology*, R.H.F. Manske, Eds. (Academic Press, 1973), pp. 265–323.
6. S. Hiranuma, T. Hudlicky, Synthesis of homoharringtonine and its derivative by partial esterification of cephalotaxine. *Tetrahedron Lett.* **23**, 3431–3434 (1982).
7. R. S. Nett, W. Lau, E. S. Sattely, Discovery and engineering of colchicine alkaloid biosynthesis. *Nature* **584**, 148–153 (2020).
8. R. S. Nett, E. S. Sattely, Total biosynthesis of the tubulin-binding alkaloid colchicine. *J. Am. Chem. Soc.* **143**, 19454–19465 (2021).
9. H. Abdelkafi, B. Nay, Natural products from Cephalotaxus sp.: Chemical diversity and synthetic aspects. *Nat. Prod. Rep.* **29**, 845–869 (2012).
10. D. Xu *et al.*, Integration of full-length transcriptomics and targeted metabolomics to identify benzylisoquinoline alkaloid biosynthetic genes in *Corydalis yanhusuo*. *Hortic. Res.* **8**, 16 (2021).
11. S. Xiang, F. Himo, Enzymatic Pictet-Spengler reaction: Computational study of the mechanism and enantioselectivity of norcoclaurine synthase. *J. Am. Chem. Soc.* **141**, 11230–11238 (2019).
12. E. Eger *et al.*, Inverted binding of non-natural substrates in strictosidine synthase leads to a switch of stereochemical outcome in enzyme-catalyzed Pictet-Spengler reactions. *J. Am. Chem. Soc.* **142**, 792–800 (2020).
13. S. Xiang, F. Himo, Computational study of Pictet-Spenglerase strictosidine synthase: Reaction mechanism and origins of enantioselectivity of natural and non-natural substrates. *ACS Catal.* **10**, 13630–13640 (2020).
14. A. Zimmer *et al.*, Dating the early evolution of plants: Detection and molecular clock analyses of orthologs. *Mol. Genet. Genomics* **278**, 393–402 (2007).
15. E. J. Lee, P. Facchini, Norcoclaurine synthase is a member of the pathogenesis-related 10/Bet v1 protein family. *Plant Cell* **22**, 3489–3503 (2010).
16. A. Ilari *et al.*, Structural basis of enzymatic (*S*)-norcoclaurine biosynthesis. *J. Biol. Chem.* **284**, 897–904 (2009).
17. A. Bonamore *et al.*, An enzymatic, stereoselective synthesis of (*S*)-norcoclaurine. *Green Chem.* **12**, 1623–1627 (2010).
18. J. S. Morris, K. M. P. Caldo, S. Liang, P. J. Facchini, PR10/Bet v1-like proteins as novel contributors to plant biochemical diversity. *Chembiochem.* **22**, 264–287 (2021).
19. B. R. Lichman *et al.*, Structural evidence for the dopamine-first mechanism of norcoclaurine synthase. *Biochemistry* **56**, 5274–5277 (2017).
20. E. Callaway, DeepMind's AI for protein structure is coming to the masses. *Nature*, 10.1038/d41586-021-01968-y (2021).
21. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
22. L. Y. P. Luk, S. Bunn, D. K. Liscombe, P. J. Facchini, M. E. Tanner, Mechanistic studies on norcoclaurine synthase of benzylisoquinoline alkaloid biosynthesis: An enzymatic Pictet-Spengler reaction. *Biochemistry* **46**, 10153–10161 (2007).
23. G. Polturak, A. Aharoni, "La Vie en Rose": Biosynthesis, sources, and applications of betalain pigments. *Mol. Plant* **11**, 7–22 (2018).
24. M. P. Torrens-Spence *et al.*, Structural basis for divergent and convergent evolution of catalytic machineries in plant aromatic amino acid decarboxylase proteins. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 10806–10817 (2020).
25. N. D. Bonawitz, C. Chapple, The genetics of lignin biosynthesis: Connecting genotype to phenotype. *Annu. Rev. Genet.* **44**, 337–363 (2010).
26. Y. He *et al.*, Characterisation, expression and functional analysis of PAL gene family in *Cephalotaxus hainanensis*. *Plant Physiol. Biochem.* **156**, 461–470 (2020).
27. L. Salmela, E. Rivals, LoRDEC: Accurate and efficient long read error correction. *Bioinformatics* **30**, 3506–3514 (2014).
28. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
29. M. Ashburner *et al.*, Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
30. R. L. Tatusov *et al.*, The COG database: An updated version includes eukaryotes. *MBC Bioinform.* **4**, 41 (2003).
31. M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, M. Hattori, The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, 277–280 (2004).
32. M. Kanehisa *et al.*, From genomics to chemical genomics: New developments in KEGG. *Nucleic Acids Res.* **34**, 354–357 (2006).
33. A. Bairoch, R. Apweiler, The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
34. R. D. Finn *et al.*, The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.* **44**, 279–285 (2016).
35. Chinese Academy of Tropical Agricultural Sciences, RNA-Seq of C. hainanensis. National Center for Biotechnology Information, *NIH, Sequence Read Archive*. https://www.ncbi.nlm.nih.gov/sra/SRX12392777. Deposited 8 August 2022.