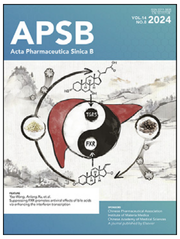
Acta Pharmaceutica Sinica B 2024;14(8):3760−3773

Chinese Pharmaceutical Association
Institute of Materia Medica, Chinese Academy of Medical Sciences

Acta Pharmaceutica Sinica B

www.elsevier.com/locate/apsb
www.sciencedirect.com



ORIGINAL ARTICLE

The chromosome-scale assembly of the *Notopterygium incisum* genome provides insight into the structural diversity of coumarins

Qien Li [a,†], Yiqun Dai [b,c,†], Xin-Cheng Huang [d,†], Lanlan Sun [c], Kaixuan Wang [c], Xiao Guo [a], Dingqiao Xu [e], Digao Wan [a], Latai An [a], Zixuan Wang [c], Huanying Tang [c], Qi Qi [c], Huihui Zeng [c], Minjian Qin [c,*], Jia-Yu Xue [d,*], Yucheng Zhao [c,*]

[a]*Tibetan Medicine Research Center of Qinghai University, Tibetan Medical College, Qinghai University, Xining 810016, China*
[b]*School of Pharmacy, Bengbu Medical University, Bengbu 233030, China*
[c]*Department of Resources Science of Traditional Chinese Medicines, School of Traditional Chinese Pharmacy, China Pharmaceutical University, Nanjing 210009, China*
[d]*College of Horticulture, Bioinformatics Center, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing 210095, China*
[e]*Key Laboratory of Shaanxi Administration of Traditional Chinese Medicine for TCM Compatibility, Shaanxi University of Chinese Medicine, Xi'an 712046, China*

Received 7 December 2023; received in revised form 17 March 2024; accepted 3 April 2024

KEY WORDS

Notopterygium incisum;
Parallel evolution;
Biosynthesis;
Pyranocoumarins;
Furanocoumarins

Abstract Coumarins, derived from the phenylpropanoid pathway, represent one of the primary metabolites found in angiosperms. The alignment of the tetrahydropyran (THP) and tetrahydrofuran (THF) rings with the lactone structure results in the formation of at least four types of complex coumarins. However, the mechanisms underlying the structural diversity of coumarin remain poorly understood. Here, we report the chromosome-level genome assembly of *Notopterygium incisum*, spanning 1.64 Gb, with a contig N50 value of 22.7 Mb and 60,021 annotated protein-coding genes. Additionally, we identified the key enzymes responsible for shaping the structural diversity of coumarins, including two *p*-coumaroyl CoA 2′-hydroxylases crucial for simple coumarins basic skeleton architecture, two UbiA prenyltransferases responsible for angular or linear coumarins biosynthesis,

*Corresponding authors.
E-mail addresses: minjianqin@163.com (Minjian Qin), xuejy@njau.edu.cn (Jia-Yu Xue), zhaoyucheng1986@126.com (Yucheng Zhao).

†These authors made equal contributions to this work.

Peer review under the responsibility of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences.

https://doi.org/10.1016/j.apsb.2024.04.005

and five CYP736 cyclases involved in THP and THF ring formation. Notably, two bifunctional enzymes capable of catalyzing both demethylsuberosin and osthenol were identified for the first time. Evolutionary analysis implies that tandem and ectopic duplications of the CYP736 subfamily, specifically arising in the Apiaceae, contributed to the structural diversity of coumarins in *N. incisum*. Conclusively, this study proposes a parallel evolution scenario for the complex coumarin biosynthetic pathway among different angiosperms and provides essential synthetic biology elements for the heterologous industrial production of coumarins.

1. Introduction

Coumarins are one of the primary secondary metabolites that are ubiquitously found in higher plants[1]. The 2*H*-1-benzopyran-2-one core structure and its constituents tetrahydropyran (THP) and tetrahydrofuran (THF) are common structural units in both natural and synthetic bioactive molecules[2]. In particular, the alignment of the THP and THF rings into the lactone core structure creates at least four subgroups of the basic THP and THF coumarin skeleton. This leads to the formation of thousands of coumarin compounds, including pyranocoumarins and furanocoumarins. Coumarins have been shown to exhibit diverse physiological and medical bioactivities, including defense against fungal infections, phytoalexins or elicitors, anticancer, anti-inflammatory, antioxidant, and calcium channel-blocking properties[2-5]. For instance, furanocoumarins, xanthotoxin, and bergapten are commonly used clinical drugs approved by the US Food and Drug Administration (FDA) for the treatment of skin diseases (Supporting Information Fig. S1)[3]. These diverse bioactivities are largely attributed to their variable structures. However, little is known regarding the biosynthetic mechanisms underlying the structural diversity of these coumarins.

Coumarins, derived from the phenylpropanoid pathway, have a conserved and widespread initial upstream synthesis in the plant kingdom (Fig. 1). It can produce key precursors, such as cinnamic acid, *p*-coumaric acid, and *p*-coumaroyl-CoA, using phenylalanine ammonia lyase (*PAL*), cinnamate 4-hydroxylase (*C4H*) and 4-coumarate: coenzyme A ligase (*4CL*)[1]. Different enzymes are involved in branching *p*-coumaroyl-CoA to form different metabolites such as lignin, flavonoids, and stilbenes[1,3]. Among the branching enzymes, two types of 2-oxoglutarate-dependent dioxygenase (*2-OGD*) family proteins, *p*-coumaroyl CoA 2′-hydroxylase (*C2′H*) and feruloyl CoA 6′-hydroxylase (*F6′H*), are known to be involved in simple coumarins skeleton (umbelliferone or scopoletin) formation (Fig. 1)[6,7]. Additionally, four *C2′H*s/*F6′H*s have been cloned from phylogenetically distinct angiosperm lineages, including the Brassicaceae, Convolvulaceae, Rutaceae, and Apiaceae[3,6-8]. Subsequently, the substitution of umbelliferone with isopentenyl groups at six or eight positions, facilitated by umbelliferone 6/8-prenyltransferase (*PT*), leads to the formation of demethylsuberosin (DMS) or osthenol, respectively[9-11]. This step is the entry point for complex coumarins synthesis, determining the formation of angular or linear coumarins. However, identified *PT*s are restricted to members of the Moraceae, Rutaceae, and Apiaceae[9-11]. This phenomenon is consistent with the fact that complex coumarins are mainly found in Moraceae, Rutaceae, and Apiaceae, whereas simple coumarins are widely distributed across at least 75 families[1,12]. Finally, the cyclization of demethylsuberosin or osthenol forms the corresponding THP and THF rings, a crucial step in shaping the basic skeleton of complex coumarins. Previous studies have shown that this cyclization is catalyzed by a CYP450 protein, with a specific CYP76F112 protein from Moraceae involved in linear furanocoumarin cyclization[1,13,14]. Recently, we found that two CYP736 family proteins are also involved in linear/angular THP and THF ring formation[15]. However, the processes by which plants generate the structural diversity of coumarins and the reasons behind the use of different enzymatic elements by different angiosperms for synthesizing the same compound remain unclear. Subsequently, structural modification steps such as hydroxylation, methoxylation, and glycosylation occur to complete the complex coumarin post-modification[4,16]. However, to date, only one type of hydroxylating or methoxylating enzyme has been identified[4,17]. Hence, the biosynthetic mechanism of coumarins, as well as the evolutionary mechanism of expanded structural diversity, remains to be elucidated.

Notopterygium incisum Ting ex H. T. Chang is an important traditional Chinese and Tibetan medicinal plant with a lengthy historical background (Fig. 2A and Supporting Information Fig. S2). It is mainly distributed in the plateau areas of Qinghai, Tibet, Shaanxi, Sichuan, Gansu, and other limited regions of China, at altitudes of 2000−4000 m in China (Supporting Information Fig. S3). Considering the abundant coumarin content and structural diversity of *N. incisum*[18], we assembled a high-quality genome of this plant to explore the mechanisms underlying the biosynthesis and evolution of the coumarin pathway. We identified all the key genes involved in the biosynthesis of simple and complex coumarin skeletons by combining genomic, metabolomic, and transcriptomic analyses. Thus, this study provides a basis for the heterologous industrial production of complex coumarins. Furthermore, it provides detailed genetic information on this species and explores the potential factors contributing to the structural diversity of coumarins in *N. incisum* and angiosperms.

2. Materials and methods

2.1. Experimental materials

N. incisum, *Notopterygium franchetii*, and cultivated *N. incisum* were collected in Guide of Qinghai Province (36°21′34″N, 101°32′6″E), and were confirmed *via* DNA molecular identification. A voucher specimen (No. CPUZYC2021013) was deposited at the Botanic Garden of China Pharmaceutical University, Nanjing, China. Genomic DNA was extracted from young leaves of *N. incisum* for sequencing. Leaves, fruits, roots, and stems of *N. incisum*, along with roots of *N. franchetii* and cultivated

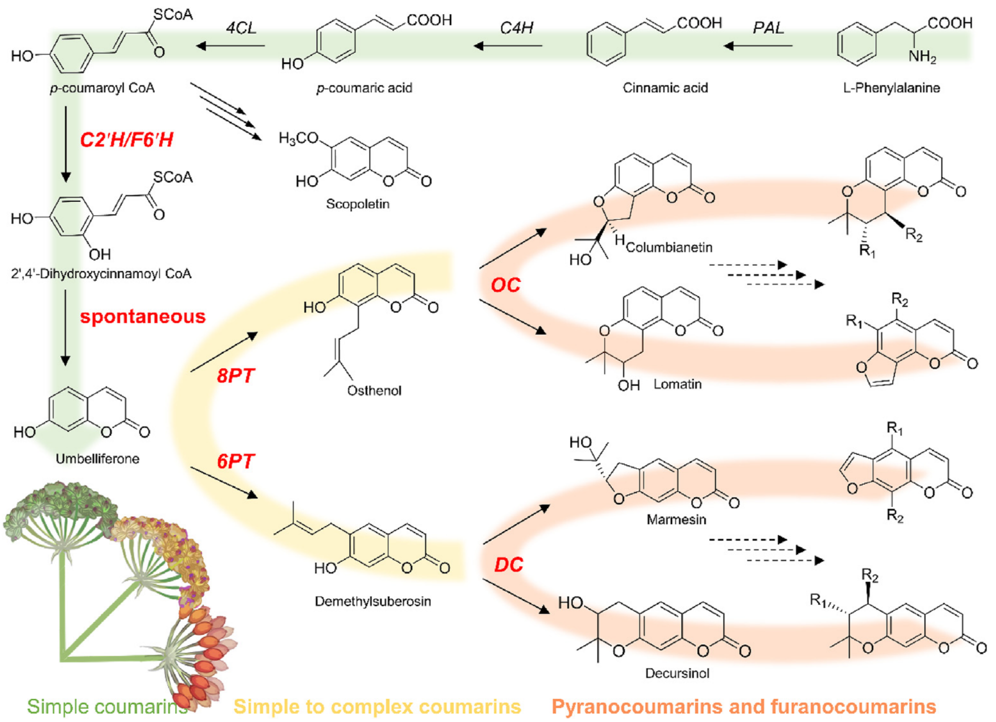


Figure 1 Proposed coumarins biosynthetic pathway. The proposed coumarins biosynthetic pathway comprises three phases: simple coumarins, simple coumarins to complex coumarins, and complex coumarins such as pyranocoumarins and furanocoumarins. Each phase was marked in different colors. Abbreviation: *PAL*, phenylalanine ammonia lyase; *C4H*, cinnamate 4-hydroxylase; *4CL*, 4-coumarate: coenzyme A ligase; *C2′H*, *p*-coumaroyl CoA 2′-hydroxylase; *F6′H*, feruloyl CoA 6′-hydroxylase; *6 PT*, umbelliferone 6-prenyltransferase; *8 PT*, umbelliferone 8-prenyltransferase; *DC*, demethylsuberosin cyclase; *OC*, osthenol cyclase. Multiple arrows represent multiple known or unknown steps.

N. incisum, were utilized for transcriptome sequencing. Solvents used for extraction and biochemical analysis were chromatographic grade. Compounds used for the enzymatic reactions and the antibiotics used for microorganism culture were purchased from Herbest (Baoji, China) and Sigma−Aldrich (Shanghai, China). Enzymes, kits used for gene fragment amplification, and vectors for *in vitro* expression were purchased from TransGen Biotech (Beijing, China).

2.2. *Genome size estimation, genome sequencing, assembly and annotation*

The genome size was estimated based on the *k*-mer distribution analysis using 72.5 Gb Illumina paired-end short reads. The *k*-mer frequency was calculated using Jellyfish 2.2.6 with *k*-mer size 19[19]. The extracted DNA of *N. incisum* was sequenced on a PacBio Sequel II platform to generate 38.93 Gb HiFi long-read data with an average length of 18.42 kb. After the initial quality control by SMRTLink v8.0, the *de novo* assembly of *N. incisum* nuclear genome firstly obtained a contig level assembly using hifiasm (v0.16.1)[20]. HiC technology was adopted to assist the scaffolding of *N. incisum* contigs, and in total of 387.7 Gb HiC data were used after FastQC (v0.11.9) filtering. At last, the polished and haplotigs-purged contigs were mostly anchored to 11 pseudomolecules using the AllHiC (v0.9.8) pipeline[20].

Protein-coding genes in the *N. incisum* genome were annotated by a combination of transcriptome-based, homology-based, and ab initio prediction approaches. Firstly, repetitive sequence annotation was combined with homology prediction based on the Repbase Library (http://www.girinst.org/repbase) and *de novo* prediction based on self-sequence alignment. In these two methods, RepeatModeler (version: open-1.0.11, http://repeatmasker.org/RepeatModeler/) and RepeatMasker (version: open-4.0.9, http://repeatmasker.org/RepeatMasker/) were employed to construct a *de novo* repeat sequence database and search repetitive sequences from genome and Repbase Library, respectively. After masking the repetitive sequences, the gene models were predicted using GeneMarkS-T (v5.1) with the transcripts assembled from RNA-Seq reads. RNA-Seq reads were mapped to the masked genome by HISAT2 (v2.0.4)[21] and assembled into transcripts using StringTie (v1.2.3)[22]. Then, the homology-based method employed GeMoMa (v1.7)[23] with reference gene model from five species, including *Daucus carota*, *Coriandrum sativum*, *Apium graveolens*, *Panax notoginseng,* and *Panax ginseng*. After that, *de novo* gene models were predicted using Augustus (v2.4) and SNAP (2006-07-28)[24]. Finally, gene models from all three methods were integrated using the EVidenceModeler (v1.1.1)[25] and updated by PASA (v2.0.2)[26].

Functional annotation of protein-coding genes was conducted by blast using Diamond (v0.9.24)[27] against the databases of NCBI NR (202009), SwissProt/TrEMBL (202005), and EggNOG (v5.0)[28]. Protein domains were searched using HMMER (v3.1) under Pfam (v33.1) database[29]. Terminal repeat retrotransposons were identified using LTRharvest and LTR_finder[30,31]. Tandem repeats were annotated with TRF (v4.09.1, https://github.com/Benson-Genomics-Lab/TRF) and MISA (v2.1, https://webblast.ipk-gatersleben.de/misa/).

2.3. *Whole-genome duplication (WGD) and evolutionary rate correction*

Identification of WGD events rely on the construction of *K*s-based age distributions for all paralogues (paranome) of *N. incisum*, *Vitis*

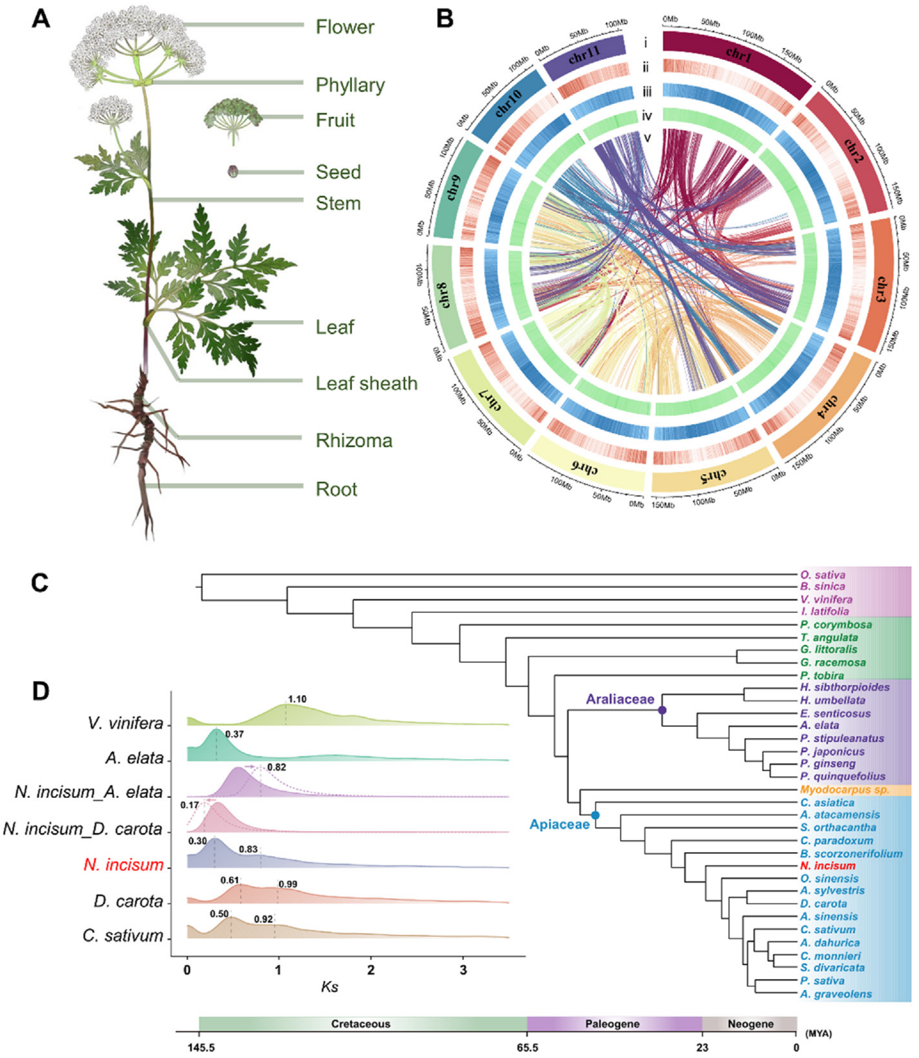


Figure 2 Genome assembly, phylogenetic tree reconstruction, and whole genome duplication analysis of *N. incisum.* (A) Morphology of *N. incisum.* (B) Genomic features of *N. incisum*. The circus plot from the outer to the inner circle represents chromosome-scale pesudochromosomes (Chr1-Chr11) (i), gene density (ii), the density of repeat sequences (iii), GC content (iv), each linking line in the center of the circus plot indicates a pair of homologous genes (v). (C) Phylogenetic relationships and divergence times between *N. incisum* and other 33 species. The species tree was generated using 1730 low-copy orthologous genes based on the maximum-likelihood method with high support. (D) *K*s distribution of anchored paralogous gene pairs and WGD events in *N. incisum*. The *K*s distribution of *N. incisum* shows two peaks, one at approximately 0.30 (α-WGD) and another at approximately 0.83 (β-WGD). The dotted lines indicate the *K*s distribution peaks of intergenomic after the evolutionary rate was corrected. The *K*s peaks representing all WGD and speciation events were labeled.

vinifera, and other three species in Apiales by using wgd (v1.1)[32]. In the WGD pipeline, coding nucleotide sequences were translated to peptide sequences first and then according to the requirements were fed into BLASTP (v2.13.0)[33] to all-*versus*-all blast with *E*-value set as 1×10^{-10}. Subsequently, MCL (v14-137) and MUSCLE (v5.1) were called to obtain a multiple sequence alignment (MSA) with protein level from each paralogous gene family[34]. After back-translating this MSA to codon alignment, the CODEML package within PAML (v4.9)[35] was employed to acquire the maximum likelihood estimate of *K*s values under the default control file and eventually build the collinear gene pairs (anchor pairs) *K*s-based age distribution by combining the weight of each anchor pairs without outliers using FastTree (v2.1.11)[36] and i-ADHoRe (v3.0)[37]. For orthologous *K*s-based age distributions, the process of MCL clustering was superseded as reciprocal best hits (RBH) searching to identify orthologous gene pairs and only one-versus-one orthologues was inferred. In addition, the difference in evolutionary rate among different species was corrected by using Ksrates (v1.1.1)[38].

2.4. *Phylogenetic tree construction and molecular dating*

To fully elucidate the phylogenetic relationships within the Apiales order, we selected 34 reported angiosperm genomes and transcriptomes (Supporting Information Table S1) to reconstruct a phylogenetic tree. Orthofinder (v2.5.5)[39] was employed to classify orthogroups based on the peptide sequence dataset from these 34 species. We filtered out 1730 low-copy orthologous gene families (Copy number <3, Coverage >90%), among which, only the longest copy of each species in the 1730 families was chosen as the refined 1730 low-copy orthologous genes subject to phylogenetic analyses. Using a de-redundancy script, redundant copies in the orthogroups were removed and fed into Gblock (v0.91b) to obtain conserved sequences. Subsequently, all conserved sequences were concatenated into a supergene. Before the phylogenetic tree construction, FastTree (v2.1.11) was used to evaluate the data matrix and select the optimal model. Phylogenetic analysis of the dataset was then performed using FastTree (v2.1.11) under the JTT model[36]. The MCMCtree of the PAML (v4.9)[35] package was used to estimate the divergence time of Apiales species based on tree topology derived from the 1730 concatenated genes and 15 fossil-based age calibrations (Supporting Information Table S2). The GTR model was used as the peptide replacement model. Posterior distributions of node ages were estimated using Markov chain Monte Carlo sampling, with samples drawn every 10 steps over 100,000 steps, followed by a burn-in of 400,000 steps[40].

2.5. *Metabolomic analysis*

A total of 42 samples were selected for metabolomic analysis, including seven biological repeats each from the roots, stems, leaves, and fruits of *N. incisum*, and the roots of *N. franchetii* and cultivated *N. incisum*. The samples were extracted with prechilled 80% methanol, and the extracts were subjected to ultra-high performance liquid chromatography with quadrupole time-of-flight mass spectrometry (UPLC−Q-TOF-MS) analysis. Chromatographic separation of the compounds was performed using an ACQUITY UPLC system (Waters, Milford, MA, USA). A C18 reversed-phase column (50 mm × 2.1 mm, 1.5 μm, Thermo Scientific, USA) was used for UPLC analysis with 30 °C column temperature. The gradient contains 0.1% formic acid (*v*/*v*, A) and acetonitrile (B) with the gradient as follows: 0 min, 10% B; 3 min, 15% B; 8 min, 38% B; 12 min, 50% B; 16 min, 95% B. The flow rate was 0.4 mL/min. The MS analysis was performed using a Synapt G2-Si Q-TOF (Waters MS Technologies, Manchester, UK). Data acquisition was performed using MassLynx V4.2 software, and subsequent data processing, including background noise removal, normalization using a reference sample, retention time correction, and peak alignment, was carried out using Progenesis QI V2.0 (Waters Corporation, Milford, MA, USA). Metabolites were annotated using public databases such as METLIN (https://metlin.scripps.edu/index.php), Lipidmaps database (http://www.lipid_maps.org/), KEGG database (http://www.genome.jp/kegg/), and HMDB database (http://www.hmdb.ca/). SIMCA 14.1 software (Umetrics, Umea, Sweden) was used for principal component analysis (PCA). Metabolite identification and quantification were performed as in our previous report[41].

2.6. *Screening of the genes involved in coumarin biosynthesis*

To screen the candidate genes involved in simple coumarin skeleton formation, we screened the functional annotation information of the genome and selected the protein sequences of the identified genes in Apiaceae plants. These sequences were then subjected to BLASTPv2.11.0 (*E*-value < 1e−5) analysis[33]. The PT of aromatic compounds belongs to the UbiA family and clusters with VTE2-1, responsible for vitamin synthesis. We used HMMER (version 3.1b2) to scan all the predicted *PT* genes in *N. incisum*, retraining those containing the ''UbiA prenyltransferase family'' (PF01040) domain. Candidate *PT*s were screened through phylogenetic analysis along with *PT* genes from *Arabidopsis thaliana* and protein sequences with verified functions in other plants. Genome, transcriptome, and metabolome data from *N. incisum* were utilized for constructing a transcriptome-metabolome co-expression matrix. A correlation network was then developed based on the Pearson correlation coefficient algorithm. Gene co-expression theory and Weighted correlation network Analysis (WGCNA)[32] were employed to analyze the candidate cyclase genes, which may have similar expression patterns to that of *PT* or *C2′H*.

2.7. *Heterologous expression in E. coli and activity assay*

The candidate *C2′H* and *PT* genes were cloned from the cDNA of *N. incisum* and then inserted into pET28a (between *BamH* I and *EcoR* I restriction sites) and pETDuet-1 (between *BamH* I and *Sac* I restriction sites) vectors, respectively[42]. The successfully sequenced recombinant plasmid was transformed into *E. coli* BL21 (DE3) competent cells. The cells were then cultured in Luria-Bertani (LB) medium containing kanamycin (50 mg/L) and ampicillin (100 mg/L), respectively. Upon induction with IPTG (0.5 mmol/L), the cells were harvested and ultrasonically lysed for activity testing or protein purification. Owing to the membrane protein characteristics and insolubility of PT in *E. coli*, the crude enzyme solution was used directly for the enzyme activity test. In contrast, C2′H was expressed in soluble form, allowing successful purification using Ni-NTA resin and further purification using the protein purification system SDL-030-F2 (SePure Instruments Co., Ltd., Suzhou, China) to obtain pure recombinant protein for enzyme activity test. All *in vitro* enzymatic activity assays were carried out on a shaking incubator (220 rpm) at 30 °C for 2 h. For the PT activity test, the reaction system contained 50 μL crude proteins protein, 200 μmol/L umbelliferone, 100 μmol/L DMAPP, and 200 μmol/L $MgCl_2$ in 200 μL of 100 mmol/L Tris-HCl (pH 7.5). For C2′H activity test, the enzyme reaction (200 μL) contained 10 μg of purified protein, 100 mmol/L Tris-HCl (pH 6.5), 0.5 mmol/L $FeSO_4$, 5 mmol/L sodium ascorbate, 5 mmol/L 2-oxoglutarate (2OG), and 1 mmol/L *p*-coumaroyl CoA or feruloyl-CoA. The reaction samples were extracted with ethyl acetate and dissolved in methanol for HPLC and LC−MS analysis.

2.8. *Functional expression of NiPTs in Nicotiana benthamiana*

Further functional characterization of *NiPTs* was achieved by transient expression in *N. benthamiana.* The open reading frames (ORFs) of *NiPT*s were introduced into the pEAQ binary plasmid through *Age*I and *Xho*I restriction enzyme sites. The correctly sequenced recombinant plasmids were transformed into *A. tumefaciens* strain GV3101 using the freeze-thaw method. The positive transformants were selected on selective LB agar plates (50 μg/mL kanamycin and 50 μg/mL rifampicin) at 28 °C. Positive transformants were inoculated into 10 mL of liquid LB cultures, which were shaken for 1 day at 28 °C. After that, the cells were pelleted through centrifugation at 6000 rpm for 5 min. Then the cell pellet was resuspended in Agrobacterium induction medium (10 mmol/L MES, 10 mmol/L $MgCl_2$, 100 μmol/L acetosyringone, pH 5.8), and incubated at 25 °C for 1 h. After measuring the concentration of the cell suspension by determining its optical density (OD_{600}), we used a needleless syringe to infiltrate these bacterial suspensions into the underside of 4−5-week-old *N. benthamiana* leaves. Three days later, inject 100 μL of umbelliferone (100 μmol/L PBS solution) into the infected area of the leaf. One day later, collect the substrate-infiltrated leaves and extract them with methanol for LC−MS analysis. *N. benthamiana* infiltrated with *A. tumefaciens* containing an empty vector was used as a negative control. Each experiment includes at least three plants as parallel experiments.

2.9. Heterologous expression in *Saccharomyces cerevisiae* and activity assay

We used *S. cerevisiae* WAT11 integrated with *A. thaliana* CYP450 reductase (AtCPR) to screen candidate CYP450 proteins. The pYES2-Ura vector carrying the CYP450 genes was transformed into *S. cerevisiae* strain WAT11, and cultured in SD medium at 30 °C until OD_{600} reached 0.8−1.0. The cells were then collected and washed with sterile water. Protein expression was induced using SG medium containing galactose after washing the cells three times. The induced yeast cells were collected and resuspended in Tris-HCl buffer for whole-cell catalysis. All the reactions were performed in a shaking incubator (220 rpm) at 30 °C for 2 h. The reaction system contained 500 μmol/L NADPH and 1 mmol/L substrates in 200 μL of 100 mmol/L Tris-HCl (pH 7.5). The reaction samples were extracted with ethyl acetate and dissolved in methanol for HPLC and LC−MS analysis.

2.10. HPLC and LC−MS analysis

A Shimadzu LC-2010C system (Shimadzu, Japan) was used for HPLC analysis. Samples were separated on an Hedera ODS-2 C18 column (4.6 mm × 250 mm, 5 μm). The mobile phase consisted of water containing 0.1% formic acid (*v*/*v*, A) and methanol (B). The flow rate was 0.5 mL/min, and the column temperature was 30 °C. For all enzymatic products, the detection wavelength was 340 nm. The gradient elution program for detecting C2′H activity was as follows: 0 min, 10% B; 5 min, 15% B; 15 min, 60% B; 22 min, 60% B. The gradient elution program for detecting PT activity was as follows: 0 min, 40% B; 5 min, 65% B; 15 min, 75% B; 22 min, 95% B. The gradient elution program for detecting cyclase activity was as follows: 0 min, 30% B; 5 min, 65% B; 12 min, 95% B; 14 min, 95% B; 16 min, 30%; 25 min, 30%. MS analysis was performed on Agilent 6545 LC/Q-TOF equipped with a heated ESI source (Agilent Technologies, USA). The parameters were as follows: MS survey scan of 100−2000 Da; sheath gas temperature, 350 °C; ion spray voltage, 3500 V; and collision energy, 44 V. MassHunter Qualitative Analysis software was used for observing and processing LC−MS data.

3. Results

3.1. Genome sequencing, assembly and annotation

To elucidate the biosynthetic pathway of complex coumarins in *N. incisum*, we first assembled the nuclear genome of *N. incisum*, combining approximately 72.1 Gb Illumina short reads, approximately 38.93 Gb PacBio HiFi reads, and approximately 387.71 Gb HiC data (Supporting Information Table S3). After deredundancy of the two haplotypes, the total length of the final assembled genome was 1.64 Gb, slightly smaller than the estimated 1.75 Gb using *k*-mer analysis (Supporting Information Fig. S4 and Table S4). Collectively, 97.94% of the contigs were anchored to 11 pseudochromosomes of a haplotype (Fig. 2B, Supporting Information Fig. S5, and Tables S5 and S6), corresponding to the number of chromosomes recorded in The Chromosome Counts Database (CCDB, https://ccdb.tau.ac.il/) ($2n = 2x = 22$). To assess the integrity of the assembly, Benchmarking Universal Single-Copy Orthologs (BUSCO) were employed to blast against our genome, and the assessment indicated 98.7% completeness (Supporting Information Tables S7 and S8) of the embryophyte_odb10 genes recovered for assembly. Using Ab initio, homology-based, and transcriptome-assisted annotation approaches, we annotated 60,021 protein-coding genes. These genes were compared with protein sequences in seven databases, including KEGG, Pathway, Nr, UniProt, GO, Pfam, and InterPro, with 93.03% of the genes assigned to putative functional annotations (Supporting Information Table S9). Furthermore, the annotation also demonstrated that *N. incisum* contained up to 75.79% of repetitive sequences. Among these, long terminal repeat retrotransposons were the most prevalent type of transposable elements, accounting for 44.86% of the genome, including 12.35% *Gypsy* and 32.38% *Copia* retroelements (Supporting Information Table S10).

3.2. Phylogenetic position and polyploidy events of *N. incisum*

To precisely determine the evolutionary relationship between *N. incisum* and other species in Apiaceae, 1730 refined low-copy orthologous genes from *N. incisum* and 33 other angiosperms (Table S1) were used to construct phylogenetic relationships based on the maximum-likelihood (ML) method (Fig. 2C). *N. incisum* was resolved as an early-diverging lineage within Apioideae, emerging subsequently to the divergence of *Chamaesium paradoxum* and *Bupleurum scorzonerifolium*. Additionally, *N. incisum* was observed to be sister to the remaining Apioideae taxa. Using this resolved phylogeny and 15 fossil calibrations (Table S2), we inferred that the Apiaceae crown group of Apiaceae originated approximately 49 million years ago (MYA), and *N. incisum* diverged approximately 22 MYA (Supporting Information Table S11), probably representing the emergence time of the *Physospermopsis* clade[43].

Whole genome duplication events are considered the pivotal drivers of genome evolution[44]. To identify WGDs that occurred in the *N. incisum* genome, we performed a comprehensive analysis using other Apiales species and *V. vinifera* as references. By calculating the synonymous substitution rate (*K*s) of paralogous genes located in collinear genomic block anchor pairs within *N. incisum* and eight other reported genomes in Apiales, we observed distinct *K*s density distributions in *N. incisum*, with two prominent peaks at 0.37 and 0.83 (Fig. 2D). Remarkably, these peaks were consistently observed in all other Apiaceae taxa (Supporting Information Figs. S6−S8). Although different Apiaceae plants exhibited these two peaks at slightly different positions owing to differential evolutionary rates, the adjusted species divergence peak (0.17) confirmed that speciation within Apiaceae occurred subsequent to the occurrence of the two WGDs. The results support the notion that Apiaceae plants underwent two additional rounds of shared whole genome duplications after the eudicot-shared whole-genome triplication event (γ-WGT), referred to as α-WGD and β-WGD, respectively[45-48]. Nevertheless, conflicting views exist regarding the occurrence time and taxa of the two WGD events, specifically whether they occurred in the common ancestor of Apiaceae and Araliaceae, or were specific to Apiaceae[45-48]. As our analysis shows, the *K*s peak representing the WGD in the Araliaceae plant *Aralia elata* (0.37) is much smaller than the peak corresponding to the divergence of Apiaceae and Araliaceae (0.83), thus indicating that Apiaceae and Araliaceae diverged earlier than all WGDs detected in the two families, and the two families have no shared WGDs since the γ-WGT.

Therefore, the two WGDs detected in *N. incisum* are likely specific to Apiaceae.

3.3. Metabolite profile of the primary compounds in N. incisum

Global non-targeted metabolomics was initially conducted to characterize the metabolite profiles of coumarins in *N. incisum* and *N. franchetii,* the two main sources of the Chinese pharmacopoeia. The samples included the leaves, fruits, roots, and stems of *N. incisum* as well as the roots of *N. franchetii* and cultivated *N. incisum* (Supporting Information Fig. S9). PCA analysis of the samples from *N. incisum* indicated that the underground parts (roots) were distinctly from the aerial parts (leaves, fruits, and stems) (Supporting Information Fig. S10). Moreover, roots exhibited more peaks and a higher abundance in total ion chromatography, implying that the roots may have specifically accumulated more coumarins (Supporting Information Fig. S11). Targeted metabolomic analysis was performed to identify the candidate metabolites involved in coumarin biosynthesis in *N. incisum*. Through comparison with reference standards (t_R and MS data) or matching with theoretical data or commercial libraries[41], we identified 39 compounds from all tissues of *N. incisum* (Fig. 3A, Supporting Information Fig. S12 and Table S12), with 20 of them being coumarins. Therefore, we chose nine representative coumarins in *N. incisum*, namely nodakenin, notopterol, cnidilin, isoimperatorin, decursinol, columbianetin, osthenol, bergapten, and psoralen, as marker compounds to investigate their abundance in different groups (Supporting Information Fig. S13). As depicted in Fig. 3B, most of the marker compounds were highly abundant in the roots, which aligns with the fact that traditional Chinese medicine uses *N. incisum* root as a medicinal component[49]. Additionally, the general abundance of marker compounds in wild *N. incisum* was higher than that in the cultivated *N. incisum* (Fig. 3C and Supporting Information Fig. S14). This may explain why *N. incisum* has been successfully cultivated. However, people still seek wild *N. incisum* because of its high quality. We also found that some compounds such as nodakenin and isoimperatorin were highly abundant in *N. franchetii* (Fig. 3C), which may explain why they were chosen as alternatives to *N. incisum*. However, all the detected compounds in cultivated *N. incisum* had a lower content than in wild *N. incisum*, implying that the wild imitation strategy may be a good way to

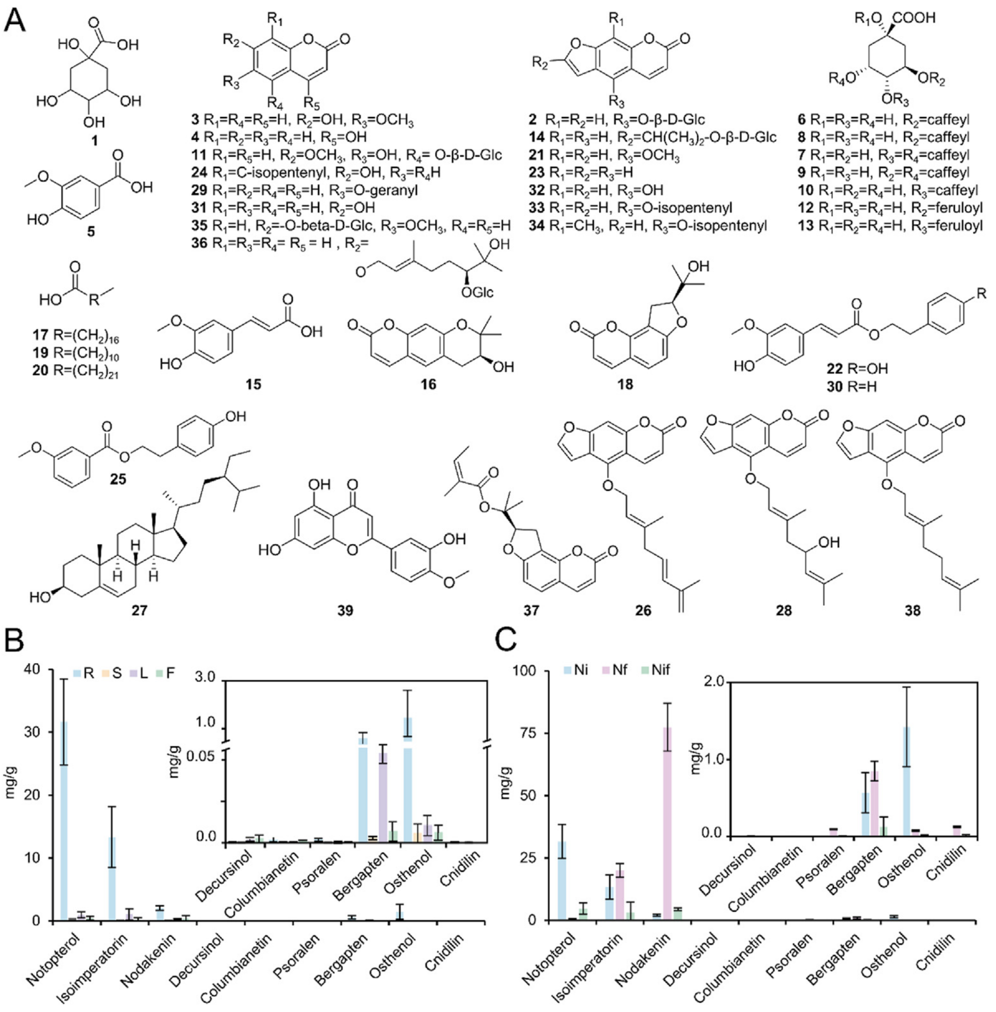


Figure 3 Metabolite profile of the main compounds in *N. incisum.* (A) Chemical structures of the compounds identified in *N. incisum*. (B) Contents of the nine main coumarins in *N. incisum* at different tissues. R: roots, S: stems, L: leaves, F: fruits. (C) Contents of the nine main coumarins in *N. incisum* (Ni), *N. franchetii* (Nf), and cultivated *N. incisum* (Nif). Error bars represent the ±SEM mean of three biologically independent samples ($n = 3$).

obtain high-quality medicinal materials during the habituated culture process (Fig. 3B).

3.4. Identification of the genes involved in simple coumarin biosynthesis

The biosynthetic mechanism of simple coumarins is relatively clear[9,50-53]; hence, we identified 54 genes potentially encoding enzymes involved in simple coumarins biosynthesis. It includes 5 *PAL* genes, one *C4H* gene, 25 *4CL* genes, 12 *C2′H* genes, and 11 *PT* genes (Supporting Information Fig. S15). Given the pivotal roles of 2-OGD and PT enzymes in driving the phenylpropanoid pathway towards coumarin biosynthetic pathway[53], we mainly focused on identifying these two types of genes in *N. incisum*. Based on the metabolite profile of coumarins in *N. incisum*, we speculated that the genes NincChr3G00107490.1 and NincChr11G00498940.1, which were highly expressed in the roots, were most likely involved in coumarin biosynthesis (Fig. 3, and Fig. S15). Additionally, the metabolome-transcriptome correlation analysis indicated that both genes were significantly correlated with isoimperatorin, the main coumarin product of *N. incisum* (Fig. 4A). Therefore, these two genes were selected for functional evaluation. After recombinant protein expression, purification (Supporting Information Fig. S16), and activity tests, only NincChr3G00107490.1 was found to hydroxylate *p*-coumaroyl-CoA to form a simple coumarin skeleton, umbelliferone (Fig. 4B and Supporting Information Fig. S17A). However, when we tested their activities with another potential substrate (feruloyl-CoA) of the 2-OGD family of proteins, NincChr11G00498940.1 displayed corresponding activity and formed scopoletin (Fig. 4C and Fig. S17B). We therefore designated NincChr3G00107490.1 and NincChr11G00498940.1 as *NiC2′H* and *NiF6′H*, respectively. Amino acid sequence alignment of *NiC2′H* and *NiF6′H* with all the identified 2-OGD genes indicated that the two genes contained a highly conserved Fe(II)-binding motif His-X-Asp-XnHis (His233, Asp235 and His291) and the 2-oxoglutarate C5 carboxy group binding motif Arg-X-Ser (Arg301 and Ser303) (Supporting Information Fig. S18)[53]. Phylogenetic analysis indicated that *NiC2′H* and *NiF6′H* belong to the DOXC30-clade as previously described (Supporting Information Fig. S19)[54]. The two genes clustered with our previously characterized *PpC2′H* and displayed a high sequence similarity, implying the orthologous relationship of *C2′H* and functional consistency within Apiaceae[53].

Nearly half of the PTs (6/13) were significantly correlated with eight out of the nine main coumarins in *N. incisum* (Fig. 4A), implying the importance of PTs in coumarin biosynthesis. Because the coding sequences (CDS) of all the identified PTs were limited to approximately 1200 bp, three highly correlated PTs (3/6, NincChr3G00107340.1, NincChr3G00107500.1, and NincChr3G00107320.1) were selected for functional verification (Fig. S15). These genes were cloned into a prokaryotic expression vector to test their activities according to our previously described methods[42]. As shown in Fig. 4D, only NincChr3G00107340.1 and NincChr3G00107500.1 yielded significant peaks at the same retention time as the standard demethylsuberosin, a C6-umbelliferone prenylation product of PT. The products shared the same protonated molecular ion *m/z* with a 231.10, identical to the $[M+H]^+$ molecular weight of the standard demethylsuberosin (Fig. 4E). Therefore, we designated NincChr3G00107340.1 and NincChr3G00107500.1 as *Ni6PT1* and *Ni6PT2*, respectively. Furthermore, we verified *Ni6PT1* and *Ni6PT2* activity in *N. benthamiana*. When umbelliferone was used as the substrate, a new peak was generated and directly compared with the standard samples, indicating the production of DMS (Supporting Information Fig. S20). We observed that all the identified coumarin-specific PTs displayed lower activity, likely attributable to their membrane protein characteristics, especially the presence of at least six transmembrane regions interspersed in the CDS of PTs (Supporting Information Fig. S21)[9-11,42]. *Ni6PT1* and *Ni6PT2* both belong to the Apiaceae-specific UbiA PT clade and differ from the PTs in the Moraceae clade (*FcPT1a* and *FcPT1b*), implying that these two PT groups are derived from distinct ancestors, despite both belonging to the UbiA family (Supporting Information Fig. S22)[11].

3.5. Identification of the genes crucial for complex coumarin biosynthesis

The cyclization of the PT products (demethylsuberosin and osthenol) is crucial for complex coumarin biosynthesis, which eventually leads to the formation of two types of complex coumarins: furanocoumarins and pyranocoumarins. Complex coumarins can be classified into four types based on their linear and angular configurations: linear furanocoumarins, linear pyranocoumarins, angular furanocoumarins, angular pyranocoumarins[1,3]. However, only a few cyclases responsible for this final step have been identified. One is a CYP76 family protein from Moraceae and the other is a CYP736 family protein from Apiaceae[13,15]. Given that *N. incisum* also belongs to Apiaceae, we speculated that the cyclases in *N. incisum* were likely members of the CYP736 proteins, too. Therefore, all 44 CYP736 proteins of *N. incisum* were identified as potential candidates. We first excluded some proteins with short CDS (<1000 bp). Based on metabolite profiles of the main compounds in *N. incisum* (Fig. 3), seven genes with high expression in the roots were selected for functional verification (Fig. 5A). After the activity test, we found that four genes (NincChr3G00106470.1, NincChr3G00121510.1, NincChr6G00302030.1, and NincChr6G00302100.1) exhibited enzymatic activity toward demethylsuberosin, with the same retention time as decursinol and marmesin (Fig. 5B and Supporting Information Fig. S23). Furthermore, we used osthenol to test their activities, and two genes, NincChr6G00302030.1 and NincChr6G00302100.1 displayed the corresponding activities (Fig. 5C). Hence, NincChr3G00106470.1 and NincChr3G00121510.1 were identified as monofunctional enzymes, with demethylsuberosin as the substrate, whereas NincChr6G00302030.1 and NincChr6G00302100.1 were bifunctional enzymes that can catalyze both demethylsuberosin and osthenol. To ensure the inclusion of other potential cyclases, genes with lower expression, such as NincChr6G00302130.1, NincChr6G00302140.1, NincChr6G00302160.1, and NincChr7G00344610.1 were cloned to test their activities. Interestingly, NincChr7G00344610.1 displayed monofunctional enzymatic activity toward osthenol and could only produce columbianetin in yeast (Fig. 5C and Supporting Information Fig. S24). Hence, NincChr3G00106470.1 and NincChr3G00121510.1 were named as demethylsuberosin cyclase (*NiDC1* and *NiDC2*); NincChr7G00344610.1 was named as osthenol cyclase (*NiOC1*); and NincChr6G00302030.1 and NincChr6G00302100.1 were named as osthenol/demethylsuberosin cyclase (*NiOD1* and *NiOD2*), respectively. Notably, NiODs are the first enzymes with dual functions ever reported to date. These enzymes fulfill the missing steps in complex coumarin skeleton biosynthesis and provide essential enzyme elements for the synthetic biology of complex coumarins.

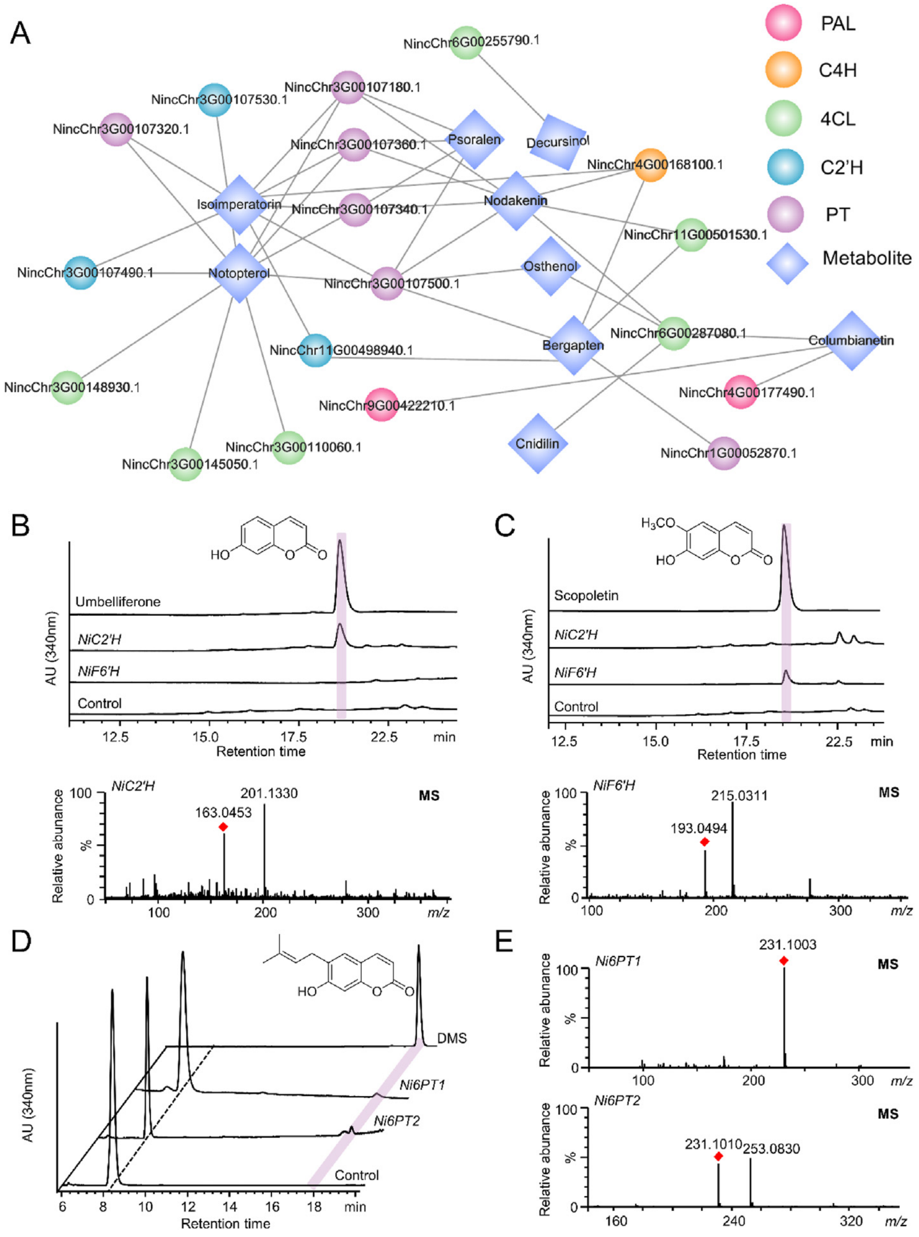


Figure 4 Functional verifications of *N. incisum* 2-OGD and PT genes. (A) The metabolome-transcriptome correlation analysis of *N. incisum.* The candidate genes in simple coumarin biosynthesis are listed. (B, C) Functional verifications of *N. incisum* 2-OGD genes. (D, E) Functional verifications of *N. incisum* PT genes. The products were detected using liquid chromatography (LC, at 340 nm) and LC−MS in positive ionization mode. The LC maps of standard umbelliferone, scopoletin, DMS (demethylsuberosin), and their corresponding chemical structures are shown in the figures. Boiled enzymes are used as a control. The red boxes indicate molecular ion peaks.

3.6. The evolution of the CYP736 family in N. incisum and the emergence of cyclase activities

Convergent or parallel evolution has been proposed to account for the independent acquisition of furanocoumarins by different plants[4,11,13]. The cyclase identified from Moraceae is a CYP76 protein, whereas the cyclases from Apiaceae identified in this study are classified as CYP736 proteins (Fig. 5), further supporting this hypothesis[13]. We explored the evolutionary history of these cyclases using a combined phylogenetic and comparative genomic approach. Additionally, considering that *N. incisum* cyclases display diverse activities (*NiDC*, *NiOD,* and *NiOC*), the mechanisms underlying the functional divergence of these cyclases were also studied.

By constructing a phylogeny comprising all CYP450 proteins from 18 species (Supporting Information Fig. S25 and Table S13), we examined the evolutionary positions of *N. incisum* cyclases and their relationships with other CYP450s. Phylogenetic analysis placed *N. incisum* cyclases in a monophyletic branch that included a minimum of 15 *N. incisum* CYP736 proteins (Fig. 6A). A deeper examination of this branch revealed that all genes within it belonged to the Apiaceae plants, suggesting that Apiaceae-specific duplications led to the emergence of five cyclase genes. Among these 15 CYP736 proteins, the overwhelming majority were from

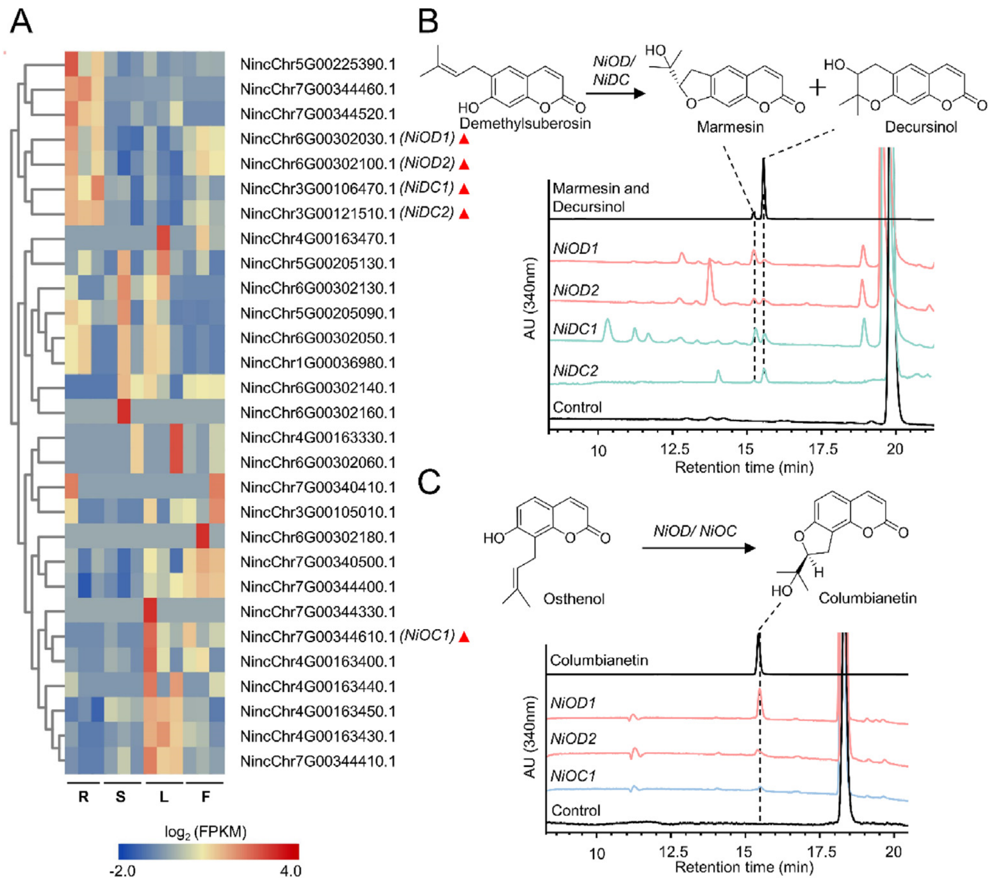


Figure 5 Functional verifications of *N. incisum* cyclases. (A) The heatmap shows the relative expression levels of the candidate genes in different tissues of *N. incisum*. The genes that we verify activities in this text are marked with a red triangle. R, roots; S, stems; L, leaves; F, fruits. (B) LC analyses of the cyclase activities toward demethylsuberosin at 340 nm. (C) LC analyses of the cyclase activities toward osthenol at 340 nm. *NiOD*s identified in the work are shown in rose red, *NiDC*s in green, *NiOC* in blue, and control (boiled enzymes) in black. The LC maps of standard marmesin, decursinol, columbianetin, and their corresponding chemical structures are shown in the supplementary figures.

chromosomes 3 and 6, and an often-occurring topology was that genes on chromosome 3 always formed sister groups with genes on chromosome 6 of *N. incisum*. The difference between *N. incisum* CYP736s on the two chromosomes was that those on chromosome 6 also grouped with genes from other species, suggesting an ancestral orthology among Apiaceae species, whereas *N. incisum* CYP736s on chromosome 3 only clustered with *N. incisum* CYP736s on chromosome 6. Such topologies suggest that the *N. incisum* CYP736s on chromosome 6 may have emerged earlier than the paralogues on chromosome 3 and that the CYP736s on chromosome 3 may represent duplicates of those on chromosome 6.

This duplication event is likely an ectopic duplication that covers a genomic block encoding several genes (Supporting Information Fig. 6B), as suggested by the phylogeny and the lack of collinearity between the two loci. Additionally, at least one tandem duplication event was inferred to have occurred before the ancestral chromosome (chromosome 6), because only two monophyletic groups containing genes from chromosomes 3 and 6 were observed. Tandem duplication likely occurred on chromosomes 3 and 6 after the ectopic duplication. This series of complex duplication events eventually resulted in the identification of four characterized cyclase genes on these two chromosomes. The fifth cyclase, located on chromosome 7, may have resulted result from a WGD/chromosomal segmental duplication, as this gene showed well-preserved synteny with the cyclase genes on chromosome 6 (Fig. S26). This duplication likely occurred before the ectopic duplication between chromosomes 3 and 6, as suggested by its basal position in the CYP736 phylogeny (Fig. 6A).

The evolutionary histories of the five cyclases were inferred, with *NiDC*s *via* WGD/chromosomal segmental duplication, *NiOC*s *via* ectopic duplication, and *NiOD*s on chromosome 6 through tandem duplications (Fig. 6B). However, the mechanisms underlying this divergence in enzymatic activity, such as sequence mutations, require further investigation. By comparing the amino acid sequences of the 15 *N. incisum* CYP736s (Supporting Information Fig. S27), we identified conserved and variable sites using four phylogenetically early-diverging *N. incisum* CYP736s as references (NincChr3G00105000.1, NincChr3G00105010.1, NincChr3G00105660.1, and NincChr6G00302180.1). Together with the evolutionary relationship, we were able to further infer the detailed process of amino acid variation (Fig. 6B and Fig. S27), which could help explain the divergence in enzymatic activities. Collectively, 18 variation sites were identified as

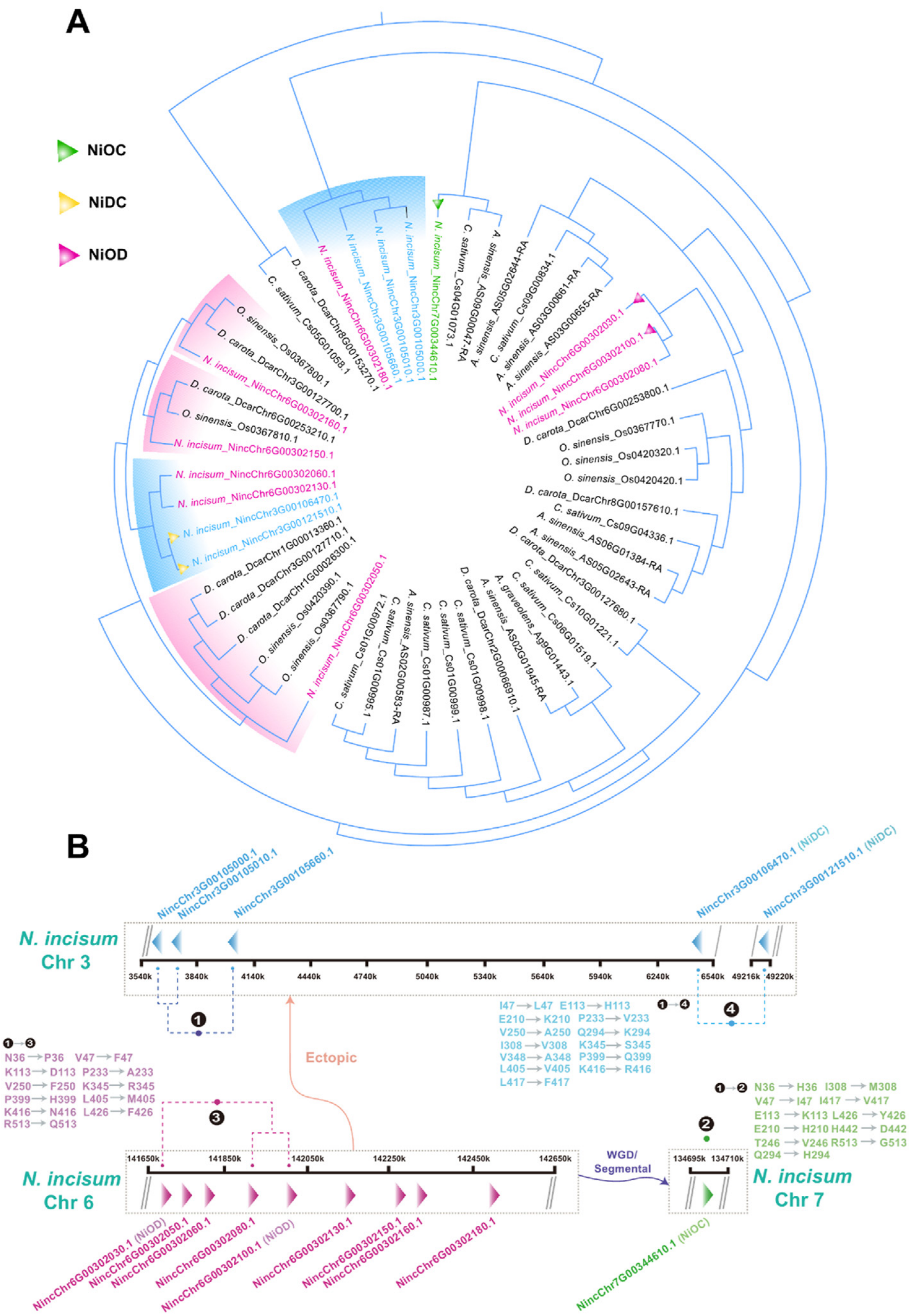

A
NiOC
NiDC
NiOD
N. incisum_NincChr7G00344610.1
C. sativum_Cs04G01073.1
A. sinensis_AS09G00047-RA
A. sinensis_AS05G02644-RA
C. sativum_Cs09G00834.1
A. sinensis_AS03G00661-RA
A. sinensis_AS03G00655-RA
N. incisum_NincChr6G00302030.1
N. incisum_NincChr6G00302100.1
N. incisum_NincChr6G00302080.1
D. carota_DcarChr6G00253800.1
O. sinensis_Os0367770.1
O. sinensis_Os0420320.1
O. sinensis_Os0420421.1
D. carota_DcarChr8G00157610.1
C. sativum_Cs09G04336.1
A. sinensis_AS06G01384-RA
A. sinensis_AS05G02643-RA
D. carota_DcarChr3G00127680.1
C. sativum_Cs10G01221.1
C. sativum_Cs06G01519.1
A. graveolens_Ag9G01443.1
A. sinensis_AS02G01945-RA
D. carota_DcarChr2G00068910.1
C. sativum_Cs01G00998.1
C. sativum_Cs01G00999.1
C. sativum_Cs01G00987.1
A. sinensis_AS02G00583-RA
C. sativum_Cs01G00995.1
C. sativum_Cs01G00972.1
N. incisum_NincChr6G00302050.1
O. sinensis_Os0367790.1
O. sinensis_Os0420390.1
D. carota_DcarChr1G00026300.1
D. carota_DcarChr3G00127710.1
D. carota_DcarChr1G00013360.1
N. incisum_NincChr3G00121510.1
N. incisum_NincChr3G00106470.1
N. incisum_NincChr6G00302130.1
N. incisum_NincChr6G00302060.1
N. incisum_NincChr6G00302150.1
O. sinensis_Os0367810.1
D. carota_DcarChr6G00253210.1
N. incisum_NincChr6G00302160.1
D. carota_DcarChr3G00127700.1
O. sinensis_Os0367800.1
C. sativum_Cs05G01058.1
D. carota_DcarChr8G00153270.1
N. incisum_NincChr6G00302180.1
N. incisum_NincChr3G00105660.1
N. incisum_NincChr3G00105010.1
N. incisum_NincChr3G00105000.1
B
NincChr3G00105000.1
NincChr3G00105010.1
NincChr3G00105660.1
NincChr3G00106470.1 (NiDC)
NincChr3G00121510.1 (NiDC)
N. incisum
Chr 3
3540k 3840k 4140k 4440k 4740k 5040k 5340k 5640k 5940k 6240k 6540k 49216k 49220k
1
1→4
4
I47→L47 E113→H113
E210→K210 P233→V233
V250→A250 Q294→K294
I308→V308 K345→S345
V348→A348 P399→Q399
L405→V405 K416→R416
L417→F417
1→3
N36→P36 V47→F47
K113→D113 P233→A233
V250→F250 K345→R345
P399→H399 L405→M405
K416→N416 L426→F426
R513→Q513
Ectopic
3
1→2
N36→H36 I308→M308
V47→I47 I417→V417
E113→K113 L426→Y426
E210→H210 H442→D442
T246→V246 R513→G513
Q294→H294
2
141650k 141850k 142050k 142250k 142450k 142650k
WGD/
Segmental
134695k 134710k
N. incisum
Chr 6
N. incisum
Chr 7
NincChr6G00302030.1 (NiOD)
NincChr6G00302050.1
NincChr6G00302060.1
NincChr6G00302080.1
NincChr6G00302100.1 (NiOD)
NincChr6G00302130.1
NincChr6G00302150.1
NincChr6G00302160.1
NincChr6G00302180.1
NincChr7G00344610.1 (NiOC)

Figure 6 Evolution of the CYP736 gene family in *N. incisum*. (A) Phylogenetic tree of CYP736 gene family in Apiaceae. Fifteen *N. incisum* genes are derived from chromosomes 3 (five genes labeled in blue), 6 (nine genes labeled in pink), and 7 (one gene labeled in green), respectively. The blue sector indicates that CYP736 genes on chromosomes 3 and 6 of *N. incisum* are paralogues with a sister relationship. Meanwhile, the pink sector shows CYP736 genes on chromosome 6 of *N. incisum* do not group with genes on chromosome 3, but clustered with genes from other species in Apiaceae. (B) The inferred evolutionary process of CYP736 gene family in *N. incisum*. The dotted line, orange arrow, and dark blue arrow are represented in tandem duplication, ectopic duplication, and WGD (segmental) duplication event, respectively. Numbers 1−4 show amino acid mutations during/after the duplication events, and are possibly related to the functional divergence of different cyclases in *N. incisum* (OC, DC, and OD).

potentially critical for the neo- and sub-functionalization of the 15 *N. incisum* CYP736s. Of the 18 amino acid sites, 13 (L47, H113, K210, V233, A250, K294, V308, S345, A348, Q399, V405, R416 and F417) may be responsible for DC activity, 11 (H36, I47, K113, H210, V246, H294, M308, V417, Y426, D442 and G513) for OC, and 12 (P36, F47, D113, A233, F250, R345, T348, H399, M405, N416, F426, and Q513) for dual OD activity.

In summary, all duplication events and sequence mutations were specific to Apiaceae, and some were possibly limited to *N. incisum*. The limited systematic grades for these duplication events agree with the inference of multiple and independent origins of the complex coumarin biosynthetic pathway in angiosperms, suggesting diversified plant biosynthetic product evolution.

3.7. Data availability

The genome data of this study had been deposited into the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database under the BioProject accession number PRJNA1017385. The accession numbers in the NCBI database of the genes whose functions have been verified in this work are as follows: *NiC2′H* (OR820159), *NiF6′H* (OR820160), *Ni6PT1* (OR820161), *Ni6PT2* (OR820162), *NiDC1* (OR820163), *NiDC2* (OR820164), *NiOC1* (OR820165), *NiOD1* (OR820166), *NiOD2* (OR820167). Other data regarding this study is included in the supporting information.

4. Discussion

Coumarins are widely distributed throughout the plant kingdom, with thousands of species from at least 75 families containing coumarins[12]. However, these typically refer to the simple coumarins. Complex coumarins, such as furanocoumarins and pyranocoumarins, are mainly limited to Apiaceae, Rutaceae, Moraceae, and Fabaceae[1]. The main differences between simple and complex coumarins lie in the prenylation and cyclization steps (Fig. 1), which establish the basic skeleton of complex coumarins. All previously identified PTs and cyclases were limited to the Apiaceae, Rutaceae, and Moraceae families[9,10,13,55,56]. Genes involved in simple coumarin biosynthesis have been identified in the Cruciferaceae, Convolvulaceae, Rutaceae, and Apiaceae[3]. Therefore, the functional evolution of PT and cyclases in the coumarin pathway is essential for the formation of complex coumarins.

The 2-OGD family proteins are responsible for simple coumarin formation by directing metabolic flux from the phenylpropanoid pathway to the coumarin biosynthesis pathway. The family mainly includes two functionally different kinds of enzymes, C2′H and F6′H. C2′H has a broad substrate recognition ability. Besides producing umbelliferone, it could also catalyze feruloyl-CoA to form scopoletin, while F6′H could only accept feruloyl-CoA as a substrate[8]. Owing to the following reasons: (1) the identified C2′Hs are mainly from Apiaceae and Rutaceae, and F6′Hs are mainly from Cruciferaceae and Convolvulaceae[6-8,57]; (2) Umbelliferone produced by C2′H is a key intermediate for complex coumarins biosynthesis[7]; (3) Simple coumarins in plants mainly represented by scopoletin and its derivatives and (4) The formation of simple coumarins in plants rich in complex coumarins is catalyzed by C2′H, while those plants catalyzed by F6′H can only accumulate simple coumarins[8], thus we deduce that the absence/presence of *C2′H* genes is also a prior indicator for the production of simple coumarins or complex coumarins. In summary, a species possessing only *F6′H* could produce simple coumarins, whereas one possessing *C2′H* has the capability of producing complex coumarins. However, further evidence is required, as only a limited number of *C2′H* and *F6′H* genes have been identified from a limited number of species[6,7,53,57]. Additionally, C2′H and F6′H have no significant difference in protein identity, and their functions are usually interchangeable (Fig. S18)[8].

Another noteworthy phenomenon that warrants mention is the multiple origins of the coumarin pathway in angiosperms[56]. The identification of PTs from different species, as well as the genes involved in the hydroxylation of complex coumarins, suggests the possibility of parallel or convergent evolution contributing to the multiple origins of the coumarin pathway[4,10,13,11,56]. Furthermore, although many PTs and hydroxylases have been identified in Apiaceae and Moraceae, no C-PTs have been identified in Rutaceae or Fabaceae, even though these families are rich in complex coumarins[9-11,56]. These results imply that the PTs in Rutaceae and Fabaceae may have ancestors different from those in Apiaceae and Moraceae. In this study, we found that cyclases from *N. incisum* (Apiaceae) belonged to the CYP736 family, whereas those from Moraceae belonged to the CYP76 family[13,15], which supports their independent origins (Fig. S25). Additionally, the genes involved in the hydroxylation step of xanthotoxin, a representative furanocoumarin in plants, are CYP71AZ1/6 in Apiaceae, CYP82D64 in Rutaceae, and CYP82C2/4 in Brassicaceae[4,58]. Hence, we speculate that the cyclases in Rutaceae and Fabaceae may not belong to the CYP76 or CYP736 families. Our inference may provide limited assistance to researchers striving to explore the complex coumarin biosynthetic pathways in Rutaceae and Fabaceae. As observed in the parallel evolution of PTs, it is not surprising that cyclases arose independently in distantly related taxa[4].

Coumarin metabolic profiles differ significantly across angiosperms, not only in their abundance but also in their structural characteristics. These structural differences encompass variations in linear and angular configurations, as well as furan- and pyran-type structures. A prevailing hypothesis suggests that angular coumarins originated from linear coumarins[1]. This hypothesis is mainly based on the following observations: (a) no angular coumarins exist in a plant species alone, but always co-exist with the linear coumarins[1]; (b) linear coumarins are discovered in more families including Apiaceae, Rutaceae, Moraceae, and Fabaceae, whereas angular coumarins are limited to certain species of Apiaceae[9]; and (c) the angular coumarins can function synergistically with linear coumarins, which further improves the toxicity of coumarins to insects[59]. However, to date, direct evidence supporting this hypothesis is still lacking. The functional evolution of *PT*s is likely to provide support to this hypothesis, but far too few cases (*PsPT2* and *PpPT2*) can be utilized to study this subject in Apiaceae[10,15]. In addition, certain linear PTs exhibited extremely weak angular activity, which is consistent with (a)[10,15]. Regarding furanocoumarins and pyranocoumarins, even less is known about the mechanism of their structural diversity, because very few cyclases have ever been identified[13,15]. Although according to our analyses, different cyclases are probably derived from different means of gene duplications (tandem, ectopic, and WGD/segmental), ultimately yielding different coumarins structures (Fig. 6), mutations in key amino acid sites that control substrate and/or product specificity may be more vital for the structural diversity. According to our recent study, the acidity and alkalinity of the reaction solution significantly affected the formation of furanocoumarins and pyranocoumarins[15]. However, the precise mechanisms require further investigation at the protein level.

Acknowledgments

This research was supported by the Western Light Project of the Chinese Academy of Sciences and the Young Qihuang Scholars Project of the National Administration of Traditional Chinese Medicine. This research was also funded by the open foundation of Shaanxi University of Chinese Medicine state key laboratory of R&D of Characteristic Qin Medicine Resources (SUCM-QM202202, China), the fund of Traditional Chinese Medicine Institute of Anhui Dabie Mountain (TCMADM-2023-18, China), and “Double First-Class” University Project of China Pharmaceutical University (CPU2022QZ27, China). It was also supported by the key project at the central government level: the ability establishment of sustainable use for valuable Chinese medicine resources (2060302, China) and the open research fund of Yunnan characteristic plant extraction laboratory (YKKF2023002, China).

Author contributions

Qien Li: Investigation, Writing – original draft. Yiqun Dai: Data curation, Formal analysis. Xin-Cheng Huang: Data curation, Formal analysis, Software. Lanlan Sun: Data curation, Formal analysis, Methodology. Kaixuan Wang: Data curation, Formal analysis. Xiao Guo: Resources. Dingqiao Xu: Investigation, Methodology. Digao Wan: Investigation, Resources. Latai An: Resources. Zixuan Wang: Validation. Huanying Tang: Methodology. Qi Qi: Data curation. Huihui Zeng: Methodology. Minjian Qin: Supervision. Jia-Yu Xue: Software, Supervision. Yucheng Zhao: Supervision, Writing – original draft, Writing – review & editing.

Conflicts of interest

The authors declare no conflicts of interest.

Appendix A. Supporting information

Supporting information to this article can be found online at https://doi.org/10.1016/j.apsb.2024.04.005.

References

1. Bourgaud F, Hehn A, Larbat R, Doerper S, Gontier E, Kellner S, et al. Biosynthesis of coumarins in plants: a major pathway still to be unravelled for cytochrome P450 enzymes. *Phytochemistry Rev* 2006; **5**:293–308.
2. Taylor RD, MacCoss M, Lawson ADG. Rings in drugs. *J Med Chem* 2014;**57**:5845–59.
3. Rodrigues JL, Rodrigues LR. Biosynthesis and heterologous production of furanocoumarins: perspectives and current challenges. *Nat Prod Rep* 2021;**38**:869–79.
4. Limones-Mendez M, Dugrand-Judek A, Villard C, Coqueret V, Froelicher Y, Bourgaud F, et al. Convergent evolution leading to the appearance of furanocoumarins in citrus plants. *Plant Sci* 2020;**292**: 110392.
5. 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* 2021;**26**:248–59.
6. Kai K, Mizutani M, Kawamura N, Yamamoto R, Tamai M, Yamaguchi H, et al. Scopoletin is biosynthesized *via ortho*-hydroxylation of feruloyl CoA by a 2-oxoglutarate-dependent dioxygenase in *Arabidopsis thaliana*. *Plant J* 2008;**55**:989–99.
7. Vialart G, Hehn A, Olry A, Ito K, Krieger C, Larbat R, 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* 2012;**70**: 460–70.
8. Sun XX, Zhou DY, Kandavelu P, Zhang H, Yuan QP, Wang BC, et al. Structural insights into substrate specificity of feruloyl-CoA 6′-hydroxylase from *Arabidopsis thaliana*. *Sci Rep* 2015;**5**:10355.
9. Karamat F, Olry A, Munakata R, Koeduka T, Sugiyama A, Paris C, et al. A coumarin-specific prenyltransferase catalyzes the crucial biosynthetic reaction for furanocoumarin formation in parsley. *Plant J* 2014;**77**:627–38.
10. Munakata R, Olry A, Karamat F, Courdavault V, Sugiyama A, Date Y, et al. Molecular evolution of parsnip (*Pastinaca sativa*) membrane-bound prenyltransferases for linear and/or angular furanocoumarin biosynthesis. *New Phytol* 2016;**211**:332–44.
11. Munakata R, Kitajima S, Nuttens A, Tatsumi K, Takemura T, Ichino T, et al. Convergent evolution of the UbiA prenyltransferase family underlies the independent acquisition of furanocoumarins in plants. *New Phytol* 2020;**225**:2166–82.
12. Stringlis IA, de Jonge R, Pieterse CMJ. The age of coumarins in plant-microbe interactions. *Plant Cell Physiol* 2019;**60**:1405–19.
13. Villard C, Munakata R, Kitajima S, van Velzen R, Schranz ME, Larbat R, et al. A new P450 involved in the furanocoumarin pathway underlies a recent case of convergent evolution. *New Phytol* 2021;**231**: 1923–39.
14. Hamerski D, Matern U. Elicitor-induced biosynthesis of psoralens in *Ammi majus* L. suspension cultures. Microsomal conversion of demethylsuberosin into (+)marmesin and psoralen. *Eur J Biochem* 1988;**171**:369–75.
15. Zhao Y, He Y, Han L, Zhang L, Xia Y, Yin F, et al. Two types of coumarins-specific enzymes complete the last missing steps in pyran- and furanocoumarins biosynthesis. *Acta Pharm Sin B* 2024;**14**: 869–80.
16. Jian X, Zhao Y, Wang Z, Li S, Li L, Luo J, et al. Two CYP71AJ enzymes function as psoralen synthase and angelicin synthase in the biosynthesis of furanocoumarins in *Peucedanum praeruptorum* Dunn. *Plant Mol Biol* 2020;**104**:327–37.
17. Hehmann M, Lukacin R, Ekiert H, Matern U. Furanocoumarin biosynthesis in *Ammi majus* L. Cloning of bergaptol *O*-methyltransferase. *Eur J Biochem* 2004;**271**:932–40.
18. Liu X, Jiang S, Xu K, Sun H, Zhou Y, Xu X, et al. Quantitative analysis of chemical constituents in different commercial parts of *Notopterygium incisum* by HPLC–DAD-MS. *J Ethnopharmacol* 2009;**126**:474–9.
19. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* 2011;**27**: 764–70.
20. 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* 2019;**5**:833–45.
21. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* 2019;**37**:907–15.
22. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc* 2016;**11**:1650–67.
23. Keilwagen J, Hartung F, Grau J. GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq Data. *Methods Mol Biol* 2019;**1962**:161–77.
24. Hoff KJ, Stanke M. Predicting genes in single genomes with AUGUSTUS. *Curr Protoc Bioinformatics* 2019;**65**:e57.
25. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol* 2008;**9**:R7.
26. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith Jr RK, Hannick LI, et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* 2003;**31**:5654–66.
27. Buchfink B, Reuter K, Drost HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* 2021;**18**:366–8.
28. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 2019;**47**:D309–14.
29. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: the protein families database in 2021. *Nucleic Acids Res* 2021;**49**:D412–9.
30. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinf* 2008;**9**:18.
31. Ou S, Jiang N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mobile DNA* 2019;**10**:48.

32. Zwaenepoel A, Van de Peer Y. Wgd-simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics* 2019;**35**: 2153−5.

33. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinf* 2009;**10**:421.

34. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;**32**:1792−7.

35. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 2007;**24**:1586−91.

36. Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;**5**: e9490.

37. Proost S, Fostier J, De Witte D, Dhoedt B, Demeester P, Van de Peer Y, et al. i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res* 2012;**40**:e11.

38. Sensalari C, Maere S, Lohaus R. Ksrates: positioning whole-genome duplications relative to speciation events in *K*s distributions. *Bioinformatics* 2022;**38**:530−2.

39. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 2019;**20**:238.

40. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* 2007;**56**:564−77.

41. Xu D, Lin H, Tang Y, Huang L, Xu J, Nian S, et al. Integration of full-length transcriptomics and targeted metabolomics to identify benzylisoquinoline alkaloid biosynthetic genes in *Corydalis yanhusuo*. *Hortic Res* 2021;**8**:16.

42. Han L, Zhang L, He Y, Liao L, Li J, Xu S, et al. Three carbon-/oxygen-prenyltransferases responsible for furanocoumarin synthesis in *Angelica dahurica*. *Ind Crops Prod* 2023;**200**:116814.

43. Han X, Li C, Sun S, Ji J, Nie B, Maker G, et al. The chromosome-level genome of female ginseng (*Angelica sinensis*) provides insights into molecular mechanisms and evolution of coumarin biosynthesis. *Plant J* 2022;**112**:1224−37.

44. Van de Peer Y, Mizrachi E, Marchal K. The evolutionary significance of polyploidy. *Nat Rev Genet* 2017;**18**:411−24.

45. Iorizzo M, Ellison S, Senalik D, Zeng P, Satapoomin P, Huang J, et al. A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nat Genet* 2016;**48**:657−66.

46. Song X, Wang J, Li N, Yu J, Meng F, Wei C, et al. Deciphering the high-quality genome sequence of coriander that causes controversial feelings. *Plant Biotechnol J* 2020;**18**:1444−56.

47. Li MY, Feng K, Hou XL, Jiang Q, Xu ZS, Wang GL, et al. The genome sequence of celery (*Apium graveolens* L.), an important leaf vegetable crop rich in apigenin in the Apiaceae family. *Hortic Res* 2020;**7**:9.

48. Song X, Sun P, Yuan J, Gong K, Li N, Meng F, et al. The celery genome sequence reveals sequential paleo-polyploidizations, karyotype evolution and resistance gene reduction in apiales. *Plant Biotechnol J* 2021;**19**:731−44.

49. Liu GQ, Dong J, Wang H, Hashi Y, Chen SZ. Comparison of two species of *Notopterygium* by high-performance liquid chromatography-photodiode array detection-electrospray ionization tandem mass spectrometry. *Eur J Mass Spectrom* 2012;**18**:59−69.

50. Sui Z, Luo J, Yao R, Huang C, Zhao Y, Kong L. Functional characterization and correlation analysis of phenylalanine ammonia-lyase (PAL) in coumarin biosynthesis from *Peucedanum praeruptorum* Dunn. *Phytochemistry* 2019;**158**:35−45.

51. Zhao Y, Jian X, Wu J, Huang W, Huang C, Luo J, et al. Elucidation of the biosynthesis pathway and heterologous construction of a sustainable route for producing umbelliferone. *J Biol Eng* 2019;**13**:44.

52. Liu T, Yao R, Zhao Y, Xu S, Huang C, Luo J, et al. Cloning, functional characterization and site-directed mutagenesis of 4-coumarate: coenzyme A ligase (4CL) involved in coumarin biosynthesis in *Peucedanum praeruptorum* Dunn. *Front Plant Sci* 2017;**8**:4.

53. Yao R, Zhao Y, Liu T, Huang C, Xu S, Sui Z, 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* 2017;**95**:199−213.

54. Kawai Y, Ono E, Mizutani M. Evolution and diversity of the 2-oxoglutarate-dependent dioxygenase superfamily in plants. *Plant J* 2014;**78**:328−43.

55. Roselli S, Olry A, Vautrin S, Coriton O, Ritchie D, Galati G, et al. A bacterial artificial chromosome (BAC) genomic approach reveals partial clustering of the furanocoumarin pathway genes in parsnip. *Plant J* 2017;**89**:1119−32.

56. Munakata R, Olry A, Takemura T, Tatsumi K, Ichino T, Villard C, et al. Parallel evolution of UbiA superfamily proteins into aromatic *O*-prenyltransferases in plants. *Proc Natl Acad Sci U S A* 2021;**118**:e2022294118.

57. Matsumoto S, Mizutani M, Sakata K, Shimizu B. Molecular cloning and functional analysis of the ortho-hydroxylases of *p*-coumaroyl coenzyme A/feruloyl coenzyme A involved in formation of umbelliferone and scopoletin in sweet potato, *Ipomoea batatas* (L.) Lam. *Phytochemistry* 2012;**74**:49−57.

58. Krieger C, Roselli S, Kellner-Thielmann S, Galati G, Schneider B, Grosjean J, et al. The CYP71AZ P450 subfamily: a driving factor for the diversification of coumarin biosynthesis in Apiaceous plants. *Front Plant Sci* 2018;**9**:820.

59. Berenbaum M. Coumarins and caterpillars: a case for coevolution. *Evolution* 1983;**37**:163−79.