Horse Chestnut Tree Genome Reveals the evolutionary mechanism of Aescin and Aesculin biosynthesis

Wei Sun

Qinggang Yin

Key Laboratory of Plant Resources and Beijing Botanical Garden, Institute of Botany, Chinese Academy of Sciences

Huihua Wan

Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences

Chao Xiong

Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences

Chong Xie

State Key Laboratory of Chemical Resource Engineering, Beijing University of Chemical Technology

Xiangxiao Meng

Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences

Caixia Wang

Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences

Ranran Gao

Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences

Weiqiang Chen

Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences

Ziyan Xie

Northeast Forestry University

Zheyong Xue

John Innes Centre

Hui Yao

Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences and Peking Union Medical College

Peng Sun

Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences

Xuehua Xie

Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences

Zhigang Hu

Hubei University of Chinese Medicine

David Nelson
University of Tennessee Health Science Center
Zhichao Xu (zcxu@nefu.edu.cn)
Northeast Forestry University
Xinxiao Sun
Beijing University of Chemical Technology https://orcid.org/0000-0001-5590-6473
Shilin Chen
China Academy of Chinese Medical Sciences https://orcid.org/0000-0002-0449-236X

Article

Keywords:

Posted Date: November 30th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2297848/v1

Additional Declarations: There is **NO** Competing Interest.

Version of Record: A version of this preprint was published at Nature Communications on October 13th, 2023. See the published version at https://doi.org/10.1038/s41467-023-42253-y.

Abstract

Medicinal trees provide a main resource for diverse medicinal compounds. However, the biosynthesis of tree metabolites and their pathway evolution has gained limited understanding. Horse chestnut (*Aesculus chinensis*) is an important medicinal tree and its seeds are rich in aescins, barrigenol-type triterpenoid saponins (BAT), and aesculin, a coumarin glycoside, which are effective in the therapy of chronic venous insufficiency and asthenopia (eye strain). To understand the biosynthesis of these compounds, herein, we assembled a 470.04-Mb high-quality horse chestnut genome and characterized an *Aesculus*-specific whole-genome duplication (WGD) event. Spatial metabolome imaging, co-expression, and biosynthetic gene cluster analyses indicated that the *Aesculus*-specific WGD event led to the formation of two gene clusters (BGCs) including oxidosqualene cyclase, cytochrome P450 monooxygenase, cellulose synthase-derived glycotransferases, and BADH acyltransferases. Further biochemical investigation revealed the roles of *AcOCS6*, *AcCYP716A278*, *AcCYP716A275*, *AcCSL1*, and *AcBAHD3* genes distributed between these two BGCs in catalyzing the formation of aescins. To understand the evolution of BAT pathways, the collinearity analysis showed the collinear BGC segments could be traced back to early-diverging angiosperms, then the essential gene-encoding enzymes necessary for the BAT biosynthesis were recruited before the split of *Aesculus*, *Acer*, and *Xanthoceras*. Meanwhile, we identified three UDP-glucosyltransferases and demonstrated their involvement in the biosynthesis of aesculin via a *de novo* synthesis. Taken together, these findings provide important information in understanding the evolution of gene clusters associated with medicinal tree metabolites.

Introduction

Plants synthesize a wide variety of structurally diverse medicinal compounds, such as noscapin, casbenes, avenacins, cucurbitacins, and taxadiene, are exclusive to specific lineage or species. It is interesting that the biosynthesis of numerous compounds has been discovered to result from a biosynthetic gene cluster (BGC), which is characterized by the physically colocalized genes involved in a common metabolite pathway (1–6). With the growing body of genome sequence information available, multiple investigations have been completed to understand the collinear BGC organization within and across species(7–9). Examples include triterpene thalianols in *Arabidopsis thaliana* and close relatives, benzylisoquinoline alkaloids in *Papaver*, steroidal glycoalkaloids in the Solanaceae, triterpene cucurbitacins in the Cucurbitaceae, and momilactone diterpenoids in the grasses (9–13). Evidence from herbaceous plants showed that complex genetic events such as whole genome duplication (WGD), tandem duplication, gene relocation, chromosomal inversion, lateral gene transfer impact on the assembly, maintaining, and diversification of physically clustered and non-homologous, but functionally coordinated genes during the evolutionary process (13–17). By contrast, little is known about genetic mechanisms leading to the exclusive production of pharmaceuticals in trees.

Aesculus L. (horse chestnut) in soapberry family (Sapindaceae) is native to the temperate northern hemisphere. Its seeds produce medicinal barrigenol-type (BAT) triterpenoid saponins (escin Ia and escin Ib, collectively referred to as aescin) and coumarins (aesculin and fraxin) (18–21). Escin Ia belongs to

BAT triterpenoids derived from an oleanane (β-amyrin) backbone, which further undergoes region-specific C-16, C-21, C-22, C-24, and C-28 oxygenation to become protoescigenin, the main aglycone of aescins. A glucuronosyl moiety exists at the C-3 position of protoescigenin with additional diglucoside and 2-methylcrotonoylation and acetylation at the C-21 and C-22 positions (22, 23). Aescin preparations have been administered orally, intravenously (sodium aescinate containing escin Ia and escin Ib), and topically in clinical trials for treatment of chronic venous insufficiency, oedema, and haemorrhoids(24, 25). Aesculin (6,7-dihydroxycoumarin 6-*O*-glucoside) and fraxin (6-methoxy-7,8-dihydroxycoumarin 8-*O*-glucoside) are formed from the glycosylation of coumarin (26, 27). Aesculin from *Aesculus* together with digitoxin from *Digitalis purpurea* have been used as well-known eye drops in Europe, alleviating asthenopia, eye pain, and diplopia (28, 29).

To date, the biosynthesis of escin Ia, escin Ib, and aesculin in horse chestnut tree and in this lineage remains uninvestigated. This unknown limit the efforts to increase their production medicinal application. In this study, to understand the biosynthesis of these compounds, we completed a chromosome-scale assembly of a 470.04-Mb genome sequence of *A. chinensis*. We detected an ancient remnant of a whole-genome duplication that occurred approximately 30.8 million years ago (MYA) in *Aesculus*, and found that it might have contributed to two triterpenoids-related BGCs. One consists of one full-length 2,3-oxidosqualene cyclase (OSC) gene, seven cytochrome P450 monooxygenase (CYP450) genes, and five BAHD acyltransferase genes, and the other consists of two CYP450 genes and one cellulose synthase-like (CSL) gene. Based on these data, we functionally analyzed five candidate genes involving aescin biosynthesis. Meanwhile, we utilized a comparative genomics-driven approach to understand the organization and evolution of related clusters across angiosperms. Furthermore, we identified two uridine diphosphate (UDP)-dependent glycosyltransferases (UGTs) that were enzymatically characterized to catalyze the formation of aesculin.

Results And Discussion

Metabolite profiling of A. chinensis in different tissues and fruit structures

We applied matrix-assisted laser desorption/ionization (MALDI)-mass spectrometry imaging (MSI) to explore the spatial distribution of triterpenoid and coumarin metabolites in epicarps, sarcocarps, and seeds of fruits (Fig. 1A). MALDI-MSI analysis showed that fraxin (*m/z* 409.0543 $[C_{16}H_{18}O_{10} + K]^+$) mainly accumulated in the epicarp while triterpenoid saponins such as aesculiside A (*m/z* 1129.484 $[C_{54}H_{88}O_{23} + K]^+$), aesculioside G (*m/z* 1183.531 $[C_{56}H_{88}O_{24} + K]^+$), escin Ia (*m/z* 1169.515 $[C_{55}H_{86}O_{24} + K]^+$), escin V (*m/z* 1157.515 $[C_{54}H_{86}O_{24} + K]^+$), and isoescin b (*m/z* 1139.505 $[C_{54}H_{84}O_{23} + K]^+$) mainly accumulated in the seeds (Figs. 1B and 1C, Table S1). Additionally, LC-MS was performed to understand metabolite profiles not only in stems, flowers, leaves, epicarps, sarcocarps, and seeds but also in different developmental stages of these tissues indicated by collection times. The contents of aesculin and fraxin were significantly higher in leaves and flowers than in epicarps, sarcocarps, and seeds ($p < 0.05$) (Fig. 1D, Table S2-3). The content of protoescigenin, the main aglycogen of aescin, was the highest in stems and

seeds collected in October. The contents of escin Ia, escin Ib, and total triterpenoids were the highest in seeds collected in October, followed by seeds collected in September, August, and July, then by flowers collected in May. ($p < 0.05$) (Fig. 1E, Table S2-3), while these two compounds were hardly detected in leaves collected from March to October. These data indicate that the accumulation of these compounds depends upon tissues and the developmental stage of seeds and flowers.

Genome assembly and gene annotation of A. chinensis

The size of the *A. chinensis* genome was predicted to be 481.90 Mb using flow cytometry (Figure S1). Genome survey of *A. chinensis* based on 17 *k*-mer frequency of Illumina short reads showed that *A. chinensis* genome is 504.28 Mb with a small heterozygous peak and an obvious repetitive peak, suggesting it has a low level of heterozygosity (about 0.37%) (Figure S2). In addition, long-read sequencing using Oxford Nanopore Technologies (ONT) obtained 34 Gb data with a ~ 68× coverage and an N50 length of 11.74 kb (Table S4). After error-correction, trimming, and assembling, filtered ONT reads were assembled into 656 contigs with a total size of 470.02 Mb, an N50 length of 2.05 Mb, and the longest contig of 8.58 Mb, covering 97.50% of the estimated nuclear genome size (Table S5). The assembled contigs comprised many sequencing errors because of the low accuracy of ONT sequencing. Accordingly, this contig-level assembly was further polished three times using Illumina short reads, and 97.4% (1,573 out of 1,614) plant single-copy orthologs were identified using Benchmarking Universal Single-Copy Orthologs (BUSCOs) estimation, indicating a high degree of completeness in the polished genome of *A. chinensis* (Table S6). In addition, Hi-C chromosome conformation capture sequencing generated 344,838,272 raw paired-end reads, of which 58.05% (189,382,086) were mapped to the contig assembly as unique paired-end reads. Among these unique reads, 174,576,440 were captured to guide the pseudo-chromosome assembly. A total of 461.08 Mb (98.09%) of the assembled genome was anchored to 20 pseudo-chromosomes (2n = 40) (Fig. 2A, Figure S3). Table S5 lists the detailed characteristics of the *A. chinensis* genome.

Approximately 59.14% (276,695,955 bp) of repetitive DNAs in the *A. chinensis* genome (Table S5) were annotated in accordance with the transposable element (TE) content of reported Sapindaceae genomes, namely *Xanthoceras sorbifolium* (56.39%) and *Dimocarpus longan* (52.87%). Of these repetitive elements, 24.37% (114,021,841 bp) TEs were long terminal repeat (LTR) retrotransposons, of which 98.1% belonged to the G*ypsy* superfamily (32.7%) and *Copia* superfamily (65.4%) (Table S7). Moreover, 36,557 protein-coding genes were identified via integrating *ab initio* gene predictions, homologous protein searches, and the *de novo* assembled transcripts from RNA-seq reads. In addition, 35,790 (97.9%) genes could be located on the 20 pseudo-chromosomes. The identified orthologs covered 95.4% of the embryophyta BUSCOs, indicating that the annotated genome is largely complete (Table S8). We further identified orthologous groups of proteomes from *A. chinensis* and 13 other Rosids species, including *Arabidopsis thaliana*, *Brassica rapa*, *Citrus clementina*, *Cucumis sativus, D. longan, Glycine max*, *Gossypium raimondiim*, *Malus domestica*, *Populus trichocarpa*, *X. sorbifolium*, *Theobroma cacao*, and *Vitis vinifera*, and harvested a total of 21,299 orthologous groups covering 497,631 genes. We then compared the genomes of candidate plant species to obtain gene families that are significantly

expanded in *A. chinensis* or that are unique to *A. chinensis* (Figure S4). Functional prediction showed that the expanded gene families are especially enriched in the KEGG pathways of secondary metabolites, such as terpenoid biosynthesis (KO00900: terpenoid backbone biosynthesis, KO00909: sesquiterpenoid and triterpenoid biosynthesis) and phenylpropanoid biosynthesis (KO00940) (Figure S5).

Phylogenomic Dating And Whole-genome Duplication Analysis

A total of 139 single-copy genes from 13 Rosids species were selected to construct a high-confidence phylogenetic tree. The phylogenetic trees from both concatenated nucleotide and protein sequences of single-copy genes supported a close relationship of *A. chinensis* with other sequenced Sapindaceae species, *X. sorbifolium* Bunge and *D. longan* Lour. (Fig. 2B). In these analyses, *A. chinensis* was found as a sister lineage to *X. sorbifolium*, and both further formed a sister group to *D. longan*. Molecular dating of the tested lineages using the nucleotide sequences of single-copy genes and fossil age calibrations inferred that the divergence of families Sapindaceae and Rutaceae of Sapindales occurred at approximately 36.3 MYA with 95% confidence interval (CI) from the range of 21.8 MYA to 50.4 MYA. The split between *A. chinensis* and *X. sorbifolium* occurred at approximately 32.5 MYA with a 95% CI from 17.5 MYA to 45.4 MYA (Fig. 2B).

Intragenomic collinear analysis identified at least one whole-genome duplication (WGD) event in the *A. chinensis* genome (Fig. 2D). Collinearity analyses between *A. chinensis* and *X. sorbifolium*, and between *A. chinensis* and *C. clementina* showed that two paralogous segments in *A. chinensis* corresponded to one orthologous region in *X. sorbifolium* and *C. clementina*, respectively (Fig. 2D, Fig. 2E, Figure S6). These results supported that the species-specific WGD event might occur in *A. chinensis* after its divergence with speciation from the common ancestor of *A. chinensis* and *X. sorbifolium*. In addition, the distributions of synonymous substitutions per synonymous site (K_S) for paralogous genes and anchor pairs in collinear regions of *A. chinensis* showed a clear peak at approximately 0.24 and a minor peak around 1.75, suggesting that the *A. chinensis* genome might have experienced two WGD events (Fig. 2C, Figure S7). Previous studies have suggested that no WGD event occurred in *C. clementina* (30) and *D. longan* (31) after the ancestral gamma triplication (γ-WGT) event. Our analysis confirmed that only one K_S peak was detected in *C. clementina* and *X. sorbifolium*, at approximately 1.5 and 1.75, respectively. This ancestral K_S peak, shared by *A. chinensis*, *C. clementina*, and *X. sorbifolium* genomes, represents the γ-WGT event. The K_S distribution of orthologs between *A. chinensis* and *X. sorbifolium* showed one K_S peak at around 0.25, slightly larger than the K_S value of paralogs in *A. chinensis*. This again suggests that the recent WGD event in *A. chinensis* happened after the divergence between *A. chinensis* and *X. sorbifolium*. Using the divergence time and mean K_S value of orthologs between *A. chinensis* and *X. sorbifolium*, between *A. chinensis* and *C. clementina*, and between *A. chinensis* and *V. vinifera*, we estimated that the species-specific WGD event in *A. chinensis* (*Aα*) occurred at around 30.8 ± 1.33 MYA (Fig. 2B). Further, we identified 3,358 genes that retained duplicates from the recent *Aα* event in *A. chinensis*. The functional annotation and GO enrichment analyses showed that the retained duplicates

might be related to the responses of *A. chinensis* to various stimuli (e.g., biotic stimulus, acid chemical, and stress) and regulation of these responses (Table S9).

The origin of large and diverse taxonomic lineages can be related to ancestral polyploidy events (32). WGD is the major evolutionary force for the production of phenotypic diversity, speciation, and domestication (33, 34). The examples of WGD-derived plant phenotypic and metabolic diversity contributing to species-specific gene expansion include oil biosynthesis in wild olive trees, ursane triterpene synthesis in loquat (*Eriobotrya japonica*), camptothecin production in camptotheca tree (*Camptotheca acuminata*), and triptolide content in thunder god vine (*Tripterygium wilfordii*) (35–37). To investigate the impact of the *A. chinensis*-specific WGD event on aescin and aesculin biosynthesis, we systematically calculated *Ks* for each duplicated paralogous gene pair, emphasizing the upstream pathway genes involved in terpene biosynthesis (i.e., *AACT*, *HMGS*, *HMGR*, *MVA*, *MVK*, *PMK*, *MVD*, *IDI*, *DXS*, *DXR*, *MCT*, *CMK*, *MDS*, *HDS*, *HDR*, *FPS*, *SQS*, and *SQE*) and coumarin biosynthesis (*PAL*, *C4H*, *C3H*, *4CL*, and *COMT)* (Table S10). We found that the *Aα* WGD only led to retention of duplicates in the terpene pathway, suggesting that metabolic flux may have shifted toward triterpenoid metabolism, resulting in aescin production.

Biosynthetic gene cluster and weighted gene co-expression network analyses (WGCNA) for discovering the Escin Ia Pathway

Aescins are triterpene saponins, biosynthesized from 30-carbon intermediate 2,3-oxidosqualene (Figure S8) by sequential actions of multiple enzymes, including oxidosqualene cyclase, cytochrome P450 monooxygenase (CYPs), glycotransferases (UGTs), and acyltransferases (BAHDs) (38, 39). The pentacyclic triterpene aescin skeleton is derived from dammarenylcation and D and E ring expansion en route to β-amyrin. β-amyrin scaffold further undergoes site-specific oxidation catalyzed by P450s, forming diverse non-glycosylated aglycones, which are collectively referred to as protoaescigenin (18). Acylation and glycosylation of protoaescigenin contribute to the structure diversification of aescins. After automated annotation of the whole genome and manual revision based on characteristic domains, we identified 21 OSCs, 162 CYP450s, 81 BAHDs, and 173 UGTs in the *A. chinensis* genome (Tables S11-14). All these P450s have been named and are shown in Table S12. To find gene clusters involved in triterpenoid biosynthesis in the *A. chinensis* genome, we searched for biosynthetic gene clusters (BGCs) containing skeleton OSCs and /or tailoring enzymes known to act in such metabolism. This led to discovery of four BGCs containing OSCs and one BGC containing CYP716 genes, known as AcClusters I-V, which are then implicated in triterpenoid synthesis in the *A. chinensis* genome (Figure S9). Syntenic gene analysis of *A. chinensis* paralogs revealed that AcCluster II is physically syntenic with AcCluster I (Fig. 3A, Figure S9). Of these, AcCluster I is located on chromosome 15 and contains two *OSC* homologs (*AcOSC6* and partial *AcOSC9*), six CYPs (*AcCYP716A274*, *AcCYP716A278*, *AcCYP716BX1*, *AcCYP716BX3*, *AcCYP716BX6*, and *AcCYP716A276*), and five *BAHD*s (*AcBAHD1-AcBAHD5*) within a 350-kb region, harboring possible catalytic steps, including 2,3-oxidosqualene cyclization, oxidation, and acylation. RNA-seq data further showed that one OSC (*AcOSC6*), two CYPs (*AcCYP716A278* and *AcCYP716BX1*), and three BADH genes (*AcBAHD1*, *AcBAHD3*, and *AcBAHD5*) were abundantly expressed

in *A. chinensis* seeds, in agreement with high aescin accumulation in the seeds (Fig. 3B, Table S15). AcCluster II is located on chromosome 8 and consists of two CYP450 genes (*AcCYP716A275* and *AcCYP716BX*) and one cellulose synthase-like gene (*AcCSL1*), which putatively participate in protoaescigenin formation via hydroxylation and glucuronidation (Table S15). These three genes also exhibit a seed-specific expression pattern and are co-expressed with the candidate genes localized in AcCluster I (Fig. 3B, Table S15). We reconstructed the phylogenetic relationship of *AcCSL1* with characterized and related genes from different plants available in the literature (40). *AcCSL1* is clustered with a clade of characterized proteins belong to CslM subfamily involved in glucuronidation at the 3-C position of oleanane-type triterpenoids, suggesting a similar function in *A. chinensis* (Figure S10, S11). Further in-depth examination revealed that no glycosyltransferases are found in the surrounding regions of either AcCluster I or AcCluster II. We then performed WGCNA (Figure S11) to retrieve potential genes encoding glycosyltransferases that catalyze the conversion of the glucuronosyl moiety of escin Ia to diglucoside and identified nine putative UDP-glycosyltransferases (Table S16) that may attach glycosyl chains to protoescigenin to form escin Ia.

Biochemical identification of AcOSC6, AcCYP716A278, AcCYP716A275, AcCSL1, **and** AcBAHD3

Heterologous overexpression of full-length *AcOSC6* (β-amyrin synthase, AcBAS) in *Nicotiana benthamiana*, which does not naturally generate β-amyrin, led to the production of β-amyrin in plants as measured by GC-MS/MS and confirmed by comparing mass spectral fragmentation with an authentic standard (Fig. 3C, 3D and Figure S12). The function of AcOSC6 matches its phylogenetic relationships, as it clusters with other OSCs that exhibit such BAS activity (Figure S13). To further explore whether seed-specific *CYP716* genes from AcClusters I and II are able to catalyze the early steps of the escin Ia pathway, we co-expressed *CYP716A278* and *CYP716A275* with *AcOSC6*, as well as *AstHMGR* to increase metabolic flux to terpenoids, using *N. benthamiana* infiltration system. Indeed, both exhibited detectable activity toward β-amyrin, creating new peaks at 13.6 and 14.9 min, respectively. According to the matching MS fragmentation pattern at 14.9 min, the new compound produced by *AstHMGR/AcOSC6/AcCYP716A278* is 21β-hydroxy-β-amyrin, which is known to be produced by *AstHMGR/AsbAS1/GmCYP72A69* (41) (Fig. 3C, 3E and Figure S12). AcCYP716A278 and GmCYP72A69 also were respectively transferred into engineered yeast strains *Y1-20-6* producing β-amyrin (42), leading to identical catalytic activity as observed in *N. benthamiana* (Figure S14). Co-expression of *CYP716A275* and *AcOSC6* in *N. benthamiana* at 13.6 min leads to the production of 16α-hydroxy-β-amyrin, as similarly verified by comparison to the known production by *AsbAS1/GmCYP716Y1* (43) (Fig. 3C, 3F and Figure S12). Several CYP families have been reported to act in triterpenoid oxidation, including CYP93, CYP716, and CYP72 (44–46). In *Medicago truncatula*, CYP716A subfamily members are involved in catalyzing the conversion of β-amyrin to 28-hydroxy-β-amyrin (47–49). Here, we showed that AcCYP716A275 catalyzes the region-specific C-16 oxidation of β-amyrin, consistent with broader function of the CYP716A subfamily in triterpenoid oxidation. To date, two CYP72A subfamily members in the CYP72 clan: oat AsCYP72A475 and soybean GmCYP72A69, have been characterized as C-21 triterpene oxidases based on their activity towards oleanane-type triterpenoids (34, 50). GmCYP72A69 oxidized soyasapogenol B and β-amyrin at position C-21, while AsCYP72A475, as a triterpene C-21 hydroxylase, hydrolyzes 12,13β-

epoxy and 16β-hydroxy-β-amyrin at the C-21 site. However, we found that a CYP716A subfamily member rather than CYP72A subfamily member carries out C-21 hydroxylation for escin Ia biosynthesis in *A. chinensis*. This indicates convergent evolution of C-21 oxidation of oleanane-type triterpenoids by independent evolution within the CYP716(A) and CYP72(A) (sub)families. Identical compounds produced by distant species may originate from different enzymes, but more often arise via the same pathway, regardless of whether these enzymes are homologous (51).

The attachment of a hydrophilic carbohydrate fragment to the triterpenoid skeleton enhances its pharmaceutical properties and water solubility (52–54). Recent reports have identified a series of cellulose synthase-derived glycosyltransferases (CSyGT) that transfer the glucuronic acid moiety to the C-3 position of triterpenoid aglycones in four leguminous plants: *Glycyrrhiza uralensis*, *G. max*, *Lotus japonicus*, and *Spinacia oleracea*(40, 55). To examine whether AcCSL1 functions as a glucuronic acid transferase, *in vivo* substrate-feeding and *in vitro* yeast assays were performed with the previously characterized *GmCSyGT1* as a positive control. The substrate-feeding experiment, carried out with recombinant expression in *N. benthamiana*, demonstrated that both GmCSyGT1 and AcCSL1 acted on protoescigenin, leading to observation of a peak with a molecular ion at *m/z* 681.3 Da, which equals the molecule weight of [protoescigenin + COOH] plus glucuronic acid (176) (Fig. 3C, 3G). Enzymatic assays with recombinant yeast microsomes containing AcCSL1 or GmCSyGT1 obtained the same results (Figure S15, S16).

Pharmacodynamic studies have shown that acylation at C-21 and C-22 positions with diangeloyl groups increased cytotoxicity of aescins (56). A few BAHDs have been identified as major acyltransferases for acylation of thalianol-derived tricyclic in Arabidopsis, tetracyclic cucurbitacins in Cucurbits, and pentacyclic triterpenoids in spinach and *Boswellia* trees (55, 57, 58). To decouple the activity of these transcribed BAHDs (*AcBAHD1*, *AcBAHD3*, and *AcBAHD5*) in AcCluster I towards substrate protoescigenin with acetyl-CoA as the donor, we performed *in vitro* enzyme assays using recombinant proteins from *E. coli* and confirmed that AcBAHD3 could utilize acetyl-CoA as an acyl donor to acetylate protoescigenin, as shown in MS with a molecular ion at *m/z* 593.4 Da, which equals to the molecule weight of $[\text{protoescigenin + COOH}]^-$ (551.3) plus 42 (Fig. 3H). AcBAHD3 then provides a second example of a BAHD family member that targets acylation of pentacyclic triterpenes, supplementing a previous report that the BAHD family member BsAT1 from *Boswellia serrata* was able to catalyze C3a-*O*-acetylation of α-boswellic acids (BA), βBA, and 11-keto-βBA, thus forming all the major C3a-*O*-acetyl-BAs (3-acetyl-aBA, 3-acetyl-bBA, and 3-acetyl-11-keto-bBA) (58). Phylogenetic reconstruction further confirmed that the BsAT1 is the closest orthologs of AcBAHD3 and AcBAHD1, 2 and 5, which are placed with the clade IIIa BAHDs representing ATs that act on distinct acceptor but predominantly utilize acetyl-CoA as donor (Figure S17). Our results demonstrate that the AcCluster I and AcCluster II BGCs contribute to biosynthesis of barrigenol-type triterpenoid (BAT), which we then name the BAT BGC.

Evolution And Organization Of Barrigenol-type Triterpenoid Biosynthetic Gene Clusters

The origin of plant BGCs stems from assembly of genes after gene duplication, neofunctionalization, and genomic relocation, not via horizontal gene transfer from microbes (59). The abundant of genome now available can provide crucial clues into the mechanisms underlying how genes encoding enzymes from a common biosynthetic pathway are assembled into formation of BGC(9). Comparative genomics in combination with functional characterization indicate that early arising BGCs underwent dynamic change of auxiliary enzymes to generate structurally diverse triterpenoids within and between Arabidopsis or between cucurbits (59). Yet, little is known about the evolutionary trajectories of triterpenoid BGCs in other plant clades.

Our functional data demonstrated that β-amyrin synthase (BAS), CYP716, BAHD IIIa, and CSL genes distributed in AcCluster I and AcCluster II are crucial enzymes catalyzing BAT saponin biosynthesis. The mean K_S value of 17 paralogous gene pairs localized in the syntenic blocks of AcCluster I and AcCluster II was 0.27, which is similar to the K_S value of the paralogs in *A. chinensis* more generally (0.24), suggesting that the duplication event of AcCluster I and AcCluster II might originate from the *A. chinensis*-specific WGD event, *Aα* WGD. Syntenic analysis showed that AcCluster I and AcCluster II are conserved among Hippocastanoideae species, including *A. chinensis*, *Acer yangbiense*, and *X. sorbifolium*, in accordance with the high accumulation of BAT saponins in these species (60–62). The *CYP716* genes, which are specifically enriched *A. chinensis*, *Acer yangbiense*, and *X. sorbifolium*, presumably are responsible for the diversity of BAT triterpenoid biosynthesis in these species.

Further large-scale analysis of evolutionary dynamics for this BAT BGC among 21 published plant genomes including early-diverging angiosperm, monocots, and eudicots was conducted to examine how this syntenic region may have evolved (Fig. 4A). An approximately 1-Mb segment of 64 genes (Fig. 4B, Fig. 4C), including one cycloartenol synthase (CAS) gene and one BAHD IIIb gene in the early-diverging angiosperm *Amborella trichopoda*, was traced as the ancestral region. Syntenic regions are also present in monocots (*Oryza sativa* and *Zostera marina*), but these lineages only contain one *CAS* gene, not the BAT biosynthesis related BAS, CYP716, CSL, and BAHD genes. The CYP716A member first appeared in the corresponding region of species within Ranunculales – i.e., *Papaver somniferum*. However, the syntenic region is not present in all the examined Superasterids species. After the split with Superasterids species, almost all the syntenic segments retain at least one *OSC* gene – i.e., other than Cruciferae representatives such as *A. thaliana* and *B. rapa* in Superrosids. Remarkably, a sizeable species-specific tandem duplication of ten complete and ten partial *OSC* genes was found in the conserved syntenic region of the *V. vinifera* genome. Complex insertion, deletion, duplication, and arrangements of *OSC*, *CYP716*, *BAHD*, and *CSL* genes in the syntenic regions of Superrosids species were observed. The *BAS/CYP716/BAHD* gene cluster was always found in the syntenic region of Superrosids species, but dynamic duplication and re-organization occurred (Fig. 4A). We propose that the complete BAT biosynthetic gene cluster *BAS/CYP716/BAHD/CSL* was first assembled in Hippocastanoideae species, – i.e., the BAT producing plants (*A. chinensis*, *Acer yangbiense*, and *X. sorbifolium*). In addition, the BGC underwent further dynamic evolution during speciation, such as the tandem gene duplication of CYP716A

and CYP716BX, and *Aesculus*-specific duplication of AcCluster I and AcCluster II from *Aa* WGD event observed here.

To further examine the evolutionary origins of OSCs, CYPs, CSLs, and BADHs, we generated a series of phylogenetic trees using the maximum likelihood method based on multiple alignments of protein sequences from the forementioned species for synteny analysis (Fig. 4D). The phylogenetic trees based on the syntenic regions of Sapindales species, including *A. chinensis*, *Acer yangbiense*, *X. sorbifolium*, *D. longan*, and *C. clementina*, along with *V. vinifera*, confirmed conservation of BASs, CYP716s, CSLs, and BADH IIIas related to triterpenoid biosynthesis. In addition, gene copy numbers and phylogenetic relationships revealed that the species-specific tandem duplication events of CYP716 family members in *A. chinensis* and *A. yangbiense* led to the expansion of CYP716A and CYP716BX subfamily members. The phylogenetic tree of kingdom-wide identification of cellulose-synthase superfamily genes suggested that *AcCSL* genes and their collinear genes from *A. yangbiense* and *X. sorbifolium* as well as the functionally-verified CSLs, which catalyze the conjugation of glucuronic acid to the triterpenoid backbone, share a close phylogenetic relationship sharing sequence identities of 44.7–55.2%. These results indicated that *CSL* genes localized in the BAT gene clusters of *A. chinensis*, *A. yangbiense*, and *X. sorbifolium* were presumably recruited to catalyze glucuronidation activity before their speciation.

Here, we proposed the evolutionary trajectories for the birth, death, and evolution of the BAT BGC among angiosperms based on their ancestral states reconstruction (Figure S18). (1) Birth in early-diverging angiosperms: The ancestor BGC contains one CAS and one BAHD IIIb in *A. trichopoda*; (2) Death in monocots: The region containing CAS and other non-BAT functional genes is retained, but with loss of BAHD IIIb; (3) Evolution in specific early-diverging eudicots: Evolution of BAS and addition of CYP716; (4) Death in Superasterids: The BGC was lost in this lineage; (5) Assembly of the BAT BGC in Hippocastanoideae: Formation of this functional BGC (BAS/CYP716/CSL/BAHD) leading to BAT production; (6) Re-organization and diversification of the BAT BGC: Dynamic evolution of this BGC in a species-specific manner via tandem duplication and WGD (Fig. 4E).

Coumarin Biosynthesis Pathway And Identification Of Ugt Synthase For Aesculin And Fraxin

Coumarins (1,2-benzopyrones) are produced via the phenylpropanoid pathway and are widely distributed in higher plants (63, 64). These flavonoids are mainly glucoconjugated and accumulated in the vacuoles. The relevant genes identified in this study from the *A. chinensis* genome include seven phenylalanine ammonia-lyases (PALs), three cinnamic acid 4-hydroxylases (C4Hs), five cinnamic acid 3-hydroxylases (C3Hs), four 4-coumaric acid:CoA ligases (4CLs), five feruloyl-CoA 6'-hydroxylase (F6'Hs), three caffeic-acid/5-hydroxyferulic-acid *O*-methyltransferases (COMTs), four caffeoyl-coenzyme A *O*-methyltransferases (CcoAOMTs), and five scopoletin 8-hydroxylases (S8Hs) (Figure S19, Table S17).

Nineteen putative AcUGTs were identified from phylogenetic analysis based on homology to the known UGTs involved in the flavonoid and coumarin biosynthesis (Table S19). These were cloned and expressed

in *E. coli* to produce recombinant proteins to investigate their enzymatic activities using esculetin and fraxetin as acceptors and UDP-glucose as a donor substrates. AcUGT71A47 and AcUGT92G7 were found to utilize fraxetin as an acceptor substrate (Fig. 5A, Figure S19), whereas AcUGT84A56 and AcUGT92G7 were found to utilize esculetin (Fig. 5B, Figure S19). MS data revealed that all these enzymes could catalyze the addition of one glucose, as indicated by increases in the molecular weight of products relatives to the corresponding aglycones by 162 Da. The enzymatic products of AcUGTs with esculetin or fraxetin as the substrates were confirmed to be aesculin or fraxin, respectively, by comparing mass spectra and retention time to that of authentic standards.

De novo **production of aesculin in** E. coli

Feeding experiments were carried out to determine the optimal glucosyltransferase for the conversion of *p*-coumaric acid to aesculin (Table S19, S20). First, we generated strains XC-1 and XC-2*E* by transforming plasmids pCS-84 (for expression of UGT84A56) and pCS-92 (for expression of UGT92G7) into *E. coli* BW25113, respectively. These were then used in feeding experiments with esculetin. After 48 h, XC-1 produced 23.0 ± 0.2 mg/L of aesculin, but XC-2 produced only 9 ± 1 mg/L of aesculin (Fig. 5E), suggesting that UGT84A56 exhibits a higher catalytic efficiency than UGT92G7 in *E. coli*. Thus, UGT84A56 was used for aesculin biosynthesis in subsequent experiments. We constructed strains XC-3 and XC-4 by transforming plasmids pZE-F6H-4CL, for expression of F6H and 4CL both from *Arabidopsis thaliana*, and pCS-84-HpaBC, for expression of UGT84A56 and HpaBC from *E coli*, into BW25113, respectively. These were then used in feeding experiments with *p*-coumaric acid. After 48 h, strain XC-3 produced 103 ± 2 mg/L of aesculin (Fig. 5F).

After achieving biosynthesis of aesculin from *p*-coumaric acid, we proceeded with expressing the entire aesculin pathway in *E. coli*. In previous work, over-expressing *ppsA* and *tkta*, as well as the feedback inhibition resistant variants *tyrA* and *aroG*, improved tyrosine production. To ensure efficient expression, we constructed plasmid pET-648T, which contains the *UGT84A56* gene and three other key genes *F6'H*, *4CL1*, and *TAL* from *Rhodotorula glutinis*. First, we generated strain XC-4 by transforming plasmids pCS-TPTA-HpaBC and pET-648T into *E. coli* BW25113 and used it for *de novo* production of aesculin as a preliminary test. However, the expected intermediates of caffeic acid and esculetin were not detected (data not shown), yet the cell culture turned black. This may be because HpaBC can convert tyrosine to L-Dopa, which is unstable and readily oxidized to melanin, and the precursor PEP or chorismite is insufficient. Thus, to enhance PEP flux and avoid inhibiting cell growth, we engineered strain BW-1 by deleting genes *pykA/pykF* in *E. coli* BW25113 and further generated strain XC-5 by transforming plasmids pCS-TPTA-HpaBC and pET-648T into BW-1. Notably, 72 h after induction, 36 ± 5 mg/L of aesculin and its intermediate product esculetin were produced (Fig. 5G, Figure S20).

Conclusion

In summary, through the genome sequencing and collinearity analysis of *A. chinensis*, we identified two WGD-originated BGCs related to the biosynthesis of barrigenol-type triterpenoid sponins for the first time.

Furthermore, one OSC, two CYP450s, one CSL, and one BAHD genes localized in these two BGCs were experimentally verified to act in the aescin biosynthesis. The evolutionary trajectories of BGCs among angiosperms revealed that the functional BGC leading to BAT biosynthesis was assembled in Hippocastanoideae species, and dynamic evolution of species-specific via tandem duplication and WGD after the speciation of *A. chinensis*. In addition, we identified the UGTs related to aesculin biosynthesis and constructed the metabolic engineering system to yield aesculin in *E. coli*. Our analysis provides novel insights in the formation and evolution of triterpenoid biosynthetic gene clusters among angiosperms.

Methods

Plant materials for sequencing

A. chinensis plants were collected from the institute of medicinal plant development (IMPLAD), the Chinese Academy of Medical Sciences (CAMS), China. *N. benthamiana* was grown for about four weeks with an 18h light/ 6 h dark in a greenhouse at the National TCM gene bank, institute of Chinese Materia Medica, China academy of Chinese Medical Sciences. DNA was extracted from the leaves of one *A. chinensis* plant, using a kit plant DNA extraction protocol (Tiangen co. Ltd., http://www.tiangen.com). The quality and quantity of the isolated DNA and RNA were assessed by electrophoresis on a 0.8% agarose gel and Epoch (Bio-Tek Instruments, https://www.biotek.com).

Estimation Of The Genome Size Using Flow Cytometry

Both flow cytometry was used to estimate the genome size of the *A. chinensis* before genome sequencing. Flow cytometry analysis for the measurement of nuclear DNA content was completed as previously described (65).

Genome Sequencing And Assembly

Short read pair-end libraries were constructed according to manufacturer's instructions, and sequenced on an Illumina Hiseq X-ten (Illumina, San Diego, CA. USA). *k*-mer analysis were For long-read sequencing, ONT twenty-kilobase insert libraries were prepared using high-quality *A. chinensis* genomic DNA with the standard ONT library preparation protocols, and sequenced using the ONT GridION X5 (v.9.4.1; Oxford Nanopore Technologies) platform (http://nanoporetech.com). Base calling of the raw ONT reads was conducted using Oxford Nanopore base caller Guppy (v.1.8.5; Oxford Nanopore Technologies) with default parameters. The ONT reads were corrected, trimmed, and assembled into genomic contigs using CANU (v.1.5) (Koren et al., 2017), Smartdenovo (https://omictools.com/smartdenovo-tool)(Schmidt et al., 2017). We utilized Illumina short reads to polish the assembled contigs three times. The completeness of the genome assembly was estimated using Benchmarking Universal Single-Copy Orthologs (BUSCO; v3)with Embryophya odb 10 dataset.

Rna Sequencing And Related Transcript Analysis

For RNA-seq, *A. chinensis* RNA samples were prepared from young leaves, branches, open flowers, exocarp and seeds using an RNeasy Plant Mini Kit (Qiagen, https://www.qiagen.com). RNAseq library preparation was done using a TruSeq RNA Library Prep Kit v.2 (Illumina, San Diego, CA, USA). Good quality libraries were sequenced using the NextSeq 500 platform. Fragments per kilobase of exon model per million reads mapped (FPKM) values for RNA-seq were counted with HISAT2 (v.2.0.5) and CUFFLINKS (v.2.2.1) after raw data was trimmed and cleaned and assembled using Trimmomatic (v.0.39) and Trinity (v.2.8.3). The WGCNA network were inferred from the FPKM values in different tissues. WGCNA module eigengene values were calculated and associated with tissues.

Hi-c Assistant Contig Clustering

A Hi-C library of young *A. chinensis* leaves was prepared by Annoroad Genomics (http://en.annoroad.com), following the standard procedure (Lieberman-Aiden et al., 2009). The Hi-C sequencing data was aligned to the assembled contigs by BWA-MEM, and then the contigs were clustered onto chromosomes with LACHESIS (http://shendurelab.github.io/LACHESIS/).

Phylogenetic Tree And Evolutionary Analysis

A maximum-likelihood (ML) phylogenetic tree of single-copy genes, which were clustered into orthologous groups, from the genomes of *A. chinensis* with *Malus domestica*, *Cucumis sativus*, *Glycine max*, *Populus trichocarpa*, *Citrus clementina*, *D. longan*, *X. sorbifolium*, *Brassica rapa*, *Arabidopsis thaliana*, *Theobroma cacao*, *Gossypium raimondiim*, and *Vitis vinifera*was carried out using RAXML 8.2.10 (Stamatakis, 2014) using the JTT + G + I substitution model for amino acids with 1000 bootstrap replicates. The divergence time of selected species was calculated using the program R8S, according to fossil divergence time points for the specific Sapindaceae and Rutaceae (21.8–50.4 MYA) split obtained from TimeTree (http://www.timetree.org). Then CAFE 2.1 was utilized for identifying contraction and expansion analysis of these gene families following species divergence prediction under a probalilistic graphical model(66). CoGe suite (www.genomeevolution.org) was applied for performing genome synteny analysis. Most of the cellulose synthase-like enzyme protein sequences used for phylogenetic analysis were from a previous study (40), with additional inclusion of SOAP5 (67) and AcCS1 reported in this study. A Neighbour joining phylogenetic tree was built using MEGA 6 after sequence alignment (Jones-Taylor-Thornton method with 1000 bootstraps)(68).

Genome Annotation

We searched repetitive sequences in the *A. chinensis* genome by *de novo*-based and homology-based strategies using RepeatModeler (v1.0.9) software. Two de novo repeat finding programs, LTR_Finder and LTR_retriever was employed for identifying and classifying repeat elements. *De novo* transcript assembly

was done using Trinity(69) (version 2.2.0) with default parameters, and peptide sequences were predicted by TransDecoder (https://github.com/ TransDecoder; version 2.1.0). MAKER annotation pipeline was used for *ab initio* predictions of protein-coding genes in the masked *A. chinensis* genome. We searched funcSwiss Prot, NR, InterPro, Pfam and KEGG using the MAKER annotation pipeline (version 2.31.9)(70). P450 annotation and nomenclature was carried out by Prof. David Nelson.

Identification Of Gene Families Related To Triterpenoid And Coumarin Biosynthesis

BLASTP-based and HMMER annotation in combination were used to find high-confidence proteins associated with triterpenoid and coumarin biosynthesis in *A. chinensis*. Protein sequences from characterized proteins in triterpenoid and coumarin were used for identifying candidate genes by BLASTP (E-value < 1e-5, identity > 50%, and coverage > 50% in the *A. chinensis* genome. Determination of candidate genes were verified by HMMER; only proteins with the representative domains were retained.

Gene Cluster Involved In Triterpenoid Biosynthesis Identification

For triterpenoid gene cluster analysis, we used http://plantismash.secondarymetabolites.org/ online pipeline and manual search. To assign gene clusters to triterpenoid biosynthesis, we surveyed if these contained one or more *OSC* genes, and annotated the physically adherent 20 genes surrounding them. The obtained gene cluster containing OSC6 was mapped to self-genome and four other genomes (*Citrus clementina*, *D. longan*, *X. sorbifolium* and *Acer yangbiense*) in order to uncover collinear segments using MCSCANX (71).

Maldi Ms Imaging

Fresh *Aesculus chinensis* fruit were harvested and immediately embedded in 10% gelatin (W/V) solutions. Initially, tissues were kept in Tissue-Tek cryomolds (25 × 20 × 5 mm) and the gelatin solution was poured over them to embed the tissues; thereafter, the molds were transferred to a -80°C freezer for 1h to form a solid block. For cryo-sectioning, the sample blocks were directly fixed on the sample holder of a cryostat (Leica, Germany), using deionized water as the adhesive. Sections of 16 μm-thickness were obtained at -20°C and thaw-mounted on indium tin oxide-coated glass slides for immediate imaging measurements. To avoid condensation, the tissue sections were dehydrated in a vacuum desiccator for approximate 10 min prior to matrix application. A Zeiss Axio M2 microscope (Zeiss, Germany) was used to obtain optical images of the sections. A laboratory-constructed automated pneumatic-assisted matrix application system was used for the uniform application of MALDI matrix solution. The matrix application system and coating procedure were as previously described with some modifications (Li, Comi, Si, Dunham & Sweedler, 2016a). Briefly, 50 mg/mL DHB dissolved in ACN: water (0.1% TFA) (7: 3, V/V) was applied for the positive mode MALDI experiments (TransMIT GmbH, Giessen, Germany), which

was operated a t atmospheric pressure and coupled to a Q Exactive HF Orbitrap mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). For homogenous deposition onto the fruit samples, the nebulizer was held 3 cm above the sample and oscillated over the plate 100 times. The flow rate was set to 6–8 mL/h and gas pressure to 50 psi, to deliver and nebulize the matrix solution, respectively.

MALDI imaging was performed using a Q Exactive Orbitrap mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) equipped with a MALDI source (TransMIT GmbH, Giessen, Germany). Single-scan spectra consisted of 100 accumulated laser shots at 1 kHz with laser focus set to “small”. MALDI images were acquired at a 50-µm spatial resolution. Instrument calibration was performed using the ESI inlet with sodium TFA clusters ranging from m/z 158.9640–1926.6365 and a quadratic calibration of the clusters. Instrument calibration was performed using the ESI portion of the instrument with NaTFA clusters and a quadratic calibration of the clusters. Data was analyzed using SMALDIControl software package (TransMIT GmbH, Giessen, Germany).

Lc-ms/ms And Gc-ms/ms Analysis

All tissues were freeze-dried under vacuum, each sample was weighed to 100mg and combined for subsequent solvent extraction; 10 mL of methanol: water (7: 3, V/V) was added into the 50mL centrifuge tubes, followed by 30 min of sonication (8891 ultrasonic cleaner, Billerica, MA). The extracts were centrifuged at 10,000 rpm for 5 min and the supernatants collected. The extract solutions were then used for subsequent LC-MS/MS analysis. LC-MS/MS for quantitation of aescins and coumarins were performed using a 6470 Triple Quadrupole mass spectrometer (Agilent Technologies, Santa Clara, CA USA) connected to an Advance 1290 UHPLC system (Agilent Technologies, Santa Clara, CA USA). Analytes were separated on an Agilent Eclipse Plus C18 column (RRHD 1.8µm, 2.1×50mm). Mobile phase A was 0.1% acetic acid and phase B was H_2O containing 0.1% acetic acid. The gradient program was conducted as follows: 0–25 min, A: B = 68:32. The injection volume was 1 µL. The flow rate was 0.30 mL/min and column temperature was 30°C. The mass spectrometer was operated in positive or negative ion mode, with sheath gas temperature at 250°C, gas flow at 11.0 L/min, and nebulizer gas at 40 psi. The capillary voltage was set at 4000 V, nozzle voltage to 500 V, and delta EMV to 200 V. Metabolites were detected in multiple reaction monitoring (MRM) mode, where two precursor-product ion MRM transitions were selected for each compound. Data was acquired and analyzed using MassHunter (Agilent Technologies, Santa Clara, CA USA) to quantify all seven metabolites for which standards were available.

For the identification of β-amyrin and corresponding hydroxylated products, GC-QQQ-MS analyses were performed on an Agilent Technology 7890 GC, coupled with a 7000C Triple Quadrupole MS (Agilent Technologies, Santa Clara, CA USA). One µL of the sample was injected into splitless mode with an injector temperature of 250°C. The column was DB-5 ms (Agilent Technologies, Santa Clara, CA USA), 30 m × 0.25 mm i.d. × 0.25-µm flm thickness, were connected by a purged ultimate union (PUU) to provide sample separation. Helium was utilized as the carrier gas at a constant flow at a flow rate of 1.2 mL/min. The temperature was set as follows: 2 min of 170°C followed by heating the column to 300°C at

20°C/min and maintaining 300°C for 11.5 min (total run time 20 min). The MS transfer line and ion source temperatures were set to 300°C and 280°C, respectively. Full mass spectra were generated by scanning within the m/z range of 50–700 amu with solvent delay of 8 min. The identification of β-amyrin, 16α-hydroxy-β-amyrin and 21β-hydroxy-β-amyrin was conducted by comparing both the retention time and mass spectra with the authentic standards β-amyrin, and the catalytic reaction product of BfCYP716Y1 and GmCYP72A69, respectively.

Gene Cloning

Total RNA was extracted using an EasyPure® Plant RNA Kit (TransGen, Beijing, China). First strand cDNA was synthesized from total RNA using the *TransScript*® Uni One-Step gDNA Removal and cDNA Synthesis Super Mix cDNA synthesis kit (TransGen, Beijing, China) according to the manufacturer's protocol. Full-length candidate genes were amplified by polymerase chain reaction (PCR) using KOD-Plus-Neo DNA polymerase (TOYOBO Bio, Japan) with specific primers (Supplementary Table) from seed first-strand cDNA. The amplified *AcOSC6, AcCYP716A275, AcCYP716A278, AcCSL1*, *AcBAHD3* and 19 *AcUGT* candidates including ORFs were cloned into the pEASY-blunt vector (TransGen, Beijing, China) and sequenced (Sangon Biotech, Shanghai, China). Previously characterized sequences including *AtHMGR* (KY284573), *GmCYP72A69* (NM_001354946), *BfCYP716Y1* (KC963423.1) and *GmCSyGT1* (LC500227.1) were synthesized (GenScript Biotech).

Functional characterization of AcOSC6 **and** CYP450 **genes in** N. benthamiana

AcOSC6 and CYP450 genes were firstly recombined into the donor vector pDONR207, then cloned into pEAQ-HT-DEST1 using LR clonase II enzyme (Thermo Fisher Scientific). pEAQ-HT-DEST constructs were electro-transformed into *Agrobacterium tumafaciens* (EHA105) strain. Positive clones were grown on solid LB plates under 50 ug.ml^{-1} Kanamycin and Rifampicin at 28℃. Subsequently, single colonies were grown in 10 mL LB liquid medium with antibiotics for 12 hours, and centrifuged at 12000rpm for 5min, with supernatant removed. The *A. tumefaciens* pellets were resuspended with 100mM MES and 100 buffers to reach OD600 = 0.2. Transformed strains carrying the candidate constructs were infiltrated solely or equally in combination with HMG CoA-reductase from *A. thaliana* into *N. benthamiana* using a syringe (34, 72). Infiltrated *N. benthamiana* leaves were collected after 4–5 days, and stored at -80°C for metabolite analysis. The infiltrated *N. benthamiana* was lyophilized, grinded and weighed 10 mg. The powder was extracted with 1 mL saponification solution (EtOH:H2O:KOH 9:1:1 v: v:w) at 65°C for 2 h with intermittent agitation. The extracts were then evaporated to dryness and reextracted with 500 µL ethyl acetate by shaking for 2 mimutes. After adding 500 µL of H_2O, samples were vortexed (2 × 10 s) and 100 µL of the upper organic phase was transferred to a fresh vial and dried. Trimethylsilylation was carried out with 50 µl of N-methyl-N-(trimethylsilyl)-trifluoroacetamide (Sigma-Aldrich). The product of trimethylsilylation was analyzed using GC-MS.

Functional characterization of CYP450 **genes in engineered yeast**

Yeast strain Y0 was derived from strain Cenpk2-1D, which was purchased from EUROSCARF, by knocking out BTS1 and ERG27 genes using CRISPR/Cas9. Gene cassettes of Ppgk-tHMG-Tadh1, Ptef1-ERG1-Tpgk, Ptdh3-ERG20 + ERG9-TCYC1 and PpgK-β-AS-Tadh1 were amplified from the corresponding plasmids. Then, these cassettes and URA3 marker gene were integrated into delta site of strain Y0. The strain producing the highest β-amyrin was designed as strain Y1-20-6. The ORF of *AcCYP716A275* and *AcCYP716A278* was subcloned into yeast expression vector pESC-HIS and recombination construct were co-transformed in yeast strain Y1-20-6. The metabolite extract was followed as above for GC-MS analysis.

Yeast-feeding Assay And Microsomes For Glucuronosylation Of Triterpenoid Aglycones

The yeast expression vector pESC-HIS constructs with *AcCSL1* or an empty vector were transformed into yeast *S. cerevisiae* WAT11. The recombinant yeast strains were first grown in 20 ml SD-His medium containing 20 g L^{-1} glucose at 30℃ for 24 h. To ensure complete digestion of glucose, 200 µL of strain solution was used with urine glucose test strips. After the glucose test, the yeast cells were harvested by centrifugation at 1000 g for 5 min and washed twice with ddH_2O. SD-His medium with 20 g L^{-1} galactose was used to resuspend the strains. 1 mL strains were used for yeast *in vivo* reaction. The substrate protoescigenin was supplemented at 100 µM into the cultures. After 16 h, ethyl acetate was used to stop the reaction, and the supernatant was concentrated at low pressure after ultrasonic centrifugation until ethyl acetate was completely volatilized, and the pellet was dissolved in 200 µL methanol and stored at 4℃ for further UPLC-MS analysis. Microsomal proteins were isolated and were dissolved in protein storage buffer (20% [v/v] glycerol, 100 mM Tris-HCl [pH 7.5]). The enzymatic activity of AcCSL1 were assayed in 100 µL of reaction volume, which contained 35 mM Tris-HCl (pH 7.4), 1 mM dithiothreitol, about 30 µg of microsomal proteins, 100 µM protoescigenin, and 1 mM UDPGA. The assays were incubated for 1 h at 30℃. The reactions were stopped by addition of 100 µL methanol, followed by brief vortex. Finally, the extracts were centrifuged for 10 min at 14,000g and analyzed by UPLC–MS.

Enzyme Assay Of Acugts And Acbahds

Full length cDNA sequences of the UGT and BAHD genes from *A. chinensis* were obtained using KOD Plus-Neo according to the kit instructions (TOYOBO, Shanghai, China). PCR products of UGT genes were purified with an EasyPure Quick Gel Extraction Kit (TransGen, Beijing, China), then digested with the corresponding restriction enzyme and ligated into similarly pre-digested pMAL-c2x vectors (New England BioLabs, USA). The pMAL expression plasmids carrying the UGT and BAHD genes were transformed into *E. coli Novablue* competent cells (Novagen, Schwalbach, Germany). Positive colonies were cultured overnight at a ratio of 1:100 to the medium at 37°C (200 rpm). The overnight culture solution was cultured again at 1:100 to the medium at 37°C (180 rpm), adding 0.3 mM isopropyl β-D-Thiogalactoside to induce protein expression until the absorbance of the bacterial solution at 600nm reached 0.6. After

being cultured for more than 16 h at 16°C (150rpm), the cell precipitate was harvested by centrifugation at 4°C and stored at − 80°C immediately until purification. The recombinant MBP-Fusion proteins were purified using maltose resin according to manufacturer's instructions (New England Biolabs, USA).In order to analyze the activity of nineteen recombinant UGTs *in vitro*, 100ul final reactions were adopted consisting of 10 mM DTT, 50mM Tris-HCl (pH 7.0), 0.1mM substrates, 1mM UDP-glucose as sugar donor, and recombinant proteins (5–10 μg). Reaction samples lacking recombinant proteins were used as blank controls. Reactions were stopped by mixing the reaction solution with 100ul 100% methanol after incubating for 1h at 37℃. Products were analyzed by HPLC after centrifugation at 12 000 rpm for 10 min and filtration through a 0.22μm micropore filter. The K_M and V_{max} of AcUGTs were measured using 1mM UDP-Glc as the sugar donor, esculetin or fraxetin as sugar acceptor; substrate concentration set in the range of 0-400μM as acceptor in Tris–HCl buffer (pH 7.0). All the kinetic samples were incubated at 37°C for 1h and repeated in triplicate. Subsequently, the related glucoside products were further identified by LC-MS, and the kinetic parameters K_M and K_{cat} were calculated by the Hyper32 program. Concerning acetylation of protoescigenin, 1 mM acetyl-CoA, 100 μM protoescigenin, and about 10 μg purified protein of AcBAHD3 in 100 μl Tris-HCl solution (50 mM, pH 8.0) at 28℃ for 2 min. The reaction was terminated with equal volume menthol, followed by brief vortex. Finally, the extracts were centrifuged for 10 min at 14,000g and analyzed by UPLC–MS.

Plasmid construction for heterologous expression

The derivative plasmids pZE12-luc (high copy), pCS27 (medium copy), and pETDuet-1 (high copy) were used for pathway assembly. Plasmids pCS-TPTA-HpaBC were constructed as described in our previous study(73). Genes *F6'H*, *AT4CL1* were amplified from *Arabidopsis thaliana* cDNA(74). Plasmid pCS-TAL was constructed as described in our previous study. Gene *RgTAL* was amplified using pCS-TAL as the template. Genes *UGT84A57* and *UGT92G7* were amplified from cDNA of *Aesculus chinensis*. The $P_L$lacO1 promoter of pZE12-luc and pCS27 was changed to BBa_J23100 (http://parts.igem.org/Part:BBa_J23100) via overlap PCR (REF). The T7 promoter and terminator of pETDuet-1 was changed to $P_L$lacO1 and T1 by inserting *$P_L$lacO1* and *T1* using *Sph*I and *Xho*I. Plasmids pCS-84 and pCS-92 were constructed by inserting *UGT84A57* and *UGT92G7* into pCS27 using *Kpn*I and *Apa*I. Plasmids pCS-84-HpaBC and pCS-92-HpaBC were constructed by inserting *HpaBC* into pCS-84 and pCS-92 using *Apa*I and *Bam*HI. Plasmid pZE-F6'H-4CL was constructed by inserting *F6'H* and *AT4CL1* into pZE12-luc using *Kpn*I, *Sph*I and *Bam*HI. Plasmid pET-648T was constructed by inserting *F6'H*, *AT4CL1*, *UGT84A57* and *RgTAL* into pETDuet-1 using *Kpn*I, *Sph*I, *Bam*HI, *Not*I and *Xho*I. The information on the plasmids and the strains used are summarized in Table S19. Primers used in this study are listed in Table S18, S20.

Feeding and De novo **Experiments**

Feeding experiments were conducted to examine the production of aesculin from esculetin and *p*-coumaric acid. Single fresh colonies were inoculated into 4 mL of LB media containing appropriate antibiotics and grown overnight at 37 °C. Subsequently, overnight cultures (1 mL) were inoculated into 50

mL of LB media and appropriate antibiotics, grown at 37 ℃ and 200 r.p.m. After 6h, the strains XC-1, XC-2 and XC-3, XC-4 were fed with esculetin and *p*-Coumaric acid (final concentration 1mM), and glucose (final concentration 10 g/L), respectively. Samples were taken at regular time intervals. Optical densities at 600 nm (OD_{600}) were measured for cell growth and the concentrations of the products were analyzed by HPLC. The dried products were dissolved with methanol and analyzed by ESI-MS.

For *de novo* production of aesculin, 1mL of overnight seed cultures were inoculated into 50 mL of M9Y media and appropriate antibiotics. The strains XC-5 and XC-6 were cultivated at 37 ℃ and 200 r.p.m for 3h and then induced with 0.5 mM isopropyl-β-D-thiogalactoside (IPTG). The induced cultures continued to grow at 37°C and 200 r.p.m. Samples were taken every 12 h. The cell densities (OD_{600}) were measured, and the concentrations of the products and the intermediates were analyzed by HPLC.

Hplc Analysis For Heterologous Expression

p-Coumaric acid and caffeic acid were purchased from Aladdin Chemical Industry, and esculetin, aesculin were purchased from Macklin Reagent, and used as the standards. Both the standards and samples were centrifuged at 12000 r.p.m for 10 min and filtered through 0.22μm film and analyzed by HPLC (HITACHI) equipped with a reverse-phase Diamonsil C18 column (Diamonsil 5μm, 250 × 4.6 mm) and UV − vis detector. Solvent A was methanol and solvent B was water with 0.1% formic acid. The column temperature was set to 40°C. The following gradient was used at a flow rate of 1 mL/min:95–50% solvent B for 20 min, 50–0% solvent B for 2 min, 0–95% solvent B for 2 min, and 95% solvent B for an additional 6 min. Quantification was based on the peak areas at specific wavelengths (310 nm for *p*-coumaric acid, 324 nm for caffeic acid, 345 nm for esculetin, and 334 nm for aesculin).

Declarations

Acknowledgement

This study was supported by scientific and technological innovation project of China Academy of Chinese Medical Sciences (CI2021A04008, CI2021A04117); the National Natural Science Foundation of China (82274037); National Key R&D Program of China (2019YFC1711100); Fundamental Research Funds for the Central public welfare research institutes (ZZ13-YQ-097); the Opening Project of State Key Laboratory of Tree Genetics and Breeding; the Key Laboratory of Plant Germplasm enhancement and Specialty Agriculture, Wuhan Botanical Garden, Chinese Academy of Sciences; and State Key Laboratory of Natural and Biomimetic Drugs.

Author contribution statement

W. S., Z.C.X. and S.L.C. conceived and designed the research. Q.G.Y., H.H.W., C.X., H.Y., and C.X. performed the experiments. X.X.M. analyzed the results. D.R.N named all of the Aesculus genome P450 sequences.

W. S., Z.C.X., X.X.S., Q.G.Y., H.H.W. wrote the manuscript. Z.G.H., C.X.W., Z.Y.X., D.R.N reviewed and edited the manuscript. All authors have read and approved the manuscript for publication.

Data availability

The raw data of genome and transcriptome sequencing reported in this paper have been deposited in the Genome Sequence Archive in BIG Data Center, Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, under BioProject number PRJCA004913 that are publicly accessible at http://bigd.big.ac.cn/gsa. The assembled genome and gene structures of *A. chinensis* have been deposited in the Figshare (https://doi.org/10.6084/m9.figshare.21350865).

References

1. T. Winzer, *et al.*, A papaver somniferum 10-gene cluster for synthesis of the anticancer alkaloid noscapine. Science (80). **336**, 1704–1708 (2012).
2. C. Zhan, *et al.*, Selection of a subspecies-specific diterpene gene cluster implicated in rice disease resistance. Nat. Plants (2020).
3. X. Qi, *et al.*, A gene cluster for secondary metabolism in oat: Implications for the evolution of metabolic diversity in plants. *Proc. Natl. Acad. Sci. U. S. A.* (2004).
4. Y. Shang, *et al.*, Biosynthesis, regulation, and domestication of bitterness in cucumber. Science (80). (2014).
5. C. Song, *et al.*, Taxus yunnanensis genome offers insights into gymnosperm phylogeny and taxol production. Commun. Biol. **4**, 4–11 (2021).
6. X. Xiong, *et al.*, The Taxus genome provides insights into paclitaxel biosynthesis. Nat. Plants **7**, 1026–1036 (2021).
7. W. Sun, Z. Xu, C. Song, S. Chen, Herbgenomics: Decipher molecular genetics of medicinal plants. *Innov.*, 100322 (2022).
8. Z. Liu, *et al.*, Formation and diversification of a paradigm biosynthetic gene cluster in plants. Nat. Commun. **11**, 1–11 (2020).
9. X. Zhou, Z. Liu, Unlocking plant metabolic diversity: A (pan)-genomic view. Plant Commun. **3**, 100300 (2022).
10. M. Itkin, Biosynthesis of Antinutritional. **175** (2014).
11. Y. Zhou, *et al.*, Convergence and divergence of bitterness biosynthesis and regulation in Cucurbitaceae. Nat. Plants **2**, 1–8 (2016).
12. X. Yang, *et al.*, Three chromosome-scale Papaver genomes reveal punctuated patchwork evolution of the morphinan and noscapine biosynthesis pathway. Nat. Commun. **12**, 1–14 (2021).
13. D. Wu, *et al.*, Lateral transfers lead to the birth of momilactone biosynthetic gene clusters in grass. Plant J. **111**, 1354–1367 (2022).

14. Z. Xu, *et al.*, Tandem gene duplications drive divergent evolution of caffeine and crocin biosynthetic pathways in plants. BMC Biol. (2020) https:/doi.org/10.1186/s12915-020-00795-3.
15. H. W. Nützmann, C. Scazzocchio, A. Osbourn, Metabolic gene clusters in eukaryotes. Annu. Rev. Genet. **52**, 159–183 (2018).
16. H. W. Nützmann, A. Huang, A. Osbourn, Plant metabolic clusters – from genetics to genomics. New Phytol. **211**, 771–789 (2016).
17. G. Polturak, Z. Liu, A. Osbourn, New and emerging concepts in the evolution and function of plant biosynthetic gene clusters. Curr. Opin. Green Sustain. Chem. **33**, 100568 (2022).
18. Z. Zhang, S. Li, X.-Y. Lian, An Overview of Genus Aesculus L.: Ethnobotany, Phytochemistry, and Pharmacological Activities. Pharm. Crop. (2014).
19. F. Wei, *et al.*, Antiinflammatory triterpenoid saponins from the seeds of Aesculus chinensis. Chem. Pharm. Bull. (2004).
20. C. R. Sirtori, Aescin: Pharmacology, pharmacokinetics and therapeutic profile. Pharmacol. Res. (2001).
21. Z. Zhang, S. Li, Cytotoxic triterpenoid saponins from the fruits of Aesculus pavia L. Phytochemistry (2007).
22. R. M. Facino, *et al.*, Mass spectrometric characterization of horse chestnut saponins (Escin). Org. Mass Spectrom. (1991).
23. E. Colson, *et al.*, Discrimination of Regioisomeric and Stereoisomeric Saponins from Aesculus hippocastanum Seeds by Ion Mobility Mass Spectrometry. J. Am. Soc. Mass Spectrom. (2019).
24. J. C. Anesth, C. Schmidt, Comparison of leg compression stocking and oral horse-chestnut seed extract therapy in patients with chronic venous insufficiency. 292–294 (1995).
25. C. R. Sirtori, Aescin: Pharmacology, pharmacokinetics and therapeutic profile. Pharmacol. Res. **44**, 183–193 (2001).
26. I. A.-L. Persson, K. Persson, Horse chestnut Aesculus hippocastanum L. (Abstract). in *Recent Prog. Med. Plants*, (2010).
27. A. Owczarek, M. A. Olszewska, Development and validation of UHPLC-PDA method for simultaneous determination of bioactive polyphenols of horse-chestnut bark using numerical optimization with MS Excel Solver. J. Pharm. Biomed. Anal. (2020).
28. F. He, W. Q. Song, G. Q. Li, Z. W. Peng, S. L. Yin, Clinical effects of Augentropfen Stulln Mono eye-drops for asthenopia after laser in situ keratomileusis. Int. J. Ophthalmol. (2010).
29. C. C. Xu, Y. Y. Xu, W. J. Xie, Q. M. Wang, M. X. Shen, Study on the effects of Augentropfen Stulln Mono eye-drops on the tear dynamics of dry eyes after excimer laser surgery. Int. Eye Sci. (2014).
30. G. A. Wu, *et al.*, Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. Nat. Biotechnol. **32**, 656–662 (2014).
31. Y. Lin, *et al.*, Genome-wide sequencing of longan (Dimocarpus longan Lour.) provides insights into molecular basis of its polyphenol-rich characteristics. Gigascience **6**, 1–14 (2017).

32. L. Carretero-Paulet, Y. Van de Peer, The evolutionary conundrum of whole-genome duplication. Am. J. Bot. **107**, 1101–1105 (2020).
33. L. Tu, *et al.*, Genome of Tripterygium wilfordii and identification of cytochrome P450 involved in triptolide biosynthesis. Nat. Commun. **11** (2020).
34. A. Leveau, *et al.*, Towards take-all control: a C-21β oxidase required for acylation of triterpene defence compounds in oat. New Phytol. **221**, 1544–1555 (2019).
35. T. Unver, *et al.*, Genome of wild olive and the evolution of oil biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E9413–E9422 (2017).
36. W. Su, *et al.*, Polyploidy underlies co-option and diversification of biosynthetic triterpene pathways in the apple tribe. *Proc. Natl. Acad. Sci. U. S. A.* **118**, 1–11 (2021).
37. D. Zhao, *et al.*, De novo genome assembly of Camptotheca acuminata, a natural source of the anti-cancer compound camptothecin. Gigascience (2017).
38. K. Haralampidis, M. Trojanowska, A. E. Osbourn, Biosynthesis of triterpenoid saponins in plants. Adv. Biochem. Eng. Biotechnol. (2002).
39. R. Thimmappa, K. Geisler, T. Louveau, P. O'Maille, A. Osbourn, Triterpene biosynthesis in plants. Annu. Rev. Plant Biol. (2014).
40. S. Y. Chung, *et al.*, A cellulose synthase-derived enzyme catalyses 3-O-glucuronosylation in saponin biosynthesis. Nat. Commun. (2020).
41. R. Yano, *et al.*, Metabolic switching of astringent and beneficial triterpenoid saponins in soybean is achieved by a loss-of-function mutation in cytochrome P450 72A69. Plant J. **89**, 527–539 (2017).
42. C. Wang, *et al.*, Efficient production of glycyrrhetinic acid in metabolically engineered Saccharomyces cerevisiae via an integrated strategy. Microb. Cell Fact. **18**, 1–15 (2019).
43. T. Moses, *et al.*, Combinatorial biosynthesis of sapogenins and saponins in Saccharomyces cerevisiae using a C-16α hydroxylase from Bupleurum falcatum. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 1634–1639 (2014).
44. S. Sommerwerk, *et al.*, The ancient CYP716 family is a major contributor to the diversification of eudicot triterpenoid biosynthesis (2017).
45. T. Moses, *et al.*, OSC2 and CYP716A14v2 Catalyze the Biosynthesis of Triterpenoids for the Cuticle of Aerial Organs of Artemisia annua. 1–17 (2015).
46. A. Banerjee, P450s controlling metabolic bifurcations in plant terpene specialized metabolism (2017).
47. K. Tamura, *et al.*, Cytochrome P450 Monooxygenase CYP716A141 is a Unique b -Amyrin C-16 b Oxidase Involved in Triterpenoid Saponin Biosynthesis in Platycodon grandiflorus. **58**, 874–884 (2017).
48. R. C. Misra, S. Sharma, A. Garg, C. S. Chanotiya, Two CYP716A subfamily cytochrome P450 monooxygenases of sweet basil play similar but nonredundant roles in ursane- and oleanane-type pentacyclic triterpene biosynthesis (2017).

49. T. Moses, *et al.*, OSC2 and CYP716A14v2 catalyze the biosynthesis of triterpenoids for the cuticle of aerial organs of Artemisia annua. Plant Cell **27**, 286–301 (2015).
50. R. Yano, *et al.*, *Received Date: 08-Jul-2016 Revised Date : 12-Oct-*2016 (2016) https:/doi.org/10.1111/tpj.13403.
51. E. Pichersky, E. Lewinsohn, Convergent evolution in plant specialized metabolism. Annu. Rev. Plant Biol. **62**, 549–566 (2011).
52. P. Jing, *et al.*, Effect of glycosylation patterns of Chinese eggplant anthocyanins and other derivatives on antioxidant effectiveness in human colon cell lines. Food Chem. (2015).
53. J. M. Augustin, V. Kuzina, S. B. Andersen, S. Bak, Molecular activities, biosynthesis and evolution of triterpenoid saponins. Phytochemistry **72**, 435–457 (2011).
54. L. Chebil, C. Humeau, A. Falcimaigne, J. M. Engasser, M. Ghoul, Enzymatic acylation of flavonoids. Process Biochem. (2006).
55. A. Jozwiak, *et al.*, Plant terpenoid metabolism co-opts a component of the cell wall biosynthesis machinery. Nat. Chem. Biol. (2020).
56. P. K. Chan, Acylation with diangeloyl groups at C21-22 positions in triterpenoid saponins is essential for cytotoxcity towards tumor cells. Biochem. Pharmacol. (2007).
57. A. C. Huang, T. Jiang, Y. Liu, Y. Bai, J. Reed, A specialized metabolic network selectively modulates Arabidopsis root microbiota. **6389** (2019).
58. A. Kumar, P. Srivastava, G. Srivastava, N. Kumar, C. S. Chanotiya, BAHD acetyltransferase contributes to wound-induced biosynthesis of oleo-gum resin triterpenes in Boswellia. 1403–1419 (2021).
59. Z. Liu, *et al.*, Drivers of metabolic diversification: how dynamic genomic neighbourhoods generate new biosynthetic pathways in the Brassicaceae. New Phytol. **227**, 1109–1123 (2020).
60. D. Wang, *et al.*, New natural barrigenol-like triterpenoid isolated from the husks of Xanthoceras sorbifolia Bunge. Nat. Prod. Res. **32**, 997–1003 (2018).
61. X. Zhang, S. Zhang, Y. Yang, D. Wang, H. Gao, Natural barrigenol–like triterpenoids: A comprehensive review of their contributions to medicinal chemistry. Phytochemistry **161**, 41–74 (2019).
62. W. Xu, *et al.*, Barringtogenol c-type triterpenoid saponins from the stem bark of Norway maple (acer platanoides). Planta Med. **86**, 70–77 (2020).
63. F. Bourgaud, *et al.*, Biosynthesis of coumarins in plants: A major pathway still to be unravelled for cytochrome P450 enzymes. Phytochem. Rev. **5**, 293–308 (2006).
64. G. Vialart, *et al.*, A 2-oxoglutarate-dependent dioxygenase from Ruta graveolens L. Exhibits p-coumaroyl CoA 2-hydroxylase activity (C2H): A missing step in the synthesis of umbelliferone in plants. Plant J. **70**, 460–470 (2012).
65. W. Sun, *et al.*, The genome of the medicinal plant Andrographis paniculata provides insight into the biosynthesis of the bioactive diterpenoid neoandrographolide. Plant J. (2019) https:/doi.org/10.1111/tpj.14162.

66. Z. Xu, *et al.*, The genome of Corydalis reveals the evolution of benzylisoquinoline alkaloid biosynthesis in Ranunculales. Plant J. **111**, 217–230 (2022).
67. A. Jozwiak, *et al.*, Plant terpenoid metabolism co-opts a component of the cell wall biosynthesis machinery. Nat. Chem. Biol. **16**, 740–748 (2020).
68. K. Tamura, G. Stecher, D. Peterson, A. Filipski, S. Kumar, MEGA6: Molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. (2013) https:/doi.org/10.1093/molbev/mst197.
69. I. Amit, *et al.*, genome from RNA-Seq data. **29**, 644–652 (2013).
70. A. Manuscript, G. Annotation, *NIH Public Access* (2015).
71. Y. Wang, *et al.*, MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. (2012) https:/doi.org/10.1093/nar/gkr1293.
72. R. S. Nett, W. Lau, E. S. Sattely, *Discovery and engineering of colchicine alkaloid biosynthesis* (Springer US, 2020).
73. X. Li, *et al.*, Establishing an Artificial Pathway for Efficient Biosynthesis of Hydroxytyrosol. ACS Synth. Biol. **7**, 647–654 (2018).
74. Y. Lin, X. Sun, Q. Yuan, Y. Yan, Combinatorial biosynthesis of plant-specific coumarins in bacteria. Metab. Eng. **18**, 69–77 (2013).

Figures

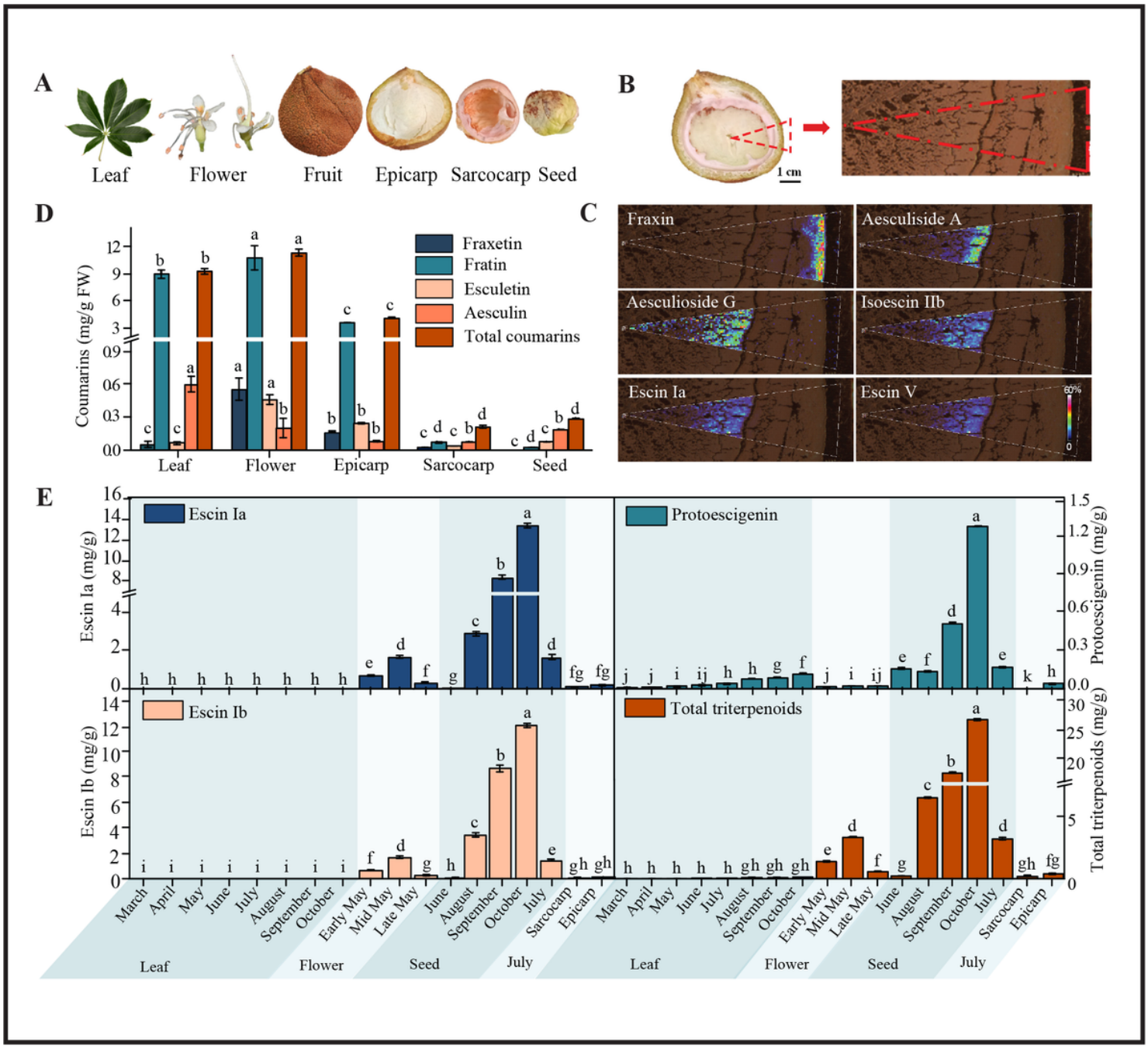

Figure 1

The profiling of triterpenoids and coumarins in different organs of *A. chinensis* during development. (A) Phenotypes of leaf, flower, and anatomical parts of fruit. (B) Image and optimal micrograph of the cross section of fruit. (C) MALDI-MSI images of triterpenoids and coumarins in the cross section of *A. chinensis* fruit. Ion images of Fraxin, Aesculiside A, Aesculioside G, escin Ia, Escin V. (D) Absolute quantification of fraxetin, fraxin, esculetin, esculin and total coumarins in leaf, flower, epicarp, scarcocarp and seed of *A. chinensis* by LC-MS analysis. Error bars indicate the SDs of three biological triplicates. Different letters represent significant differences ($P < 0.05$) between means according to analysis of variance (ANOVA) combined with duncan's multiple range test. (E) Absolute quantification of escin Ia, escin Ib, protoescigenin and total triterpenoids of tree parts from different months through LC-MS analysis. Error

bars indicate the SDs of three biological triplicates. Different letters represent significant differences ($P < 0.05$) between means according to analysis of variance (ANOVA) combined with duncan's multiple range test.

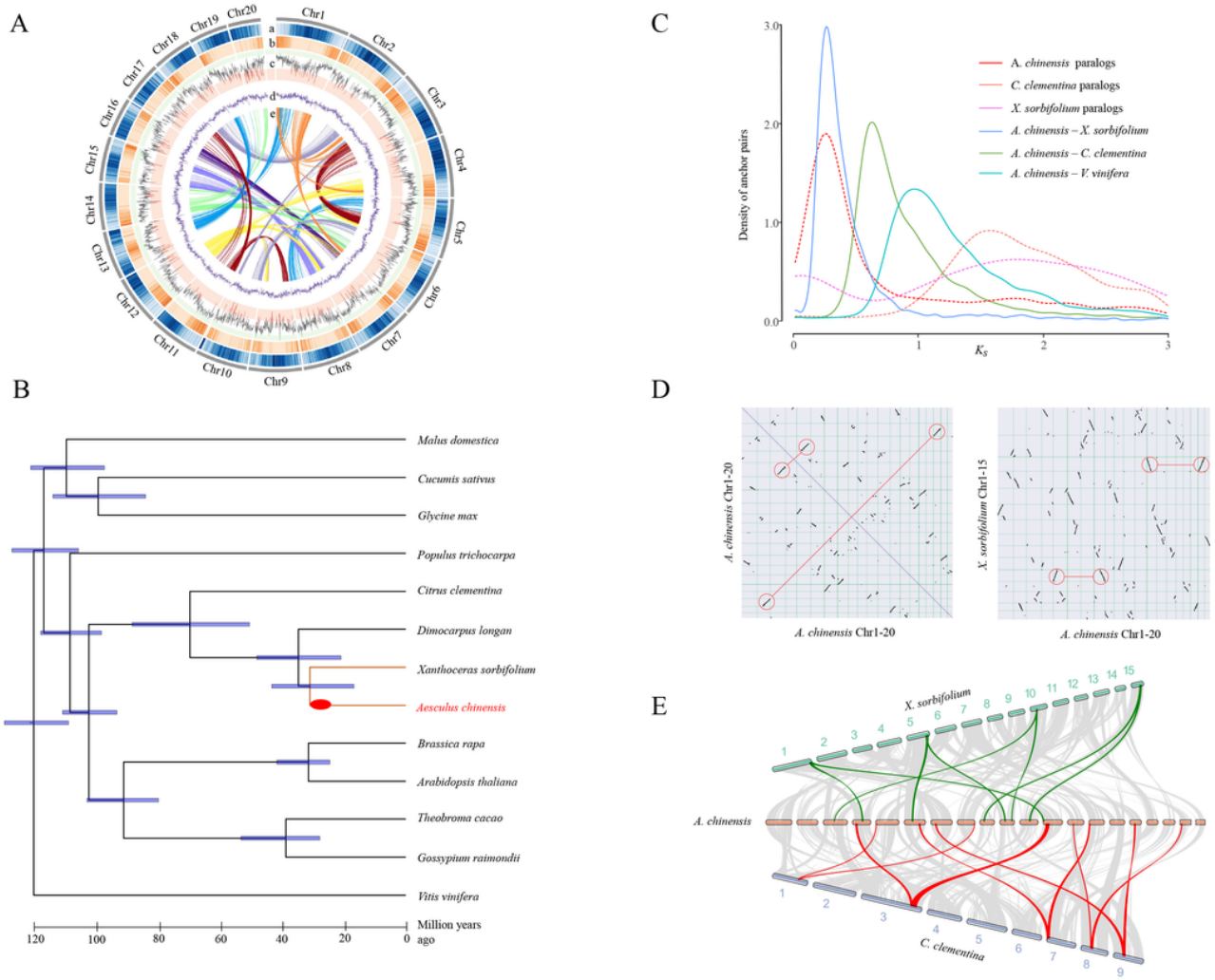

Figure 2

Evolutionary analysis of *A. chinensis* genome. (A) Chromosome level assembly of the *A. chinensis* genome. a, transposable elements; b, gene density; c, gene expression level between seed and other tissues; d, GC content; and e, synteny blocks of paralogous sequences of *A. chinensis* genome. (B) Phylogenetic tree inferred from orthologues of single gene families among selected plant taxa is shown. The 95% confidence interval for highest posterior density is shown on the internal nodes. The red circles on the *Aesculus* branch highlight an *Aesculus* specific WGD event (*Aα*). (C) The synonymous substitutions per synonymous site (K_S) distributions of orthologous and paralogous genes among *A. chinensis*, *C. clementina* and *X. sorbifolium*. (D) Comparison of chromosomes between *A. chinensis* and *A. chinensis* assemblies and *A. chinensis* and *C. clementina*. (E) Macrosynteny between genomic regions of *A. chinensis*, *C. clementina* and *X. sorbifolium*. Macrosynteny patterns between *A. chinensis* and *C.*

clementina or *X. sorbifolium* show that each *C. clementina* or *X. sorbifolium* region aligns to two syntenic regions in *A. thaliana* that experienced one round of *A. chinensis* genome duplication.

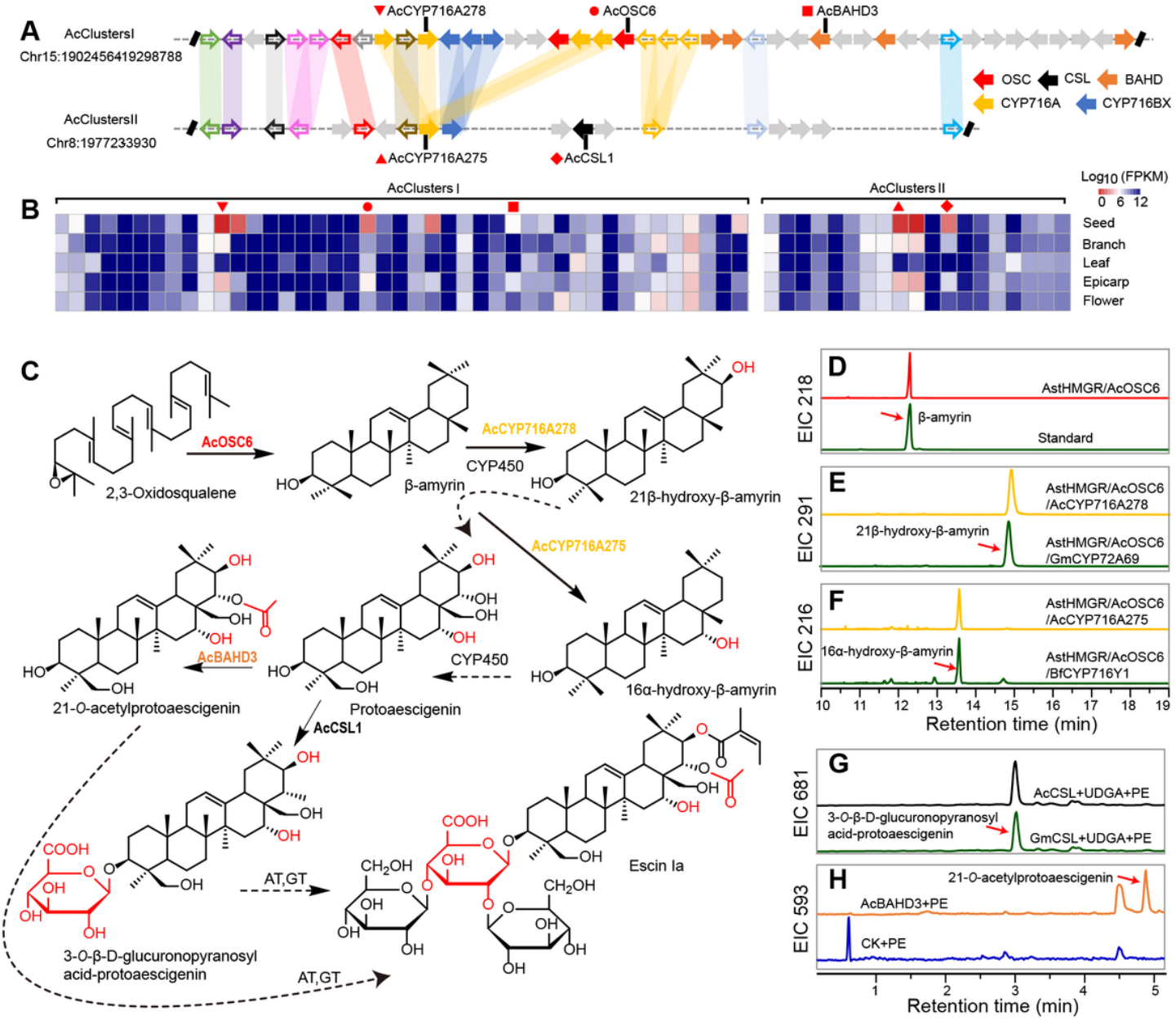

Figure 3

The mining and identification of BAT triterpenoid biosynthetic gene clusters in *A. chinensis*. (A) The collinearity of two triterpenoid biosynthetic gene clusters. AcCluster I and AcCluster II present strongly syntenic signal, and the collinearity of these two blocks are originated from *Aesculus* specific *Aa* WGD event. The thickness colored lines between syntenic genes represents the identity of related genes. (B) The expression of syntenic genes related to triterpenoid biosynthesis in different *A. chinensis* tissues, including seed, branch, leaf, epicarp and flower. (C) The proposed escin Ia biosynthetic pathway and candidate enzymes for each step. Solid arrows represent transformations verified in this study, and dotted arrows represent proposed reactions. (D, E, F) Overlay of GC-MS extracted ion chromatograms for

β-amyrin (*m/z* 218, D), 21β-hydroxy-β-amyrin (*m/z* 291, E), and 16α-hydroxy-β-amyrin (*m/z* 216, F) produced in *Nicotiana benthamiana*expressing AcOSC6 and AstHMGR (KY284573.1), and co-expressing with *AcCYP716A278*, *AcCYP716A275*,*GmCYP72A69* (41), or BfCYP716Y1(43), respectively. The soybean GmCYP72A69 has previously been reported to catalyze the hydroxylation of β-amyrin-derived scaffold at the C-21β position. BfCYP716Y1has been proved to hydroxylate β-amyrin at the C-16α position. (G, H) Overlay of LC-MS extracted ion chromatograms for 3-*O*-β-D-glucuronopyranosyl acid-protoaescigenin (*m/z* 681, G) produced in yeast WAT11 strain expressing AcCSL and GmCSL(40) using UDP-glucuronic acid as the sugar donor, and for 21-*O*-acetylprotoaescigenin (*m/z*593, H) produced in *Escherichia coli* expressing AcBAHD1. The soybean GmCSL has previously been reported to catalyze 3-O-glucuronosylation.

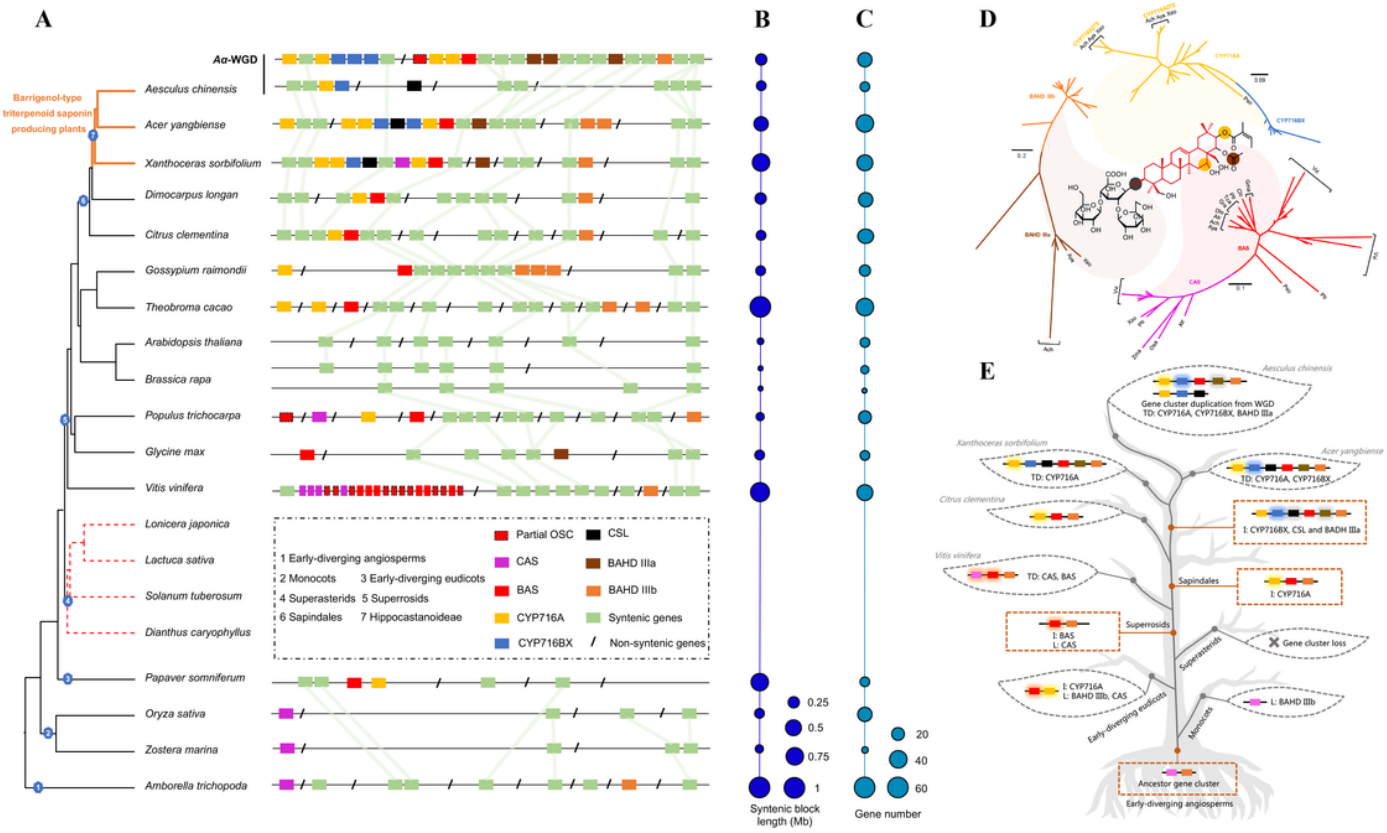

Figure 4

The evolution of BAT triterpenoid biosynthetic gene clusters among angiosperms. (A) The phylogenetic tree was constructed using OrthoFinder based on the reported genome information. The synthetic genes or non-synthetic genes were marked by different color, and the ancestor collinearity gene cluster was firstly identified in early-diverging angiosperm, *A. trichoppda*. (B) The length distribution of syntenic blocks among angiosperms. (C) The gene numbers localized in syntenic blocks among angiosperms. (D) The phylogenetic relationship of OSCs, CYP450s, and BAHDs, which localized in the BAT gene clusters. The different colors represent the correspondingly catalytic activities in escin Ia. (E) The evolutionary model of BAT triterpenoid gene clusters following lots of tandem duplication, gene insertion, gene loss, and WGD. The model illustrates that the BAT biosynthetic genes were swapped and duplicated into the

gene cluster, and subsequently evolved the accumulation of species-enriched BAT triterpenoids. I: insertion, L: loss, TD: tandem gene duplication.

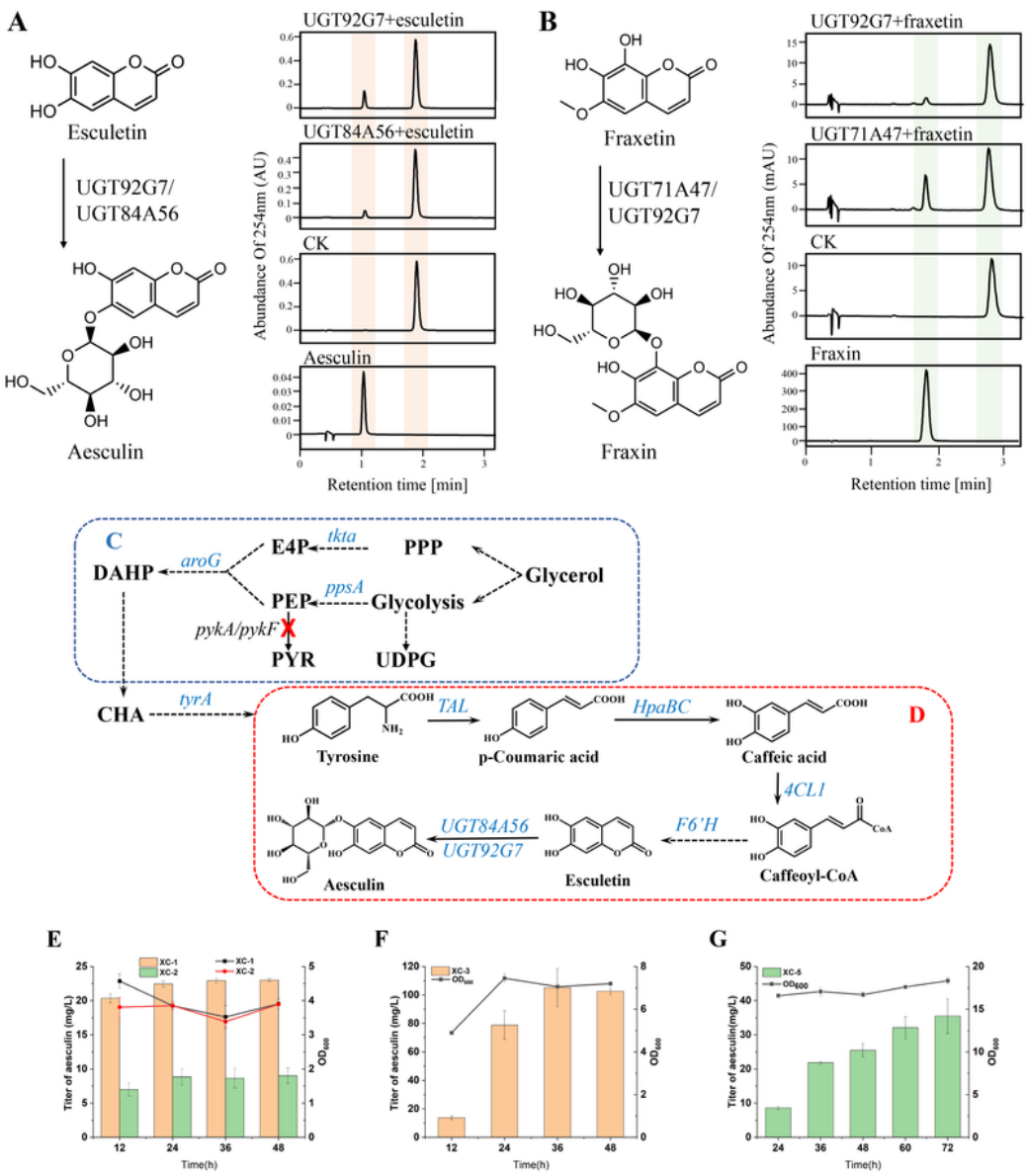

Figure 5

Functional identification of UGTs involved in aesculin and fraxin biosynthesis, and *de novo* biosynthetic pathway of aesculin. (A) UPLC spectrums for AcUGT92G7 and AcUGT84A57 enzyme activity analysis

toward esculetin. (B) UPLC spectrums for AcUGT92G7 and AcUGT71A47 enzyme activity analysis toward fraxetin. (C) Native tyrosine biosynthetic pathway in *E. coli*. (D) The artificial pathway for aesculin biosynthesis from tyrosine. Dashed arrows represent multi-step conversion. Red cross line represents pathway block; genes in blue indicate overexpression. PPP, Pentose Phosphate Pathway; UDPG, UDP-α-D-glucose; E4P, D-erythrose-4-phosphate; PEP, phosphoenolpyruvate; PYR, pyruvate; DAHP, 3-deoxy-arabino-heptulonate-7-phosphate; CHA, Chorismate. Tkta, transketolase; PpsA, phosphoenolpyruvate synthase; PykA/pykF, Pyruvate kinase; AroG, Phospho-2-dehydro-3-deoxyheptonate aldolase; TyrA, Chorismate mutase-prephenate dehydrogenase; TAL, tyrosine ammonia lyase; HpaBC, 4-hydroxyphenylacetic acid 3-hydroxylase; 4CL1, 4-coumarate-CoA ligase; F6'H, Feruloyl-CoA 6'-hydroxylase; AcUGT84A56 and AcUGT92G7, glucosyltransferases. (E) Aesculin production from esculetin using different glucosyltransferase (AcUGT84A56 and AcUGT92G7). Error bars represent standard deviations of three replicates. (F) Aesculin production from *p*-coumaric acid by strain XC-3 (BW25113, pCS-84-HpaBC and pZE-F6'H-4CL). Error bars represent standard deviations of three replicates. (G) *De novo* biosynthesis of aesculin by strain XC-6 (BW-1, pCS-TPTA-HpaBC and pET-648T). Error bars represent standard deviations of three replicates.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable16.xlsx
- SupportingOnlineMaterialforSupplementary20221121.docx