nature communications

Article

https://doi.org/10.1038/s41467-024-51285-x

The gradual establishment of complex coumarin biosynthetic pathway in Apiaceae

Received: 15 January 2024

Accepted: 5 August 2024

Published online: 10 August 2024

Check for updates

Xin-Cheng Huang[1,8], Huanying Tang[2,8], Xuefen Wei[1,8], Yuedong He[3], Shuaiya Hu[1], Jia-Yi Wu[1], Dingqiao Xu[4], Fei Qiao[5,6] ✉, Jia-Yu Xue[1] ✉ & Yucheng Zhao[2,7] ✉

Complex coumarins (CCs) represent characteristic metabolites found in Apiaceae plants, possessing significant medical value. Their essential functional role is likely as protectants against pathogens and regulators responding to environmental stimuli. Utilizing genomes and transcriptomes from 34 Apiaceae plants, including our recently sequenced *Peucedanum praeruptorum*, we conduct comprehensive phylogenetic analyses to reconstruct the detailed evolutionary process of the CC biosynthetic pathway in Apiaceae. Our results show that three key enzymes – *p*-coumaroyl CoA 2′-hydroxylase (C2′H), C-prenyltransferase (C-PT), and cyclase – originated successively at different evolutionary nodes within Apiaceae through various means of gene duplications: ectopic and tandem duplications. Neofunctionalization endows these enzymes with novel functions necessary for CC biosynthesis, thus completing the pathway. Candidate genes are cloned for heterologous expression and subjected to in vitro enzymatic assays to test our hypothesis regarding the origins of the key enzymes, and the results precisely validate our evolutionary inferences. Among the three enzymes, C-PTs are likely the primary determinant of the structural diversity of CCs (linear/angular), due to divergent activities evolved to target different positions (C-6 or C-8) of umbelliferone. A key amino acid variation (Ala161/Thr161) is identified and proven to play a crucial role in the alteration of enzymatic activity, possibly resulting in distinct binding forms between enzymes and substrates, thereby leading to different products. In conclusion, this study provides a detailed trajectory for the establishment and evolution of the CC biosynthetic pathway in Apiaceae. It explains why only a portion, not all, of Apiaceae plants can produce CCs and reveals the mechanisms of CC structural diversity among different Apiaceae plants.

In response to the arms race with aggressors, plants have evolved the ability to produce secondary metabolites. Some secondary metabolites exhibit specific therapeutic activities in humans, and were skillfully employed in treating human diseases by our ancestors who developed Ayurvedic medicine, Arabian medicine, traditional Chinese medicine, and other ancient healing practices. Coumarins represent one of the major categories of metabolites in plants. In addition to serving as protectants against phytopathogens and responding to biotic and abiotic pressures[1–3], coumarins also exhibit diverse medical bioactivities such as antibacterial, anti-tumor, anti-oxidation, anti-coagulation and anti-inflammatory properties[1,4].

These diversified biological activities potentially stem from the structural diversity of coumarins, which can be classified into five types: simple coumarins (SCs), linear furanocoumarins (FCs), angular

A full list of affiliations appears at the end of the paper. ✉e-mail: fei.qiao@catas.cn; xuejy@njau.edu.cn; zhaoyucheng1986@126.com

https://doi.org/10.1038/s41467-024-51285-x

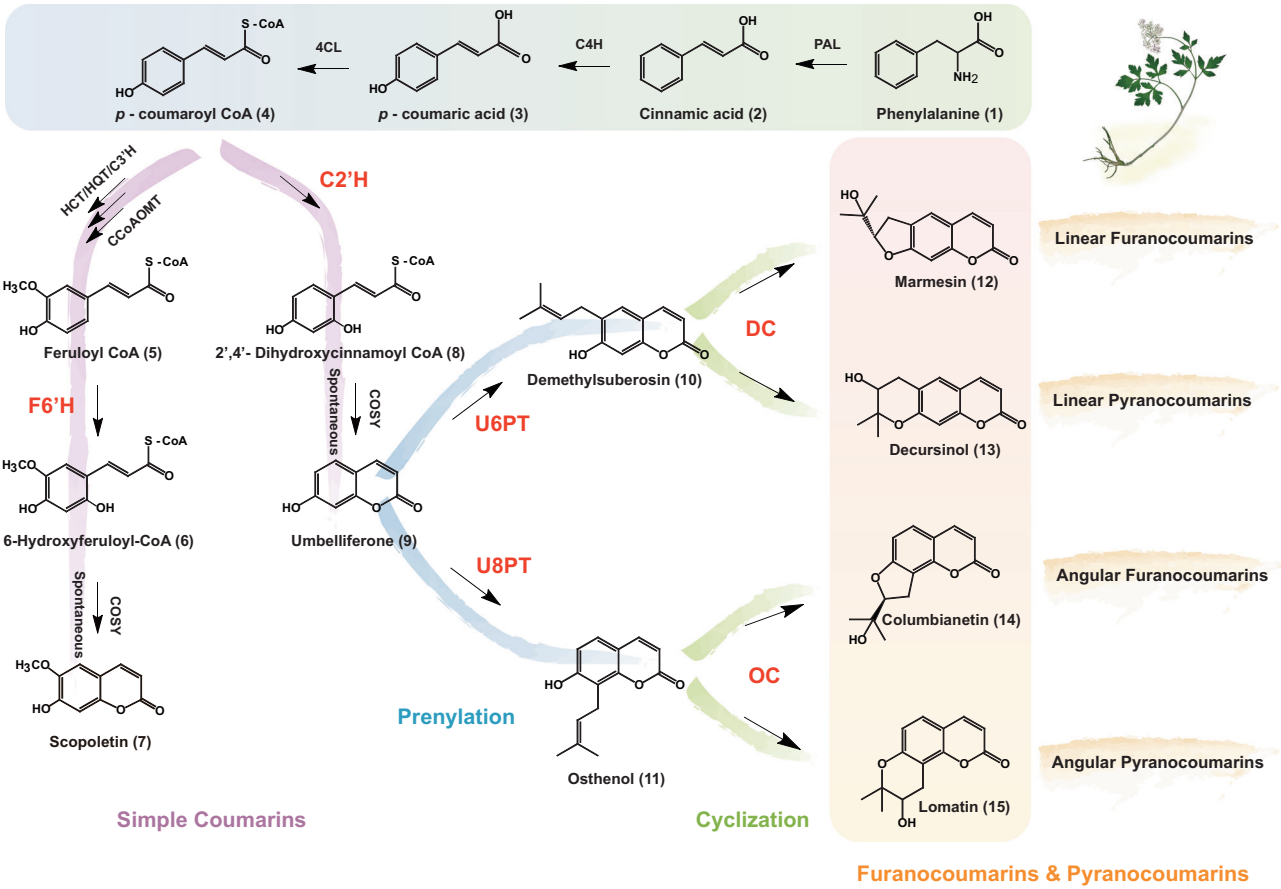

Fig. 1 | The proposed biosynthetic pathway of CC in Apiaceae. PAL phenylalanine ammonia lyase, C4H cinnamate 4-hydroxylase, 4CL 4-coumarate: coenzyme A ligase, HCT/HQT hydroxycinnamoyl CoA shikimate/quinate hydroxycinnamoyl transferase, CCoAOMT caffeoyl-CoA O-methyltransferase, C3'H cinnamoyl ester 3'-hydroxylase, C2'H *p*-coumaroyl CoA 2' -hydroxylase, F6'H feruloyl-CoA 6'-hydroxylase, COSY coumarin synthase, U6PT umbelliferone 6-prenyltransferase, U8PT umbelliferone 8-prenyltransferase, DC demelthyisuberosin cyclase, OC osthenol cyclase. The plant in the picture is the newly sequenced *P. praeruptorum* rich in all coumarin types and particularly high angular CC content.

FCs, linear pyranocoumarins (PCs) and angular PCs; and the later four can be collectively referred to as complex coumarins (CCs). Comparatively, CCs are the primary ingredients with potential medical values, and are the further biosynthesized products of the SCs. The biosynthetic process involves three distinct steps via three types of enzymes – *p*-coumaroyl CoA 2'-hydroxylase (C2'H)[5,6], C-prenyltransferase (C-PT)[7–10] and cyclases[7,11]. The detailed CC biosynthetic process (Fig. 1) initiates with C2'H catalyzing *p*-coumaroyl CoA to form umbelliferone[5,6]. At this step, C2'H competes the substrate with enzymes catalyzing the biosynthesis of scopoletin (F6'H), which will not continue to process to CCs. Due to the strict substrate-specificity, only umbelliferone can be recognized by downstream C-PTs, and continues to form demethylsuberosin or osthenol by prenylation, while O-prenyltransferases (O-PTs) lead reactions to 6',7'-epoxyauraptene or 6',7'-dihydroxybergamttin instead, without continuing to process to CCs[7,12]. At last, cyclases catalyze the cyclization of pyran or furan rings to form PCs or FCs, which were characterized more recently, completing the final missing gap in the CC biosynthetic pathway[7,11].

In nature, although SCs are ubiquitous in angiosperms, the main active ingredients, CCs are reported to primarily accumulate in four out of all 433 angiosperm families, Apiaceae, Moraceae, Rutaceae, and Fabaceae[1,13], which are phylogenetically distant to each other. This dispersed distribution of CCs across angiosperm phylogeny implies multiple and independent origins of CC biosynthesis in these four families. Coincidently, molecular evidence also supports independent evolution of CC biosynthesis in different angiosperm lineages. For instance, a phylogenetic analysis of PTs proposed that Apiaceae, Moraceae, Rutaceae PTs are derived from distinct ancestors through convergent evolution[8]. The enzymes responsible for cyclization also belong to different CYP450 families: CYP76F in Moraceae and CYP736A in Apiaceae, respectively[7,11].

Despite Apiaceae being a well-known angiosperm family for CC metabolites, the truth is that not all Apiaceae plants can produce CCs. Some famous traditional Chinese medicinal herbs, such as *Peucedanum praeruptorum*, *Angelica sinensis*, *A. dahurica*, *Saposhnikovia divaricata* accumulate CCs in roots. In contrast, other plants, for example, most vegetables and spices – *Daucus carota* (carrot), *Apium graveolens* (celery), *Oenanthe sinensis* (water dropwort) and *Coriandrum sativum* (coriander) – do not accumulate CCs. Additionally, some other famous medicinal plants like *Centella asiatica* and *Bupleurum chinense*, do not accumulate CCs either[7,14–18]. These observations imply that the CC biosynthesis likely underwent a variable mode of evolution in Apiaceae. Now, Apiaceae has amassed a rich genomic and transcriptomic data source due to the rapid advancement of sequencing technology[19–23]. We also newly sequenced *P. praeruptorum*, an Apiaceae plant rich in all coumarin types and particularly high angular CC content[24,25]. These data serve as valuable foundational information to elucidate the molecular mechanisms underlying the origin of CC biosynthesis and the diversified products. Using an evolutionary genomics research strategy, our work strived to reconstruct a detailed process of the establishment and evolution of Apiaceae CC biosynthetic pathway.

Results

Phylogeny of Apiaceae species provides a framework to study the evolution of CC biosynthesis

To provide a stable framework of species relationship for our analysis into the Apiaceae CC biosynthesis, we first reconstructed a highly resolved phylogeny of Apiaceae, utilizing all available genomes and transcriptomes in this family and 18 plants from other taxa as the outgroup (Supplementary Table 1). Among the taxa used, our newly sequenced *P. praeruptorum* using combined Hifi and HiC data (1.74 Gb in size, scaffold N50 = 157.14 Mb and 33,420 protein-coding genes annotated, our annotation captures 94.9% of the embryophyta BUSCO (odb10) genes, with 82.6% in single copy and 12.3% in duplicates, see Supplementary Figs. 1–3, Supplementary Tables 2–9, Supplementary Notes for more information) stands as a representative that highly accumulates diverse CC products.

According to orthologous gene families classified by OrthoFinder (v2.5.5)[26], we extracted a total of 1708 low copy orthologous families (OGs) and further generated a refined single-copy genes (RSCGs) using the criterion described in the Methods section. The RSCGs were then processed, and concatenated datasets of amino acids and corresponding coding sequences were constructed. Both nucleotide and amino acid datasets recovered a congruent topology for Apiaceae phylogeny with robust support (Fig. 2, Supplementary Fig. 4). In Apiaceae, *P. praeruptorum* is resolved as the sister to *S. divaricata*, and *Cnidium monnieri*, *A. dahurica* and *C. sativum* were recovered to be successively sisters, all belonging to Apioideae. All relationships received 100% bootstrap support except the sister relationship of *A. graveolens* and *Pastinaca sativa*, which received 98% support from the amino acid dataset. Therefore, a further analysis using the coalescent datasets was conducted, and obtained congruent results supporting the sister relationship, whereas incomplete lineage sorting was detected (Fig. 2, Supplementary Fig. 5), likely influencing the bootstrap value. Apioideae was the densest sampled among all subfamilies of Apiaceae, due to the fact that reported Apiaceae plants with CC metabolites mainly belong to this clade. All 13 Apioideae plants clustered as a monophyletic group, with *Sanicula orthacantha*, *Azorella atacamensis* and *C. asiatica* being successive sister groups, representing Saniculoideae, Azorelloideae and Mackinlayoideae, respectively. Molecular clock estimated the emergence of the crown group of Apiaceae to be around 49.4 million years ago (MYA), whereas the divergence of Apiaceae from other Apiales lineages could date back to 53.2 MYA. The divergence time for *P. praeruptorum* and its closest relative *S. divaricata* is around 5.5 MYA (Supplementary Fig. 6, Supplementary Table 10), which may represent the divergence time of two genera, *Peucedanum* and *Saposhnikovia*.

By mapping our collected information of the coumarin accumulation in Apiaceae species to the phylogenetic backbone, it's notable that Apiaceae plants of CC accumulation are limited in a number of tribes/genera gathering in Apoideae, instead of being evenly scattered throughout Apiaceae (Fig. 2). Therefore, the complete CC biosynthetic pathway is likely established first in later Apoideae evolution, which would explain the lack of CCs in other subfamilies of Apiaceae. Meanwhile, different plants incline to accumulate different CCs, and most Apoideae plants cannot produce all CCs except *P. praeruptorum*,

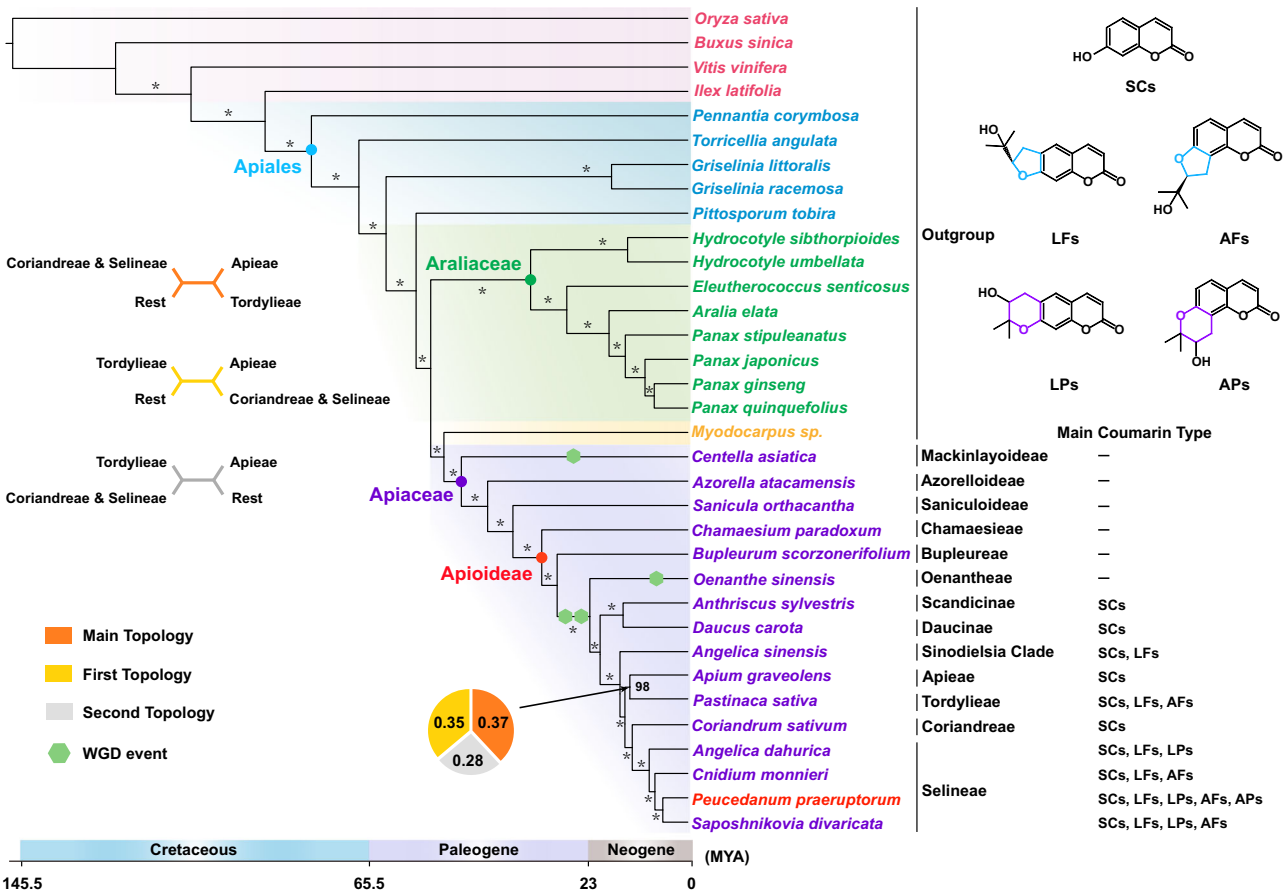

Fig. 2 | Phylogeny and the coumarins distribution of Apiaceae plants. Phylogenetic analysis and divergence time estimation were based on 1708 refined low copy orthologous gene groups from 34 angiosperms. The coumarin type and distribution information were collected from[14,66–85]. '-' no report, SCs simple coumarins, LFs linear furanocoumarins, LPs linear pyranocoumarins, AFs angular furanocoumarins, APs angular pyranocoumarins, WGD whole-genome duplication. Please see Supplementary Figs. 26–28 for the relevant analysis. Asterisks represent 100% support. Source data are provided as a Source Data file.

https://doi.org/10.1038/s41467-024-51285-x

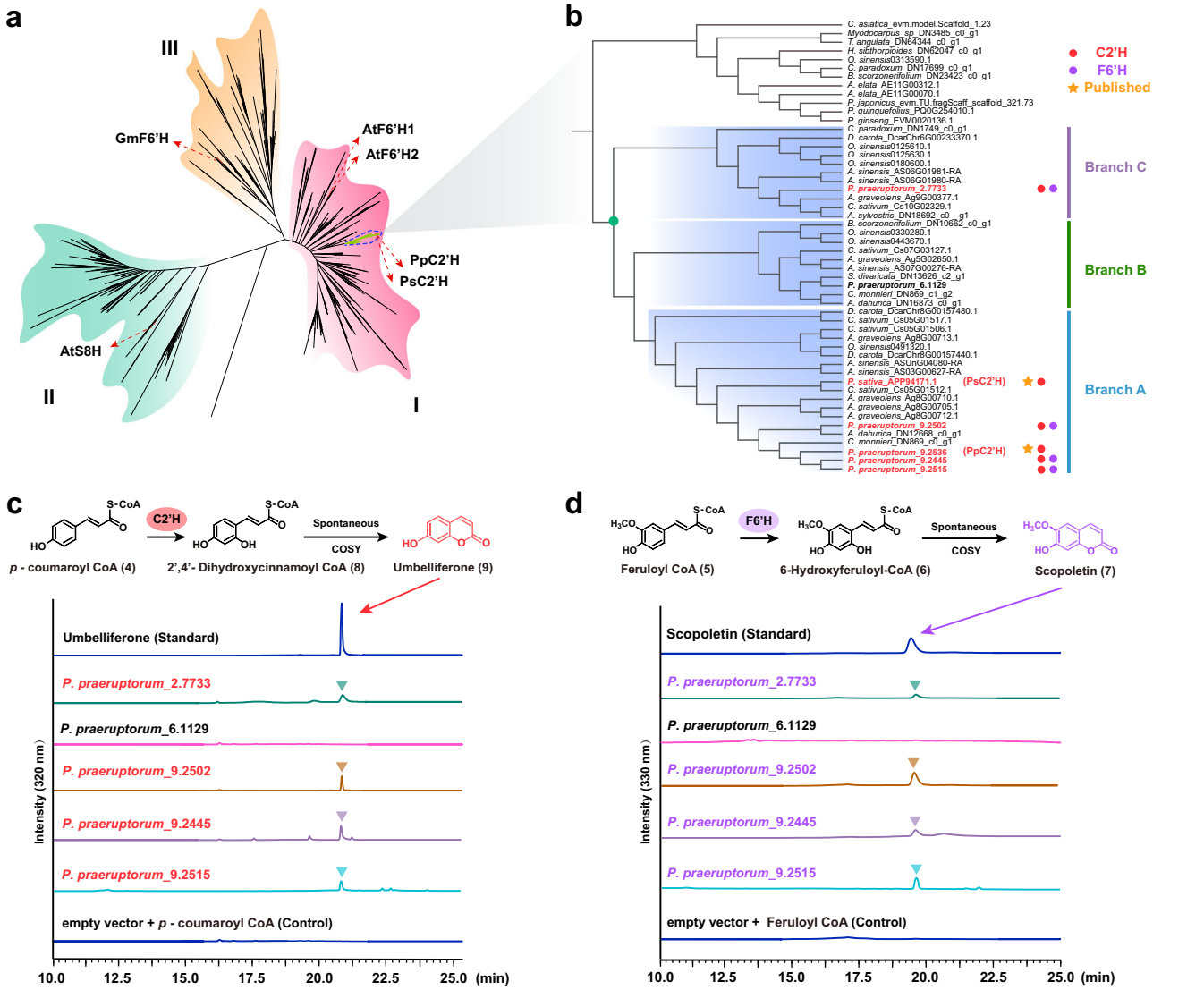

Fig. 3 | The phylogenetic analysis and functional validations of Apiaceae C2'Hs. **a** The phylogeny of 2-OGD gene family from 154 plant genomes and 17 transcriptomes, see Supplementary Fig. 29 for the detailed phylogeny. GenBank accession numbers: PsC2'H, APP94171.1; PpC2'H, ASR80916.1; AtS8H, NP_187896.1 (AT3G12900.1); GmF6'H, AHY03267.1; AtF6'H1, NP_187970.1 (AT3G13610.1); AtF6'H2, NP_175925.1 (AT1G55290.1) **b** The extracted phylogeny of Apiaceae F6'H and C2'H. The green dot indicates the origin of C2'H function. The red and purple dots show that the corresponding gene has C2'H and F6'H function, respectively. The orange pentacle represents functionally characterized genes published already. **c** The C2'H activity assays of selected Apiaceae 2-OGD genes. **d** The F6'H activity assays of selected Apiaceae 2-OGD genes. The LC maps of standards and their corresponding chemical structures are shown in the figure. The retention time of umbelliferone and scopoletin was 20.75 and 19.49 min, respectively. Source data are provided as a Source Data file.

implying an active evolutionary mode for the Apiaceae CC biosynthetic pathway.

Apiaceae C2'Hs arose during early Apioideae evolution by ectopic duplications

Given the conservative sharing of the upstream phenylpropanoid metabolism pathway among angiosperms, the downstream C2'H, C-PT and cyclase should be pivotal for exploring the mechanism underlying the establishment and evolution of CC biosynthesis in Apiaceae. The CC biosynthesis is initiated by C2'H targeting *p*-coumaroyl CoA, and the critical trigger enzyme C2'H is encoded by the 2-oxoglutarate-dependent dioxygenase (2-OGD) gene family[5,6]. To explore the origin and evolution of Apiaceae C2'H, we constructed a phylogenetic tree using all 2-OGDs extracted from 154 plant genomes and 17 transcriptomes (covering 42 angiosperm orders). The selected genomes and transcriptomes aimed to provide an adequate and balanced species sampling with a focus on Apiaceae, serving as the data source of the phylogenetic analysis (Supplementary Tables 11, 12). The phylogenetic tree can be divided into three monophyletic groups, probably due to duplication events in early angiosperm ancestors. Apiaceae C2'Hs fall into Subgroup I (Fig. 3a), whereas the other two subgroups contain *Arabidopsis thaliana* Scopoletin 8-hydroxylase (S8H) and *Glycine max* F6'H, respectively[27,28]. By comparing the gene tree to the species tree and examining the position of Apiaceae C2'Hs in Subgroup I, two rounds of Apiaceae-specific duplications could be recognized, generating three Apiaceae-specific branches, with the previously characterized C2'Hs (C2'Hs of *P. sativa* and *P. praeruptorum*)[5,13] nested in Branch A (Fig. 3b). One can observe that the members of three branches are limited in Apoideae, and include all Apoideae species; thus, we speculate that the duplications giving birth to C2'H occurred during early Apoideae evolution, which was supported by the reconciliation of gene duplication events in Apiaceae (Supplementary Fig. 7). Our following functional characterization revealed that Pp2.7733 from Branch C also exhibited enzymatic activity of C2'H by heterologous

expression in *Escherichia coli* (Fig. 3c), suggesting that the C2′H function likely arose via the earlier duplication event. Since genes derived from the two duplications were not observed to cluster on chromosomes or show synteny in between (Supplementary Fig. 8), Apiaceae C2′Hs are likely to have arisen and expanded by ectopic duplications. The two ectopic duplications likely date back to an early Apioideae ancestor. *P. praeruptorum* has four copies in Branch A due to later self-tandem duplications on Chromosome 9, and the duplicates (Pp9.2445, Pp9.2502, Pp9.2515 and Pp9.2536) all demonstrated C2′H activity as well according to our functional validations (Fig. 3c, Supplementary Figs. 9, 10), despite of some differences in quantitative activities (Supplementary Fig. 11). However, Pp6.1129 in Branch B did not exhibit either activity (Figs. 3c, 3d). As we examined the sequence of Pp6.1129, we found that this protein has truncated sequence with ten amino acids missing/variation at the N-terminus (Supplementary Fig. 12), which we infer likely resulted in the loss-of-function mutation. Since the tested proteins (except Pp6.1129) also exhibit F6′H activity (Fig. 3d) and F6′H is discovered in other species and subgroups with a wider systematic distribution, we speculate that F6′H should be the original function and Apiaceae C2′H is likely derived from F6′H. This hypothesis is also supported by the monophyletic grouping of Apiaceae C2′Hs in the phylogenetic tree, which suggests C2′H is a later-evolved activity, compared with the more widely distributed F6′Hs. At the sequence level, we discovered from the multiple sequence alignment that, C2′Hs have more conserved amino acid residues whereas F6′Hs show more diverse residue types (Supplementary Fig. 13, Supplementary Data 1).

Tandem duplications led to the rise of Apiaceae C-PTs and activity divergence of U6- and U8PT

Although C-PT functions after C2′H in the biosynthetic pathway, it generates product diversity for the first time – demethylsuberosin and osthenol by prenylation at C-6 and C-8 of umbelliferone (Fig. 1), respectively. The two products are the corresponding precursors of linear and angular CCs, making C-PT the key to elucidate the mechanism underlying the linear and angular configurations of CC products. The same 154 genomes and 17 additional transcriptomes resource (Supplementary Tables 11, 12) was used for the evolutionary analyses of C-PT. Phylogenetic analysis indicates that C-PT genes may have undergone a complicated evolutionary trajectory. Despite a small gene family (averaging a dozen members per angiosperm species), the evolutionary trajectory of PTs with the C-C activity was inferred to involve as many as seven times Apiaceae-specific duplication events (Fig. 4a, b), and tandem duplications should be the primary mechanism. For instance, seven out of ten *P. praeruptorum* PT genes (PpPTs) are found to group in a monophyletic group, suggesting their close evolutionary relationships, and they are all located together at one locus on Chromosome 9, forming a gene cluster (Supplementary Fig. 8). These results serve as solid evidence for the emerging mechanism of these PpPTs: by tandem duplications. All three characterized *P. praeruptorum* C-PTs (PpPT1-3)[7] are also located in this gene cluster, suggesting that Apiaceae C-PT enzymes are probably derived from tandem duplication events. To be precise, these tandem duplication events should not be limited to *P. praeruptorum*, as the seven clustered PpPTs scatter over five phylogenetic branches that include orthologous genes from multiple plants. By comparing with the species tree (Fig. 2), one can see the characterized C-PTs (PpPT1-3) and their orthologs are only present in Selineae (*P. praeruptorum*, *A. dahurica*, *C. monnieri* and *S. divaricata*), Tordylieae (*P. sativa*), Coriandreae (*C. sativum*), *Sinodielsia* (*A. sinensis*) and Apieae (*A. graveolens*) (Fig. 4b), and these tribes/lineages are derived from a common ancestor in early Apioideae evolution (Fig. 2)[29,30]. Therefore, the tandem duplications for the emergence of Branches A, B and C should have occurred in the common ancestor of these tribes (Supplementary Fig. 14), and thus genes in these branches evolved C-PT activity through neofunctionalization. The amino acid sequence identities between genes from the three different branches are high (at least above 75%), with 84.43% (median) between A and B, 82.35% (median) between A and C, and 81.56% (median) between B and C (Supplementary Fig. 15), suggesting these genes from different branches are close related and also implying their functional divergence is likely caused by only a few sequence variations.

To test our hypothesis, functional experiments were subsequently conducted. Genes from different branches were selected to be heterogeneously expressed in *E. coli*, and their products were analyzed by High Performance Liquid Chromatography (HPLC) and mass spectrometry (MS). By mapping our experimental results (Fig. 4c, d, Supplementary Figs. 16, 17) and previously characterized enzymes to the PT phylogeny (Fig. 4b), it can be observed that only genes in Branch A, B and C (only present in Selineae, Tordylieae, Coriandreae, *Sinodielsia* and Apieae) possess C-C bonding activity at C-6 of umbelliferone (U6PT, indicated by green dots) or at C-8 (U8PT, indicated by pink dots), while genes in other branches were tested to have O-PT activity rather than C-PT activity (indicated by blue dots). These experimental results align with our evolutionary inference that C-PTs have a more recent origin in later-diverging Apioideae lineages. In addition, it was proved that C-PT activity evolved from O-PT genes through neofunctionalization. After all, the duplication events giving birth to Branches A, B and C took place relatively later (Supplementary Fig. 14).

Branches representing U6PT and U8PT are of close phylogenetic relationship. In *P. praeruptorum*, U6PT (PpPT1) and U8PT (PpPT2) are located in the same gene cluster (Supplementary Fig. 8), indicating that they are the products of tandem duplications and have undergone neo/subfunctionalization. Sequence comparison between proteins encoded by genes in Branch A and C revealed six sites (Leu102-Phe102, Thr161-Ala161, Ile195-Val195, Phe262-Tyr262, Gly335-Ala335 and Lys388-Gln388, using PpPT1 as the reference) with significant sequence discrepancy (Fig. 4e, Supplementary Fig. 18), potentially accounting for the functional divergence of U6- and U8PTs. To determine the critical role of these sites, we changed the amino acids through the site-specific mutations to test the corresponding functional variations. The enzymatic activity assay of mutated PpPT1s and PpPT2 indicate that amino acid at position 161 is key to the transition from U6- to U8PT activity, with Thr161 accounting for the U6PT and Ala161 responsible for U8PT, as the PpPT1 with the Thr161-to-Ala161 (T161A) mutation displayed U8PT activity and PpPT2 with the Ala161-to-Thr161 (A161T) mutation displayed U6PT activity, while other mutations did not show obvious activity variations (Fig. 4f, Supplementary Fig. 19). The three-dimensional protein structures of PpPT1 and PpPT2 were modeled by Alphafold2[31], and suggested to form a complex with umbelliferone and dimethylallyl pyrophosphate (DMAPP) via molecular docking. We infer that PpPT1, with the hydrophilic Thr161, is possibly prone to bind with the hydrophilic side of umbelliferone, leaving DMAPP to the only available C-6 to generate a linear product. PpPT2, with the hydrophobic Ala161, prefers the opposite hydrophobic side of umbelliferone. Thus, DMAPP can only target C-8, and from an angular conformation accordingly (Fig. 4g).

Apiaceae cyclases originated by ectopic duplications

Cyclases responsible for the last step of CC biosynthesis in Apiaceae are members of the CYP450 superfamily. To explore the origin and evolution of Apiaceae cyclases, we extracted all CYP450 proteins from 18 plant genomes and 17 transcriptomes (fewer genomes were used because of too large numbers of CYP450 genes in plant genomes, Supplementary Tables 12, 13), and performed a phylogenetic analysis (Fig. 5a). Cyclases in Apiaceae can be classified into two types according to functions, namely demelthyisuberosin cyclase (DC) and osthenol cyclase (OC), catalyzing linear and angular products, respectively[7,11]. In *P. praeruptorum*, the recently characterized PpDC (Pp7.4765 in Branch A) and PpOC (Pp9.2436-37 in Branch C) have very similar sequences[7], and were recovered as close relatives along with

https://doi.org/10.1038/s41467-024-51285-x

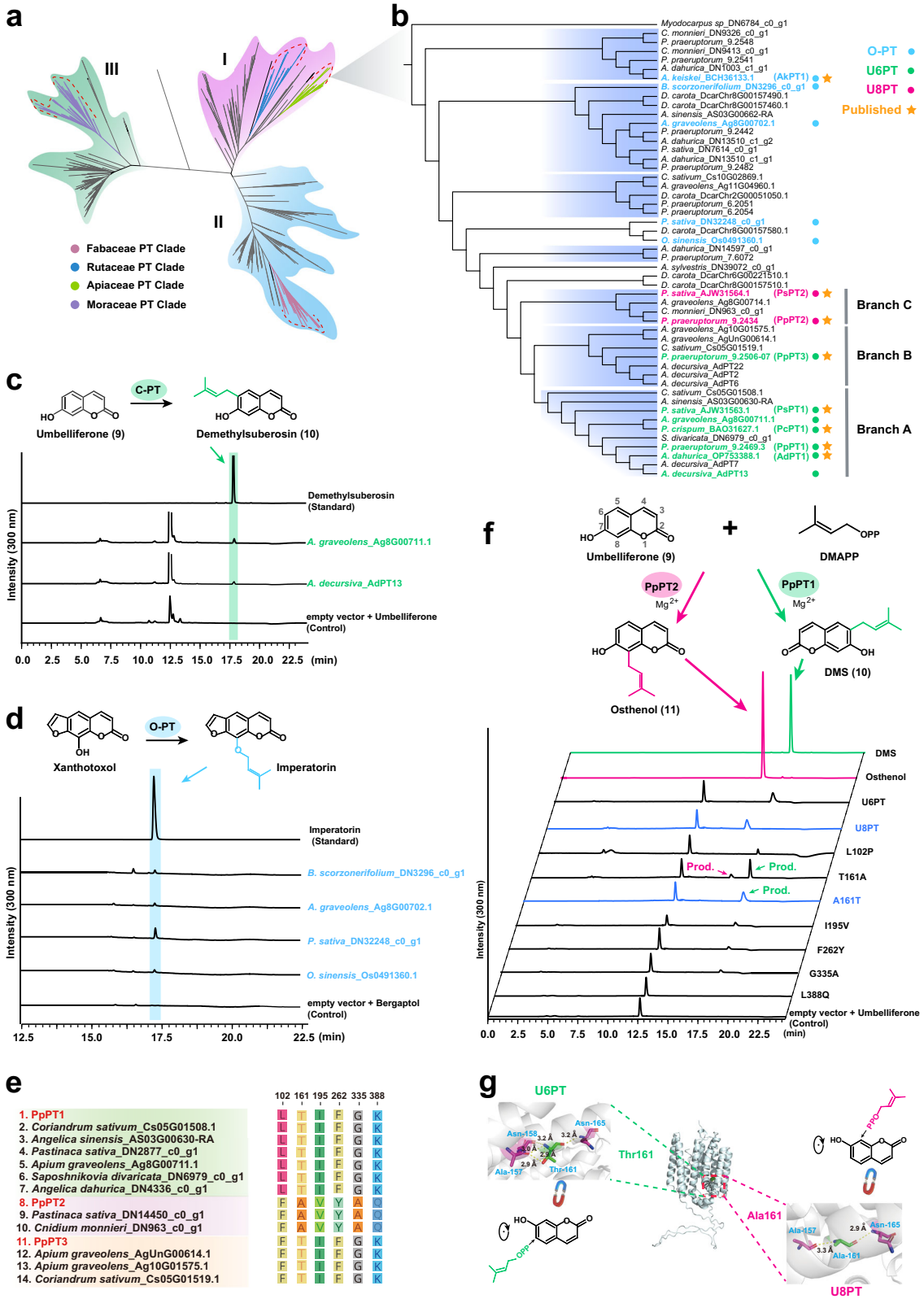

Fig. 4 | The phylogenetic analysis and functional validations of Apiaceae PTs. **a** The phylogeny of PT gene family from 154 plant genomes, see Supplementary Fig. 25 for the detailed phylogeny. **b** The extracted phylogeny of Apiaceae PTs. **c** Enzymatic activity assays of C-PTs by HPLC. **d** Enzymatic activity assays of O-PTs by HPLC. **e** Six significant discrepant amino acid sites between the U6- and U8PTs in Apiaceae. **f** Enzymatic activity assays of PpPT1 and PpPT2 with site-specific mutations by HPLC. DMAPP, dimethylallyl pyrophosphate; DMS, demethylsuberosin. **g** 3-D protein structures of PpPT1 and PpPT2, and hypothesized process of umbelliferone prenylation catalyzed by U6PT and U8PT, respectively. Essential sites (Thr161/Ala161) for functional variations were displayed in colors. Source data are provided as a Source Data file.

https://doi.org/10.1038/s41467-024-51285-x

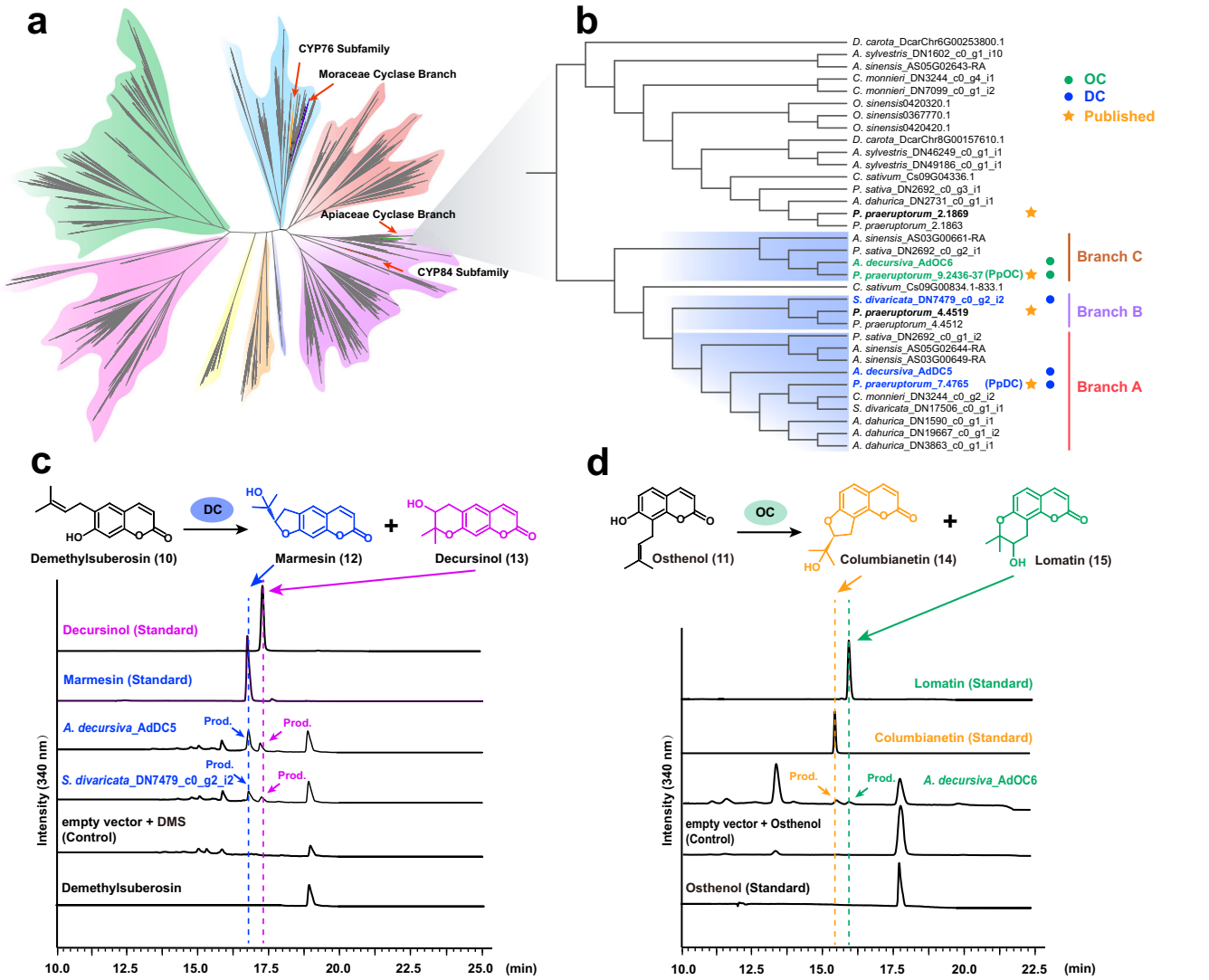

Fig. 5 | The origin and evolution of Apiaceae cyclases. a The phylogeny of CYP450 gene family from 18 plant genomes and 17 plant transcriptomes, see Supplementary Fig. 30 for the detailed phylogeny. **b** The extracted phylogeny of Apiaceae cyclases. **c** Enzymatic activity assays of DC by HPLC. The peaks in 16.7 min and 17.2 min are marmesin and decursinol, respectively. **d** Enzymatic activity assays of OC by HPLC. The peaks in 15.4 min and 16.0 min are columbianetin and lomatin, respectively. Source data are provided as a Source Data file.

other two CYP450 genes (Pp4.4512 and Pp4.4519 in Branch B) in the phylogenetic tree (Fig. 5b). While Pp4.4512 and Pp4.4519 were tested to have extremely weak cyclase activity[7], another *S. divaricate* gene (DN7479) from the same branch was proved to have stronger DC activity (Fig. 5c, Supplementary Figs. 20, 21), suggesting a potential loss of function scenario for the two *P. praeruptorum* genes. Since both Branch A and B genes exhibited DC activity (Fig. 5c, Supplementary Fig. 20), the origin of DC could be traced back to the common ancestor of the two branches. This ancestor serves as a sister group to the OC lineage (Branch C), suggesting a simultaneous divergence of DC and OC lineages. A previous study reported that genes from other branches (outside Branch A, B and C) have no cyclase activity[7]. Therefore, the cyclase activity was limited to Branch A, B and C, comprising genes from only Selineae, Apieae, *Sinodielsia* and Tordylieae, and the cyclases should consequently have originated in the common ancestor of these lineages. Comparatively, Apiaceae cyclase originated posterior to C2′H, and no earlier than C-PTs as well, and finally established complete CC biosynthetic pathway. In Apiaceae genomes that possess both DC and OC (*P. praeruptorum* and *A. sinensis*), intragenomic synteny was not detected between DC and OC genes, suggesting that Apiaceae DC and OC may be derived from an ectopic duplication (Supplementary Fig. 22). In addition, independent gene losses were inferred to have occurred in high frequency, as all three branches were reconciled to have undergone gene loss events, with the OC branch inferred to have experienced the most gene losses (Supplementary Fig. 23).

The establishment and collapse of CC biosynthetic pathway is associated with the origins and losses of enzyme-encoding genes

In summary, Apiaceae C2′H, C-PT and cyclase are the outcomes of specific duplications within Apiaceae (Figs. 3b, 4b and 5b), indicating that strict orthologs of these enzyme genes are exclusive to Apiaceae. More specifically, they are derived from distinct duplication events by means (tandem duplications and ectopic duplications), arising at different phylogenetic nodes within Apioideae. The comprehensively reconciliation of gene duplication events aided in tracing the detailed evolutionary histories of the three key enzyme genes (Supplementary Figs. 7, 14, 23), exhibiting the step-by-step completion of the Apiaceae CC biosynthetic toolbox (Fig. 6). Although C2′H has an earlier origin in Apioideae, the other two enzymes emerged successively afterwards. Till the split of Daucinae-Scandicinae lineage, cyclase finally completed

https://doi.org/10.1038/s41467-024-51285-x

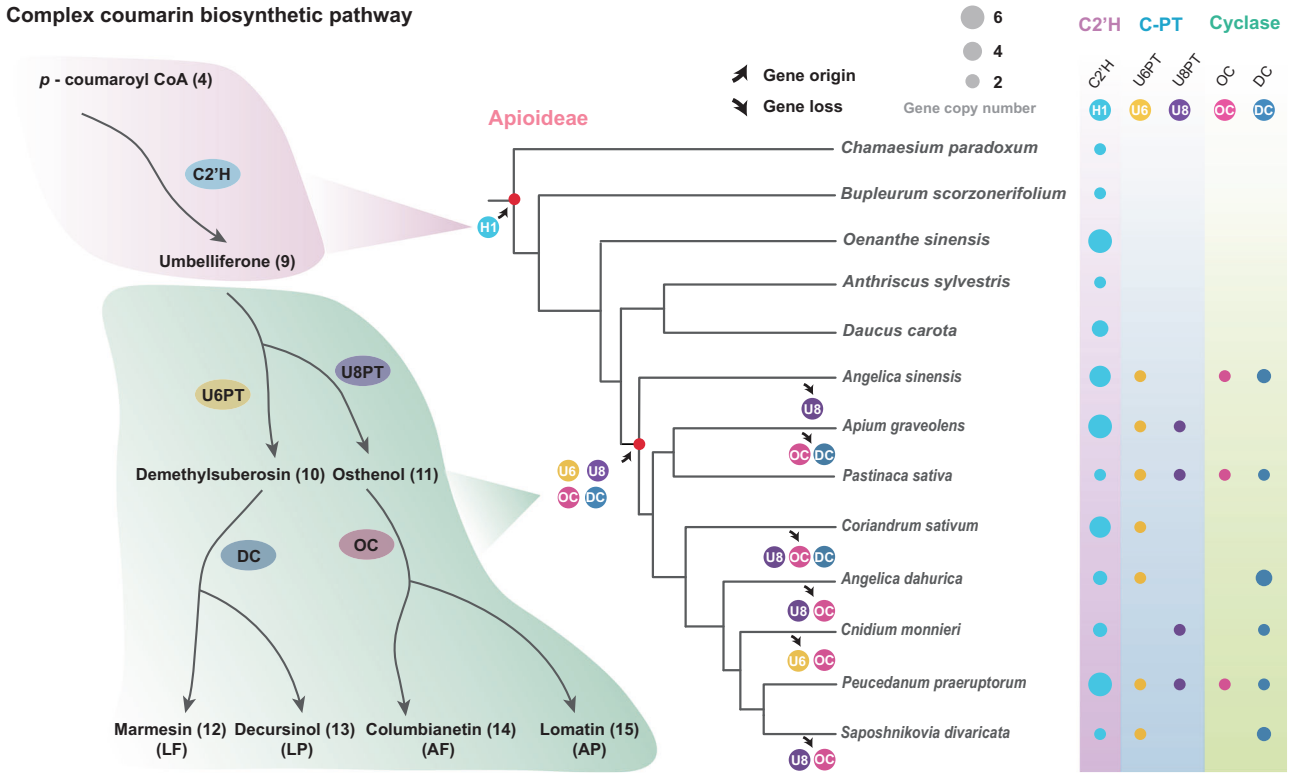

Fig. 6 | The summarized origins and loss of the key enzymes of CC biosynthesis in Apiaceae. Apiaceae C2'Hs could at least date back to the common ancestor of Apioideae, and Apiaceae C-PTs and cyclases likely have originated soon after the split of the Daucinae-Scandicinae lineage. Later secondary losses of these key enzymes are common and more occurred to U8PTs and OCs that responsible for angular CCs. H1, C2'H; U6, U6PT; U8, U8PT.

the last missing step of CC biosynthesis. Theoretically, all Apieae, Sinodielsia, Tordylieae, Coriandreae and Selineae plants should have the complete CC biosynthetic toolkit, and be capable of producing CCs. Nevertheless, secondary losses should have taken place (Fig. 6, Supplementary Figs. 7, 14, 23), and resulting in the absence of certain enzymes that may influence CC product diversity or even lead to the collapse of the pathway; for instance, celery and coriander lost all cyclases, thus cannot produce CCs at all. This reconstructed history explains presence/absence of CCs in current Apiaceae plants and aligns with the known metabolic products in these species, reflecting the genotype-phenotype consistency.

A CC BGC likely accounting for highly accumulating angular CCs in *P. praeruptorum*

Although a number of Apioideae plants have the complete toolkit for CC biosynthesis, the product contents and types vary significantly, suggesting differential biosynthetic efficiencies among taxa. In comparison, *P. praeruptorum* has evolved a stronger biosynthetic ability than other Apiaceae plants, dominantly accumulating angular products. Biosynthetic gene clusters (BGCs) are a particular yet common organization of genes involved in various metabolic pathways, such as, benzoxazinoids[32], momilactone[33], thalianol[34], polyyne protegencin[35], covering a diversity of organisms, e.g., plants, fungi and bacteria[36–38]. The significance of BGCs is speculated to be related with the efficiency of biosynthesis: the close physical distance between up- and downstream enzymes may keep them in pace during transcription[37], and facilitate the rapid processing of intermediates in pathways. We examined the genomic organization of the CC pathway genes in *P. praeruptorum*. By anchoring all PpC2'Hs, PpPTs, PpDC and PpOC to chromosomes (Fig. 7, Supplementary Fig. 8), we observed that all four PpC2'Hs, seven PpPTs (including PpPT1-3) and PpOC are located at the same locus on Chromosome 9, showing a very close physical distance to each other and forming a typical BGC; while PpDC is located on a different chromosome (Chromosome 7), distant from PpC2'Hs and PpPTs (Supplementary Fig. 8). We thus speculate that such a clustered array of genes involved in angular CC biosynthesis should facilitate the corresponding metabolic process, thus strengthens the biosynthetic efficiency of angular CCs in *P. praeruptorum*. This BGC arose by multiple rounds of tandem duplications of PpC2'Hs and PpPTs and the insertion of PpOC via ectopic duplication according to our aforementioned evolutionary inferences. Intergenomic synteny analysis detected a similar BGC on *A. sinensis* Chromosome 3 comprising the same three enzyme genes (Fig. 7). While other Apiaceae species, such as *A. graveolens*, *C. sativum*, *D. carota* and *O. sinensis*, either have incomplete BGCs or lack BGCs entirely (Supplementary Fig. 24), which are in line with fact that only SCs or low content CCs were detected in these taxa[20,21,23,39].

Discussion

Deciphering the evolution of Apiaceae CC biosynthetic pathway

CCs are the characteristic metabolites of Apiaceae, possessing high medical values and promising clinical potentials. Recently, the three key enzymes (C2'H, C-PT and cyclase) were isolated, thereby unveiling the complete CC biosynthetic pathway in Apiaceae[5,7]. Nevertheless, in particular, not all Apiaceae taxa are detected to accumulate CCs, and different species usually accumulate distinct CCs products as well[7,14–18]. To understand the molecular basis for the presence/absence and diversity of CCs in Apiaceae, we performed a comprehensive evolutionary study using all available genomes and transcriptomes of Apiaceae plants, among which, included our newly assembled *P. praeruptorum*, an Apiaceae plant with all coumarin types and high content of angular CCs. The evolutionary histories of the three key enzyme genes – C2'Hs, C-PTs and cyclases – were reconstructed based on phylogenetic analyses. Then, combined with functional experiments, we revealed that later-origins and secondary losses of these

https://doi.org/10.1038/s41467-024-51285-x

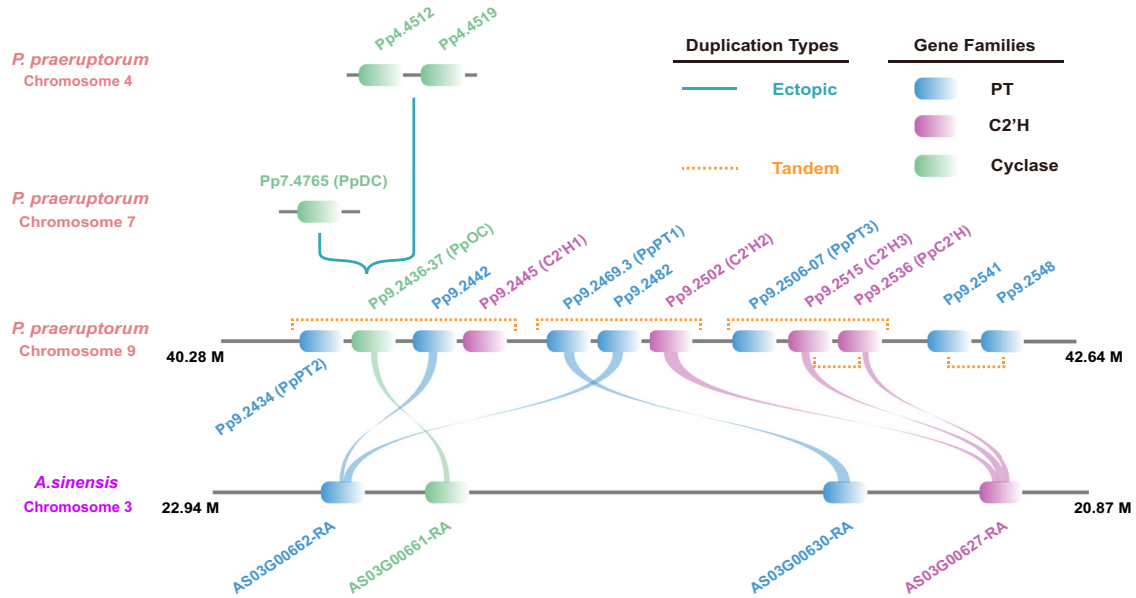

Fig. 7 | The biosynthetic gene cluster of CCs in *P. praeruptorum* and *A. sinensis*. PpC2'Hs, PpPTs and PpOC are located in a gene cluster of Chromosome 9 in the *P. praeruptorum* genome, while PpDC is located in Chromosome 7, not belonging to the gene cluster. Multiple rounds of tandem duplications involving PpC2'Hs and PpPTs can be recognized according to the gene phylogenies. *A. sinensis* has a similar gene cluster with containing C2'H, C-PT and cylcase in a collinear genomic block to that of *P. praeruptorum*.

enzymes should explain the limited distribution of CC-containing taxa within Apiaceae and the diversity of CCs products.

The three enzymes are all indispensable for the biosynthesis of CCs, among which, C-PTs are expected to make the prior contribution to the product diversity, due to the sub/neofunctionalized U6PT and U8PT activities that generate linear and angular precursors, respectively. The characterized site 161 (Ala/Thr) has been proven of high importance to enzymatic properties between U6PT and U8PT. Similarly, Glu303 in PpDC and Asp301 in PpOC define their distinct catalyzing functions, continuing reactions downstream to U6- and U8PT, respectively[7]. Therefore, the C-PT sequence diversity should be the decisive factor for product diversity. Nevertheless, our experiments only tested the variations in single amino acids, and have not extended to combinations of different candidate sites/amino acids, so there is possibility that certain simultaneous mutations at multiple sites could also change the activity of these enzymes.

Previous studies also proposed a hypothesis that angular CCs are later-evolved metabolites derived from linear CCs because angular CCs are always in coexistence with linear products[10,40]. Angular CCs are restricted to only a few plants that meanwhile produce linear CCs as well[1,2,10,40]. However, such evidence supporting the hypothesis is circumstantial. According to our evolutionary analyses, angular CCs should have arisen no later than linear CCs, as the cyclases phylogeny indicate that Apiaceae DC and OC arose simultaneously (phylogenetically sister to each other), and none of their outgroup genes display DC or OC activities along (no prior DC or OC functions evolved) (Fig. 5b).

In conclusion, our study reconstructed a detailed trajectory for the establishment and evolution of CC biosynthesis in Apiaceae. This reconstruction enhances our understanding of molecular mechanisms underlying coumarin diversity among Apiaceae plants, and should inspire the identification of key genes involved in CC biosynthesis in other plant lineages.

Parallel evolution of biosynthetic pathways of complex coumarin in angiosperms

Other than Apiaceae, CCs also accumulate in three other phylogenetically distant angiosperm lineages, namely Rutaceae, Moraceae and Fabaceae[1–4,41]. Yet, Apiaceae is by far the only lineage in which the complete biosynthetic pathway has been decoded[5,7]. As the pathway is established by genes derived from Apiaceae-specific duplication events, we speculate the three enzyme genes in other three plant lineages cannot be orthologs to the Apiaceae genes, but paralogs at most. The phylogenies of C2'H, PT and cyclase genes provide evolutionary evidence for our hypothesis. For instance, angiosperm PTs diverged into three subfamilies, and the majority of species maintain PTs in Subfamilies I and III, whereas lost their members in Subfamily II (Fig. 4a, Supplementary Fig. 25, Supplementary Table 14). Nevertheless, Apiaceae and Rutaceae only maintain low gene copies in Subfamilies I and III, but exhibit significant expansions of Subgroup II genes (Fig. 4a, b), which should have provided abundant materials to gene neofunctionalization of Apiaceae and Rutaceae PTs. As a result, Apiaceae Subgroup II genes developed diverse activities including C-C bonds at U6 and U8, and C-O bonds as well; and Rutaceae was also found to have developed genes with C-O activity (CmiPT and CpPT) in Subgroup II[9,10,12]. Therefore, it is logical to speculate that the Rutaceae C-PT genes are also likely among their sharply expanded Subfamily II members, as there are not many available options for Rutaceae in other two subfamilies. In contrast to Apiaceae and Rutaceae, the other two taxa – Fabaceae and Moraceae have absolutely lost in Subgroup II. Thus, Fabaceae and Moraceae C-PTs cannot be derived from Subgroup II, but must be among either Subgroup I or III, which happened to have expanded in these two taxa. In fact, *G. max* has 15 genes in Subgroup I (including functionally characterized GmDTs) and *Ficus erecta* has 38 genes in Subgroup III (including FcPTs) (Fig. 4a, Supplementary Fig. 25, Supplementary Table 14), perfectly supporting our inference. Taken together, Apiaceae, Fabaceae, Rutaceae and Moraceae must have developed functional C-PTs by lineage-specific duplications of different PT subfamilies, indicating the parallel evolution of C-PT activities in angiosperms[2,8,10,12].

Likewise, cyclase genes should also have arisen in parallel in different angiosperm lineages. Apiaceae cyclase genes belong to a subfamily sister to CYP84. Other than Apiaceae cyclases, only one additional cyclase (CYP76F112 in *F. carica*) in other plant lineages (Moraceae) has been identified[11], and it falls into a different phylogenetic branch, distant from Apiaceae cyclases (Fig. 5a). Such

dispersed distribution of gene phylogeny indicates that the Apiaceae and Moraceae cyclases likely originated independently in parallel.

To sum up, lineage-independent gene duplication events are likely to have provided materials for neofunctionalization. Long-term evolutionary selection likely guided a convergence in functions, and finally resulted in the recurrent origins of CC biosynthesis in distantly related angiosperm taxa.

Methods

Plant material collection, genome sequencing, assembly and annotation

The samples of *P. praeruptorum* utilized in this study were collected from the medicinal botanical garden of China Pharmaceutical University (31°54′N, 118°54′E). The plant leaves were first cleaned with 75% alcohol and subsequently with pure water for DNA extraction. Genomic DNA was extracted using the cetyltrimethylammonium bromide (CTAB) method, and a single-molecule real-time (SMRT) DNA library on the PacBio Sequel platform was employed for DNA sequencing. For the construction of the HiC DNA library, the DNA samples underwent ultrasonic crushing, end repair, A-tail addition, adapter addition, purification, PCR amplification, and then the paired-end reads (150 bp for each end) were sequenced on the Illumina Hiseq platform. Finally, a total of 188.37 Gb PacBio data and 178.09 GB HiC data were obtained collectively for subsequent genome assembly (Supplementary Table 3). The RNA used for transcriptome sequencing was extracted from the leaves and roots of *P. praeruptorum*.

Firstly, the self-correction program from Falcon (v3.1.0)[42] was utilized to correct the raw PacBio data, and the corrected reads were subsequently processed by Smartdenovo (v1.0.0)[43] to assemble the genome. Subsequently, we calculated the GC content and average depth of the assembled genome sequence using 200 kb as a window to evaluate whether there was GC bias in the sequencing data and whether there was contamination in the samples. Then, three rounds of polishing with Pilon (v1.20)[44] were applied to correct assembled mistakes caused by snps, indels and gaps based on Illumina sequencing reads. Purge haplotigs (v1.1.0)[45] was used to refine the assembly and collapse homologous regions, with the parameters: "-l 10 -m 63 -h 195" for "purge_haplotigs contigcov" ("purge_haplotigs hist" and "purge_haplotigs purge" were performed with default parameters). For the acquisition of a chromosome-level genome, about 178 Gb of HiC data were used to execute HiC chromosome conformation in conjunction with ALLHiC (v0.9.8)[46]. After performing the five steps of pruning, partition, rescue, optimization, and building, the generated scaffold-level genome was transformed into chromatin contact matrix using 3D-DNA (v170123)[47] and Juicer (v1.6)[48] and further visualized via Juicebox (v1.11.08)[49]. After manual adjustments, including correcting inversion errors and re-joining contigs, 11 longest scaffolds with a total length of 1.735 Gb (accounting for 99.68% of the total genome assembly), was selected presumably to correspond to the 11 chromosomes of the haplotype genome of *P. praeruptorum*. Finally, BUSCO (v1.0.0)[50] and LAI (v2.9.8)[51] were employed to assess integrity and continuity of assembly.

We employed a method that integrates homology-based prediction, de novo prediction, and transcriptome-based prediction for gene and functional prediction. During the homology-based prediction process, we aligned the protein sequences from five published plant genomes (*A. thaliana*, *Citrus sinensis*, *G. max*, *C. sativum*, and *D. carota*) with the *P. praeruptorum* genomes. In the RNA-seq-based prediction process, aimed at optimizing genome annotation, transcriptome data from various tissues were aligned to the assembled genome using Hisat2 (v2.0.4)[52] to identify exon regions and splice positions. Lastly, EvidenceModeler (v1.1.1)[53] was employed to generate a consensus gene set based on the aforementioned three processes.

Phylogenetic reconstruction and divergence time estimation

In the process of phylogenetic reconstruction in Apiaceae, we employed a strategy that combines three distinct approaches. For phylogenetic reconstruction based on amino acid sequences, we curated 17 genomes, which included the newly generated genome sequences of *P. praeruptorum*. Additionally, we assembled 17 transcriptomes using transcriptome sequencing reads from a public database using Trinity (v2.9.1)[54], and subsequently, Transdecoder (https://github.com/TransDecoder/TransDecoder) was selected for annotation by blasting the assembled results against the public protein database. Ultimately, non-redundant longest transcripts were obtained using cd-hit (v4.8.1)[55]. Following this, these amino acid sequences were inputted into OrthoFinder (v2.5.5)[26] to identify low copy orthologous families (OGs) using default parameters. For the generation of refined single-copy genes (RSCGs), we initially filtered out low copy orthologous families with a copy number of paralogous genes less than three in a single species. Subsequently, we conducted additional screening to ensure that at least 90% of the species in each low copy gene family had more than one gene copy. After two rounds of filtering, a Python script was employed to de-redundantly process each low copy gene family, ensuring that the orthologous genes in each family were the longest copies of the paralogous genes in each species. Multi-species concatenated nucleotide and amino acid datasets were constructed, respectively. Lastly, iqtree (v2.2.2.3)[56–58] was utilized to infer the maximum likelihood trees using these concatenated data. In the route of inferring the species tree based on coalescent gene trees, iqtree (v2.2.2.3)[57] and Astral (v5.7.1)[59] were employed to infer the maximum likelihood trees and to summarize the coalescent species tree and quartet supports with default settings (-t 2), respectively.

The divergence time of species in Apiaceae was estimated using MCMCtree package in PAML (v4.10.0)[60–62]. A total of 17 fossils (Supplementary Table 15) were chosen to calibrate the chronogram of Apiaceae plants. To ensure the rationality of sampling, we set the values of burn-in, sampfreq, and nsample to 400,000, 10, and 100,000, respectively.

Phylogenetic analyses of enzyme gene families

To explore the evolutionary trajectory of the PT and C2′H genes in angiosperms, we collected 154 plant genomes and 17 Apiales plant transcriptomes. Subsequently, we used Orthofinder (v2.5.5)[26] to identify homologous genes. Following the removal of invalid sequences, iqtree (v2.2.2.3)[57] was employed to reconstruct the gene tree. For the cyclase gene, we selected 18 genomes and 17 transcriptomes from five families (Supplementary Tables 12, 13). Blastp (v2.13.0)[63] and Hmmsearch (v3.3.2)[64] were used to identify homologous genes through homology-based searches. To further trace duplication events of the three genes in Apiaceae, we extracted the monophyly from the phylogenetic topologies in which the functionally validated genes reside. Subsequently, we ensured the accuracy and integrity of sequences based on transcripts and fed them to iqtree (v2.2.2.3)[57] after alignment.

Extraction of total RNA and acquisition of cDNA

Total RNA was extracted from the roots and leaves using the EASYspin Universal Plant RNA Kit (Aidlab, Beijing, China). The extracted RNA was then subjected to agarose gel electrophoresis to check for contamination. Subsequently, the purity and concentration were determined using an ultraviolet spectrophotometer. The extracted RNA was promptly reverse transcribed using the TransScript All-in-One First-Strand cDNA Synthesis SuperMix for PCR (TransGen Biotech, Beijing, China) to obtain cDNAs. The resulting cDNA was stored at −20 °C for subsequent experiments.

Expression and purification of recombinant PpC2′Hs

The open reading frame (ORF) of four PpC2′H genes were amplified using cDNA as a template with the PrimeSTAR® Max (Takara Bio Inc.,

Kusatsu, Japan) and specific primers (Supplementary Table 16). Subsequently, the genes were digested with NdeI and EcoRI and ligated into the pET28a plasmid to generate pET28a-C2′H, allowing the expression of a fusion protein with an N-terminal histidine tag. The recombinant plasmids were then transformed into the *E. coli* BL21 (DE3). Positive transformants were cultured in Luria-Bertani medium at 37 °C and 220 rpm until the OD_{600} reached 0.6–0.8. The culture was induced with 0.5 mM isopropyl β-d-1-thiogalactopyranoside (IPTG) for 12 h at 16 °C and subsequently centrifuged at 4000 rpm for 15 min at 4 °C. Cells were re-suspended in a buffer (20 mM HEPES, 20 mM imidazole, 500 mM NaCl, 10% glycerol, pH 7.5) and sonicated for 15 min on ice. Subsequently, the resulting lysate was centrifuged at 12,000 rpm for 60 min at 4 °C to collect the supernatant for further purification. The recombinant protein was purified using Ni-NTA columns (Smart-Lifesciences, Changzhou, China), and the adsorbed protein was eluted using an elution buffer (20 mM HEPES, 300 mM imidazole, 500 mM NaCl, 10% glycerol, pH 7.5). The purified protein was analyzed by SDS-PAGE, and the protein concentration was determined using a Micro spectrophotometer (KAIAO, Beijing). Finally, the purified protein was stored at −80 °C for future research. For the enzyme reaction (200 μL), 20 μg of purified protein was mixed with 100 mM Tris HCl (pH 7.0), 0.5 mM $FeSO_4$, 2 mM sodium ascorbate, 2 mM 2-oxoglutarate, and 200 μM substrate (*p*-coumaric acid, *p*-coumaroyl CoA, ferulic acid and feruloyl CoA). The reaction mixture was incubated at 20 °C and 300 rpm for 30 min, and then it was terminated by the addition of 200 μL of methanol. The reaction products were detected using HPLC and MS.

Cloning and functional verifications of PTs

Heterologous expression of candidate PTs in *E. coli* was conducted following a previous report[7]. The ORF of PpPTs was amplified using the specific primers (Supplementary Table 16). PCR products were cloned into the BamHI and EcoRI restriction sites of the pETDuet-1 vector using the peasy-basic seamless cloning and assembly kit (TransGen Biotech, Beijing, China). Re-constructed plasmids were then transformed into *E. coli* DH5α strain, and monoclonal colonies with positive sequencing results were further cultivated for plasmid extractions. All the recombinant plasmids were individually inserted into *E. coli* competence cells and cultured in Luria-Bertani medium at 37 °C and 220 rpm until the OD_{600} reached 0.6–0.8. Subsequently, they were induced with 0.4 mM IPTG and incubated for 12–16 h at 25 °C and 120 rpm. The bacterial cells were harvested, and recombinant proteins were used for relevant activity verification experiments. Bioinformatics studies have shown that PTs are membrane proteins, making it difficult to purify heterologously expressed proteins of AdPT. Therefore, crude proteins were used to analyze their activities[9,10,65]. Enzymatic reactions were conducted using a shaking incubator at 220 rpm and 25 °C for 5 h. The reaction broth consisted of 100 μL of crude enzyme, 200 μM umbelliferone, 200 μM $MgCl_2$, and 100 μM DMAPP in 200 μL of 50 mM Tris-HCl, pH 8.0. The supernatant was centrifuged at 15,000 rpm for 10 min, and 10 μL of the supernatant was analyzed by HPLC and MS.

Cloning and functional verifications of CYP450s

Heterologous expression of CYP450s in yeast and in vitro assays were conducted according to our previous report[7]. The ORF of CYP450s was amplified using specific primer (Supplementary Table 16). Subsequently, the genes were cloned into the BamH I/EcoR I sites of the yeast expression vector pYES2.0. The recombinant plasmids were then transformed into *Saccharomyces cerevisiae* strain WAT11. Positive transformants were screened on solid SC-U (SC dropout medium without uracil) containing 20 g/L glucose and then cultured in liquid SC-U medium until the OD_{600} reached 2-3. Subsequently, the cells were centrifuged at 4000 rpm for 10 min and washed at least three times with ddH_2O to remove glucose residue. The cell precipitate was then transferred to an SC-U medium containing 20 g/L galactose to induce the expression of the target protein. For preliminary activity screening, osthenol and demethylsuberosin were added to the cultures with a final concentration of 100 μM. After incubation at 29 °C for 4 h, methanol was added in an equal volume to terminate the reaction. The reaction supernatants were collected for HPLC-MS analysis.

LC/MS analysis

All the reactions are analyzed using a C-18 chromatographic column (4.6 × 250 mm; 2.5 μm). For AsC2′H, the gradient contains solvents A (0.1% formic acid in ultrapure water) and solvents B (methanol) with the following method: 0 min, 10% B; 5 min, 15% B; 15 min, 60% B; 22 min, 60% B. The flow rate was maintained at 0.5 mL/min. For PTs and CYPs, the gradient contains solvents A (0.1% formic acid in ultrapure water) and solvents B (methanol) with the following method: 0 min: 70:30 (v/v), 5 min: 35:65 (v/v), 12 min: 95:5 (v/v), 14 min: 95:5 (v/v), 16 min: 70:30 (v/v), 22 min: 70:30 (v/v). The flow rate was maintained at 0.5 mL/min. The detection wavelength for *p*-coumaroyl CoA, umbelliferone, DMS, osthenol, marmesin, and columbianetin is 340 nm using Shimadzu LC-2010AT. For MS analysis, Agilent Poroshell 120 SB-Aq (3.0 × 150 mm, 2.7 μm) was used, and the mobile phase consisted of 0.1% HCOOH aqueous solution and methanol. The linear elution conditions were as follows: 0 min: 85:15 (v/v), 3 min: 85:15 (v/v), 8 min: 20:80 (v/v), 12 min: 5:95 (v/v), 16 min: 85:15 (v/v), 18 min: 85:15 (v/v). The conditions of the ESI source were as follows: drying gas (N2) flow rate, 8.0 L/min; Collision energy, 35 eV; Spray voltage, 3.5 kV; Capillary temperature, 320 °C; Aux gas heater temp, 300 °C; Sheath gas flow rate, 35 arb; Aux gas flow rate, 15 arb; Sweep gas flow rate, 5 arb. All gases used, including aux gas, sheath gas, and sweep gas, were of high purity N_2. All operations and data analyses were performed in positive ion mode.

Site-directed mutation of PpPT1 and PpPT2

Specific Primers (Supplementary Table 16) for site-directed mutagenesis were designed based on the docking results and active site analysis. Site-directed mutagenesis was conducted using a PCR method with KOD-plus-neo. Mutants were extracted from *E. coli* DH5α for sequencing and the positive plasmids were subsequently transferred to *E. coli* BL21(DE3) for protein expression and activity test. Finally, functional verifications will inspect the influence of the mutation on the production of DMS and osthenol.

Statistics and reproducibility

GraphPad Prism 9.5 software was used for regular statistical analysis. A two-tailed Student's *t* test was used to calculate significant differences among samples or groups. Without special statement, all the reactions were conducted at three biological replicates ($n = 3$). The enzymes or crude proteins produced by the correspondingly empty vector were used as the negative control.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The raw genome and transcriptome sequencing data have been deposited into the National Center for Biotechnology Information, NIH, Sequence Read Archive under accession number PRJNA1095837. The assembly and annotation data reported in this paper have been upload Figshare: https://doi.org/10.6084/m9.figshare.26335837.v1. All the genes identified in this study were listed in Supplementary Table 17. Source data are provided with this paper.

References

1. Bourgaud, F. et al. Biosynthesis of coumarins in plants: a major pathway still to be unravelled for cytochrome P450 enzymes. *Phytochem Rev.* **5**, 293–308 (2006).

2. Limones-Mendez, M. et al. Convergent evolution leading to the appearance of furanocoumarins in citrus plants. *Plant Sci.* **292**, 110392 (2020).
3. Robe, K., Izquierdo, E., Vignols, F., Rouached, H. & Dubos, C. The Coumarins: Secondary Metabolites Playing a Primary Role in Plant Nutrition and Health. *Trends Plant Sci.* **26**, 248–259 (2021).
4. Rodrigues, J. L. & Rodrigues, L. R. Biosynthesis and heterologous production of furanocoumarins: perspectives and current challenges. *Nat. Prod. Rep.* **38**, 869–879 (2021).
5. Yao, R. et al. Identification and functional characterization of a p-coumaroyl CoA 2′-hydroxylase involved in the biosynthesis of coumarin skeleton from Peucedanum praeruptorum Dunn. *Plant Mol. Biol.* **95**, 199–213 (2017).
6. Vialart, G. et al. A 2-oxoglutarate-dependent dioxygenase from Ruta graveolens L. exhibits p-coumaroyl CoA 2′-hydroxylase activity (C2′H): a missing step in the synthesis of umbelliferone in plants. *Plant J.* **70**, 460–470 (2012).
7. Zhao, Y. et al. Two types of coumarins-specific enzymes complete the last missing steps in pyran- and furanocoumarins biosynthesis. *Acta Pharm. Sin. B* **14**, 869–880 (2024).
8. Munakata, R. et al. Convergent evolution of the UbiA prenyltransferase family underlies the independent acquisition of furanocoumarins in plants. *N. Phytol.* **225**, 2166–2182 (2020).
9. Karamat, F. et al. A coumarin-specific prenyltransferase catalyzes the crucial biosynthetic reaction for furanocoumarin formation in parsley. *Plant J.* **77**, 627–638 (2014).
10. Munakata, R. et al. Molecular evolution of parsnip (Pastinaca sativa) membrane-bound prenyltransferases for linear and/or angular furanocoumarin biosynthesis. *N. Phytol.* **211**, 332–344 (2016).
11. Villard, C. et al. A new P450 involved in the furanocoumarin pathway underlies a recent case of convergent evolution. *N. Phytol.* **231**, 1923–1939 (2021).
12. Munakata, R. et al. Parallel evolution of UbiA superfamily proteins into aromatic O-prenyltransferases in plants. *Proc. Nat. Acad. Sci.* **118**, e2022294118 (2021).
13. Roselli, S. et al. A bacterial artificial chromosome (BAC) genomic approach reveals partial clustering of the furanocoumarin pathway genes in parsnip. *Plant J.* **89**, 1119–1132 (2017).
14. Ivie, G. W., Beier, R. C. & Holt, D. L. Analysis of the garden carrot (Daucus carota L.) for linear furocoumarins (psoralens) at the sub parts per million level. *J. Agric Food Chem.* **30**, 413–416 (1982).
15. Zhang, T. et al. Transcriptomic and Metabolomic Differences Between Two Saposhnikovia divaricata (Turcz.) Schischk Phenotypes With Single- and Double-Headed Roots. *Front Bioeng Biotechnol* **9** (2021).
16. Laribi, B., Kouki, K., M'Hamdi, M. & Bettaieb, T. Coriander (Coriandrum sativum L.) and its bioactive constituents. *Fitoterapia* **103**, 9–26 (2015).
17. Brinkhaus, B., Lindner, M., Schuppan, D. & Hahn, E. G. Chemical, pharmacological and clinical profile of the East Asian medical plant Centella aslatica. *Phytomedicine* **7**, 427–448 (2000).
18. Liu, W. et al. Systematic Characterization and Identification of Saikosaponins in Extracts From Bupleurum marginatum var. stenophyllum Using UPLC-PDA-Q/TOF-MS. *Front Chem.* **9**, 747987 (2021).
19. Iorizzo, M. et al. A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nat. Genet.* **48**, 657–666 (2016).
20. Song, X. et al. Deciphering the high-quality genome sequence of coriander that causes controversial feelings. *Plant Biotechnol. J.* **18**, 1444–1456 (2020).
21. Song, X. et al. The celery genome sequence reveals sequential paleo-polyploidizations, karyotype evolution and resistance gene reduction in apiales. *Plant Biotechnol. J.* **19**, 731–744 (2021).
22. Han, X. et al. The chromosome-level genome of female ginseng (Angelica sinensis) provides insights into molecular mechanisms and evolution of coumarin biosynthesis. *Plant J.* **112**, 1224–1237 (2022).
23. Liu, J.-X. et al. High-quality genome sequence reveals a young polyploidization and provides insights into cellulose and lignin biosynthesis in water dropwort (Oenanthe sinensis). *Ind. Crop Prod.* **193**, 116203 (2023).
24. Liu, R., Feng, L., Sun, A. & Kong, L. Preparative isolation and purification of coumarins from Peucedanum praeruptorum Dunn by high-speed counter-current chromatography. *J. Chromatogr. A* **1057**, 89–94 (2004).
25. Chu, S. et al. Comparative analysis and chemical profiling of different forms of Peucedani Radix. *J. Pharm. Biomed. Anal.* **189**, 113410 (2020).
26. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
27. Siwinska, J. et al. Scopoletin 8-hydroxylase: a novel enzyme involved in coumarin biosynthesis and iron-deficiency responses in Arabidopsis. *J. Exp. Bot.* **69**, 1735–1748 (2018).
28. Wei, S., Zhang, W., Fu, R. & Zhang, Y. Genome-wide characterization of 2-oxoglutarate and Fe(II)-dependent dioxygenase family genes in tomato during growth cycle and their roles in metabolism. *BMC Genomics* **22**, 126 (2021).
29. Wen, J. et al. Backbone phylogeny and evolution of Apioideae (Apiaceae): New insights from phylogenomic analyses of plastome data. *Mol. Phylogenet Evol.* **161**, 107183 (2021).
30. Wen, J. et al. A transcriptome-based study on the phylogeny and evolution of the taxonomically controversial subfamily Apioideae (Apiaceae). *Ann. Bot.* **125**, 937–953 (2020).
31. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
32. Wu, D., Jiang, B., Ye, C. Y., Timko, M. P. & Fan, L. Horizontal transfer and evolution of the biosynthetic gene cluster for benzoxazinoids in plants. *Plant Commun.* **3**, 100320 (2022).
33. Kitaoka, N. et al. Interdependent evolution of biosynthetic gene clusters for momilactone production in rice. *Plant Cell* **33**, 290–305 (2021).
34. Field, B. & Osbourn, A. E. Metabolic diversification-independent assembly of operon-like gene clusters in different plants. *Science* **320**, 543–547 (2008).
35. Mullins, A. J. et al. Discovery of the Pseudomonas Polyyne Protegencin by a Phylogeny-Guided Study of Polyyne Biosynthetic Gene Cluster Diversity. *mBio* **12**, e0071521 (2021).
36. Rokas, A., Wisecaver, J. H. & Lind, A. L. The birth, evolution and death of metabolic gene clusters in fungi. *Nat. Rev. Microbiol* **16**, 731–744 (2018).
37. Liu, Z. et al. Formation and diversification of a paradigm biosynthetic gene cluster in plants. *Nat. Commun.* **11**, 5354 (2020).
38. Jensen, P. R. Natural Products and the Gene Cluster Revolution. *Trends Microbiol* **24**, 968–977 (2016).
39. Wang, Y. H. et al. Telomere-to-telomere carrot (Daucus carota) genome assembly reveals carotenoid characteristics. *Hortic. Res* **10**, uhad103 (2023).
40. Berenbaum, M. Coumarins and Caterpillars: A Case for Coevolution. *Evolution* **37**, 163–179 (1983).
41. Harborne, J. B. The natural coumarins: occurrence, chemistry and biochemistry (Book). *Plant Cell Environ.* **5**, 435–436 (1982).
42. Chin, C. S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
43. Liu, H., Wu, S., Li, A. & Ruan, J. SMARTdenovo: a de novo assembler using long noisy reads. *GigaByte* **2021**, gigabyte15 (2021).
44. Walker, B. J. et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE* **9**, e112963 (2014).

45. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinforma.* **19**, 460 (2018).
46. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).
47. Dudchenko, O. et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
48. Durand, N. C. et al. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **3**, 95–98 (2016).
49. Durand, N. C. et al. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst.* **3**, 99–101 (2016).
50. Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: Assessing Genomic Data Quality and Beyond. *Curr. Protoc.* **1**, e323 (2021).
51. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res* **46**, e126 (2018).
52. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
53. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
54. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
55. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
56. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
57. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
58. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
59. Mirarab, S. & Warnow, T. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* **31**, i44–i52 (2015).
60. Jeffares, D. C., Tomiczek, B., Sojo, V. & dos Reis, M. A Beginners Guide to Estimating the Non-synonymous to Synonymous Rate Ratio of all Protein-Coding Genes in a Genome. *Methods Mol. Biol.* **1201**, 65–90 (2015).
61. dos Reis, M. & Yang, Z. Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. *Mol. Biol. Evol.* **28**, 2161–2172 (2011).
62. dos Reis, M. & Yang, Z. Bayesian Molecular Clock Dating Using Genome-Scale Datasets. *Methods Mol. Biol.* **1910**, 309–330 (2019).
63. Boratyn, G. M. et al. BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.* **41**, W29–W33 (2013).
64. Potter, S. C. et al. HMMER web server: 2018 update. *Nucleic Acids Res* **46**, W200–W204 (2018).
65. Bu, X. L. et al. Constructing Microbial Hosts for the Production of Benzoheterocyclic Derivatives. *ACS Synth. Biol.* **9**, 2282–2290 (2020).
66. Jeong, G.-S. et al. Lignans and Coumarins from the Roots of Anthriscus sylvestris and Their Increase of Caspase-3 Activity in HL-60 Cells. *Biol. Pharm. Bull.* **30**, 1340–1343 (2007).
67. Oganesyan, E. T., Nersesyan, Z. M. & Parkhomenko, A. Y. Chemical composition of the above-ground part of Coriandrum sativum. *Pharm. Chem. J.* **41**, 149–153 (2007).
68. Ceska, O. et al. Coriandrin, a novel highly photoactive compound isolated from Coriandrum sativum. *Phytochemistry* **27**, 2083–2087 (1988).
69. Baba, K., Xiao, Y.-Q., Taniguchi, M., Ohishi, H. & Kozawa, M. Isocoumarins from Coriandrum sativum. *Phytochemistry* **30**, 4143–4146 (1991).
70. Taniguchi, M., Yanai, M. & Xiao, Y. Q. Kido, T.-i. & Baba, K. Three isocoumarins from Coriandrum sativum. *Phytochemistry* **42**, 843–846 (1996).
71. Tan, L. et al. Simultaneous determination of scopoletin and scoparone in Apium graveolens L. by HPLC. *Cent. South Pharm.* **8**, 503–506 (2010).
72. Garg, S. K., Sharma, N. D. & Gupta, S. R. A new Dihydrofurocoumarin from Apium graveolens. *Planta Med* **43**, 306–308 (1981).
73. Garg, S. K., Gupta, S. R. & Sharma, N. D. Coumarins from Apium graveolens seeds. *Phytochemistry* **18**, 1580–1581 (1979).
74. Jiang, H. et al. Qualitative Analysis of Multiple Cumarins in Angelicae Sinensis Radix Based on HPLC-Q-TOF-MS/MS. *Chin. J. Exp. Tradit. Med Form.* **25**, 157–162 (2019).
75. Soine, T. O., Abu-Shady, H. & Digangi, F. E. A note on the isolation of bergapten and imperatorin from the fruits of Pastinaca sativa L. *J. Am. Pharm. Assoc. Am. Pharm. Assoc.* **45**, 426–427 (1956).
76. Ekiert, H. & Kisiel, W. Isolation of furanocoumarins from Pastinaca sativa L. callus culture. *Acta Soc. Bot. Pol.* **69**, 193–195 (2000).
77. Ekiert, H. & Gomolka, E. Furanocoumarins in Pastinaca sativa L. in vitro culture. *Pharmazie* **55**, 618–620 (2000).
78. Lombaert, G. A., Siemens, K. H., Pellaers, P., Mankotia, M. & Ng, W. Furanocoumarins in celery and parsnips: method and multiyear Canadian survey. *J. Aoac Int.* **84**, 1135–1143 (2001).
79. Xie, N. et al. Determination of 9 Chemical Components of Coumarins in Angelica Dahurica by High Performance Liquid Chromatography and its Multivariate Statistical Analysis. *Phy Test. Chem. Anal. Part B: Chem. Anal.* **58**, 657–663 (2022).
80. Pan, M. et al. Study on separation of coumarins in Angelica dahurica by supercritical fluid chromatography. *Chin. J. Pharm. Anal.* **43**, 1120–1128 (2023).
81. Zou, J., Su, W., Pan, Y. & Cui, J. Chemical Components and Pharmacological Action for Angelica dahurica Sinensis and Predictive Analysis on its Q-marker. *Mod Trad Chin Med Materia Medica-World. Sci. Technol.* **25**, 2535–2548 (2023).
82. Jia, Y. et al. Study on HPLC fingerprint chromatogram and quantitation methed of seven coumarins in Cnidii Fructus decoction pieces. *Chin. J. Pharm. Anal.* **42**, 1371–1380 (2022).
83. Shin, E. et al. Antifibrotic activity of coumarins from Cnidium monnieri fruits in HSC-T6 hepatic stellate cells. *J. Nat. Med.* **65**, 370–374 (2011).
84. Cai, J. et al. Variation and regularity of coumarin constituents in Fructus Cnidii collected from different regions of China. *Acta Pharm. Sin.* **34**, 767–771 (1999).
85. Zhao, B., Yang, X., Yang, X. & Zhang, L. Chemical constituents of roots of Saposhnikovia divaricata. *China J. Chin. Mater. Med.* **35**, 1569–1572 (2010).

Acknowledgements

We thank reviewers for their constructive suggestions in improving our manuscript, and Dr. Zhen Li at Ghent University for discussion. This work was supported by grants from the ability establishment of sustainable use for valuable Chinese medicine resources (2060302), the University Synergy Innovation Program of Anhui Province (GXXT-2023-070), the Fundamental Research Funds for the Central Universities (KYCXJC2022003 and 2632024TD04), the open foundation of Shaanxi University of Chinese Medicine state key laboratory of R&D of Characteristic Qin Medicine Resources (SUCM-QM202202), the fund of Traditional Chinese Medicine Institute of Anhui Dabie Mountain (TCMADM-

2023-18), and the open research fund of Yunnan characteristic plant extraction laboratory (YKKF2023002).

Author contributions

J.Y.X., F.Q. and Y.Z. conceived the study. Y.Z. and H.T. collected plant samples. X.C.H. conducted phylogenetic and WGD analyses, X.C.H., X.W., S.H. and J.Y.W. and J.Y.X. conducted evolutionary analyses of C2′H, PT and cyclase genes. H.T., Y.H., D.X. and Y.Z. conducted gene functional experiments. J.Y.X. and Y.Z. drafted the manuscript. All authors read the manuscript and participated in the revision of the manuscript. All authors approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-51285-x.

Correspondence and requests for materials should be addressed to Fei Qiao, Jia-Yu Xue or Yucheng Zhao.

Peer review information *Nature Communications* thanks COLIN KIM, and the other, anonymous, reviewer for their contribution to the peer review of this work. A peer review file is available.

[1]College of Horticulture, Bioinformatics Center, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing 210095 Jiangsu, China. [2]Department of Resources Science of Traditional Chinese Medicines, School of Traditional Chinese Pharmacy, and State Key Laboratory of Natural Medicines, China Pharmaceutical University, Nanjing 210009 Jiangsu, China. [3]College of Bioscience and Biotechnology, Hunan Agricultural University, Changsha 410128 Hunan, China. [4]School of Pharmacy, Shaanxi University of Chinese Medicine, Xi'an 712046 Shaanxi, China. [5]National Key Laboratory for Tropical Crop Breeding, Sanya 572024 Hainan, China. [6]Tropical Crops Genetic Resources Institute, Chinese Academy of Tropical Agricultural Sciences, Haikou 571101 Hainan, China. [7]Medical Botanical Garden, China Pharmaceutical University, Nanjing 210009 Jiangsu, China. [8]These authors contributed equally: Xin-Cheng Huang, Huanying Tang, Xuefen Wei. ✉e-mail: fei.qiao@catas.cn; xuejy@njau.edu.cn; zhaoyucheng1986@126.com