

*This is an article that is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain. The final authenticated version is available online at: <https://doi.org/10.1111/tpj.17220>*

*This work was funded by European Research Council (DOUBLE-TROUBLE 833522)*

### **Deciphering the biosynthetic pathway of triterpene saponins in *Prunella vulgaris***

Si-Jie Liu<sup>1,†</sup>, Zhengtai Liu<sup>2,†</sup>, Bing-Yan Shao<sup>1</sup>, Tao Li<sup>1</sup>, Xinning Zhu<sup>2</sup>, Ren Wang<sup>2</sup>, Lei Shi<sup>3,\*</sup>, Sheng Xu<sup>2,\*</sup>, Yves Van de Peer<sup>1,4,5\*</sup>, Jia-Yu Xue<sup>1,\*</sup>

<sup>1</sup>College of Horticulture, Bioinformatics Center, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing 210095, China.

<sup>2</sup>Institute of Botany, Jiangsu Province and Chinese Academy of Sciences, Nanjing 210014, China

<sup>3</sup>Jiangsu Key Laboratory of Drug Design and Optimization, Department of Medicinal Chemistry, China Pharmaceutical University, Nanjing 210009, China

<sup>4</sup>Department of Plant Biotechnology and Bioinformatics, Ghent University, VIB-UGent Center for Plant Systems Biology, B-9052, Belgium

<sup>5</sup>Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria 0028, South Africa

<sup>†</sup>These authors contributed equally to this work.

\*Corresponding author. Email: xuejy@njau.edu.cn (J.Y.X.); yvpee@psb.vib-ugent.be (Y.V.d.P.); xusheng@cnbg.net (S.X.); shilei@cpu.edu.cn (L.S.)

## **Abstract**

The traditional Chinese medicinal plant *Prunella vulgaris* contains numerous triterpene saponin metabolites, notably ursolic and oleanolic acid saponins, which have significant pharmacological values. Despite their importance, the genes responsible for synthesizing these triterpene saponins in *P. vulgaris* remain unidentified. This study used a comprehensive screening methodology, combining phylogenetic analysis, gene expression assessment, metabolome-transcriptome correlation and co-expression analysis, to identify candidate genes involved in triterpene saponins biosynthesis. Nine candidate genes - two OSCs, three CYP716s and four UGT73s - were precisely identified from large gene families comprising hundreds of members. These genes were subjected to heterologous expression and functional characterization, with enzymatic activity assays confirming their roles in the biosynthetic pathway, aligning with bioinformatics predictions. Analysis revealed that these genes originated from a whole-genome duplication (WGD) event in *P. vulgaris*, highlighting the potential importance of WGD for plant metabolism. This study addresses the knowledge gap in the biosynthesis of triterpene saponins in *P. vulgaris*, establishing a theoretical foundation for industrial production via synthetic biology. Additionally, we present an efficient methodological protocol that integrates evolutionary principles and bioinformatics techniques in metabolite biosynthesis research. This approach holds significant value for studies focused on unraveling various biosynthetic pathways.

**Keywords:** Chinese herb medicine, multi-omics, triterpenoid saponins biosynthesis, OSC, CYP450, UGT

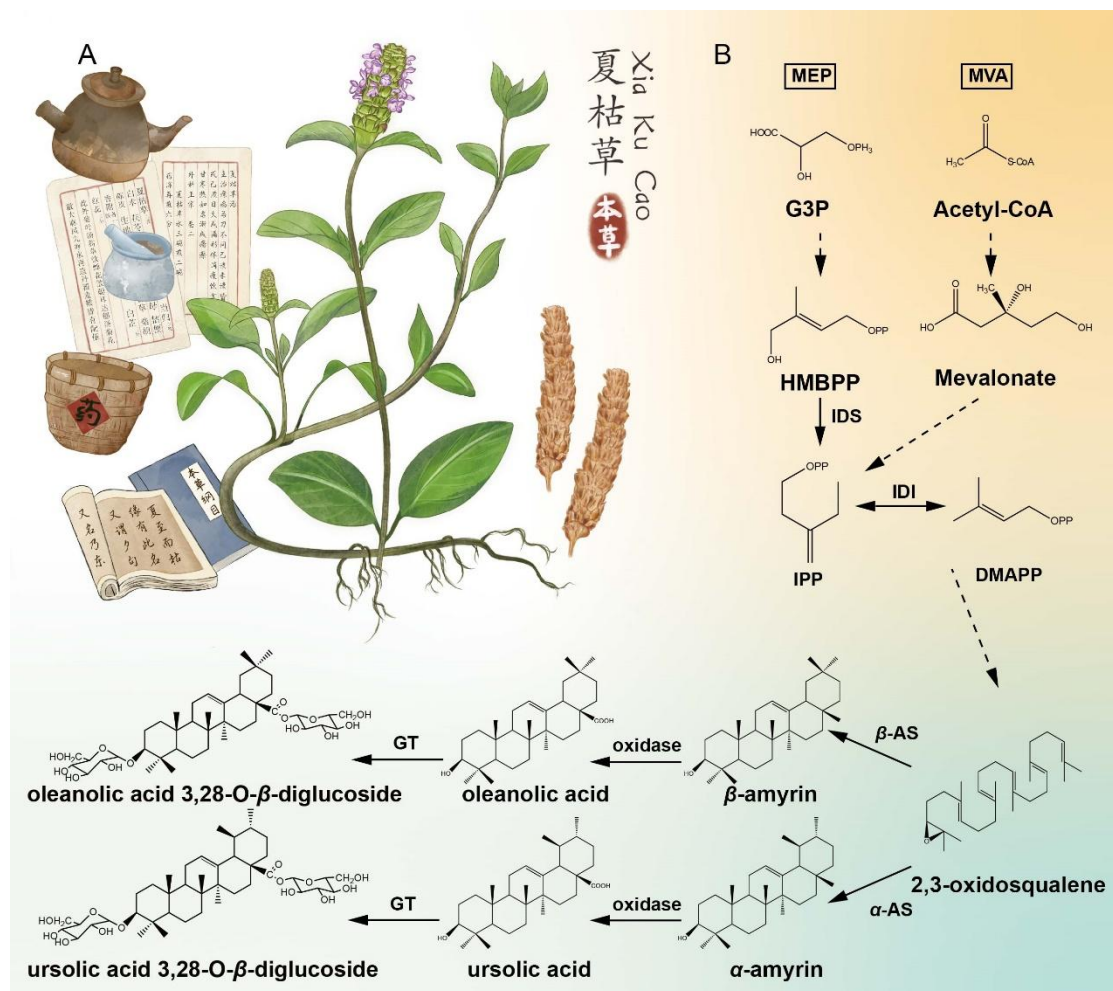
## Introduction

Lamiaceae, an angiosperm family encompassing more than 200 genera and over 3,500 species (Group *et al.*, 2016), exhibits a global distribution, and predominantly comprises herbs and vines, with occasional shrubs and small trees. The hallmark feature of this family lies in its abundant and diverse volatile oil components, which after extractions, play a crucial role in various aspects of human daily life. There have been extensive applications of Lamiaceae plants in medicinal contexts, exemplified by species such as *Scutellaria baicalensis* (Zhao *et al.*, 2019), *Salvia miltiorrhiza* (Ma *et al.*, 2021) and *Schizonepeta tenuifolia* (Liu *et al.*, 2023b). Additionally, they serve as culinary spices, including *Ocimum basilicum*, *Pogostemon cablin*, and *Perilla frutescens*. Furthermore, the edible chia seeds derived from *Salvia hispanica* contribute to the production of various food products such as sticks, cookies, noodles, and bread. *Rosmarinus officinalis* plays a vital role in perfume and soap production, while the industrial significance of peppermint (*Mentha* plants) is noteworthy.

Given the manifold role of Lamiaceae plants in human life, the genomes of 39 species, including several medicinal plants, have been sequenced to date. These genomes serve as fundamental sources, offering valuable insights into the biosynthetic pathways of diverse metabolites. Notable examples include the identification of 4'-deoxyflavonoid enzyme-encoding genes in *S. baicalensis*, leading to the decoding of the Baicalein biosynthetic pathway (Zhao *et al.*, 2019). Analysis of the *S. hispanica* genome suggests that the expansion of the *FAB2* gene family is responsible for the high omega-3 content (Wang *et al.*, 2022). Similarly, the genetic study of *S. tenuifolia* unveiled the molecular mechanism of pulegone biosynthesis through evolutionary analysis and experimental validations (Liu *et al.*, 2023a).

*Prunella vulgaris* L. (Chinese name "夏枯草", [Figure 1A](#)) is a perennial herb in Lamiaceae, with a wide distribution in southern regions of the Qinling Mountains in China and other temperate and subtropical areas worldwide (Wu *et al.*, 1900, Wang *et al.*, 2019). Harvested in early summer, the mature seeds, dried spikes, or even the entire plant of *P. vulgaris* are used in traditional Chinese medicine ([Figure 1A](#)). The triterpenoids, particularly ursolic acid and oleanolic acid, are considered the principal active constituents, with ursolic acid serving as an indicator of herb quality (China,

2010). The pharmacological effects of *P. vulgaris*, including the recordation of "cleaning liver and brightening eyes, clearing knot and detoxification" in the masterpiece on traditional Chinese medicine Compendium of Materia Medica, and the anti-inflammatory, blood pressure-lowering, and blood circulation-improving properties discovered by modern medicine, are attributed to the abundance of ursolic acid (Li et al., 2023b), making *P. vulgaris* a promising candidate for an anti-tumor agent. Furthermore, *P. vulgaris* plays a significant role in the food industry, particularly in Southern China, where it is used as a primary ingredient in herbal teas, and the annual market demands in China reach approximately 5,100 tons. Despite the wide distribution, *P. vulgaris* growing in Jiangsu is considered of high pharmacological quality, and recognized as the genuine herbal material (Yang, 2015).



**Figure 1. Morphology of *P. vulgaris* and its hypothetical pentacyclic saponins biosynthetic pathway.**

**(A)** Morphology of *P. vulgaris*. **(B)** The proposed pentacyclic saponins biosynthetic pathway in plants. MEP: methylerythritol phosphate pathway, MVA: mevalonate pathway, HMBPP: dimethylallyl pyrophosphate, G3P: Glyceraldehyde 3-phosphate; IDS: Isoprenyl diphosphate synthase, IDI: isopentenyl diphosphate isomerase;  $\beta$ -AS:  $\beta$ -amyryn synthase;  $\alpha$ -AS:  $\alpha$ -amyryn

synthase; GT: glycosyltransferase. Dashed lines represent multiple steps, solid lines represent a single step.

In the biosynthetic pathway of triterpenoid compounds in plants, the process initiates with the mevalonic acid (MVA) pathway and, to some extent, the 2-methyl-D-erythritol-4-phosphate (MEP) pathway, both contributing to the production of the common precursor, isopentyl diphosphate (IPP) (Krokida *et al.*, 2013). The subsequent step involves the formation of the triterpene basic skeleton precursor, 2,3-oxidosqualene, catalyzed by squalene oxide cyclase (OSC). The final steps require the involvement of oxidase and glycosyltransferase, typically encoded by members of the cytochrome P450 (CYP450) (Zhou *et al.*, 2019) and uridine diphosphate (UDP)-glycosyltransferases (UGT) (Erthmann *et al.*, 2018) gene families to produce triterpene saponins (Figure 1B). However, the precise enzymes catalyzing these last three steps in *P. vulgaris* remain poorly understood, given that the genes encoding these enzymes belong to gene families with multiple members, especially the vast CYP450 and UGT families, which contain hundreds of members in each angiosperm species.

Although attempts have been made to sequence the genome of *P. vulgaris* and to study the biosynthesis of triterpene saponins, only genes encoding OSCs were identified and characterized, whereas CYP450s or UGTs were only predicted without functional verification (Bryson *et al.*, 2023, Zhang *et al.*, 2024). To unravel the complete biosynthetic pathway of ursolic acid and oleanolic acid saponins in *P. vulgaris*, we collected samples from Nanjing, Jiangsu, where the cultivated *P. vulgaris* is considered of high pharmacological quality and assembled a chromosome-level genome. Through comprehensive analyses integrating genomic, transcriptomic, and metabolic data, coupled with functional characterizations through experiments, we aimed to identify key enzyme-encoding genes involved in the process. This study not only unveils the complete triterpene saponin biosynthetic pathway in *P. vulgaris*, but also presents an efficient methodology for identifying unknown genes.

## Results

### Genome and metabolome of *P. vulgaris*

*P. vulgaris*, a diploid ( $2n = 2x = 28$ ) species (Hanlin *et al.*, 2022), was estimated to have a genome size of approximately 690 Mb, based on K-mer distribution ( $K = 19$ )

analysis and exhibiting a heterozygosity of 0.79% (Supplementary Figure S1). Our preliminary assembly based on Nanopore and Illumina reads generated 104 contigs (contig N50 = 15.3 Mb), totaling a size of 683.87 Mb (Supplementary Table S1). Subsequent HI-C data anchored all contigs onto 14 scaffolds, representing 14 haplotype pseudochromosomes ranging from 30.1 Mb to 79.9 Mb in length (Supplementary Figure S2). Gene annotation generated 45,001 protein-coding gene models, and 98.1% of BUSCO groups (embryophyta\_odb10) were completely captured after annotation, indicating a high quality and completeness of genome assembly. Targeted metabolomic analysis was employed to characterize the metabolite profile of terpenoids in *P. vulgaris*. Samples comprised the roots, stems, leaves, seeds, and spikes. Through comparison with a reference standard library, we identified 63 terpenoid compounds from all tissues, including 11 diterpenoids, five sesquiterpenoids, 42 triterpenoids, and five triterpenoid saponins. Based on cluster analysis of the 63 terpenoid compounds (Supplementary Figure S3), all tissues, apart from the spikes, exhibit richness in various oleanane and ursane-type triterpenoids, whereas *P. vulgaris* spikes accumulate nearly all triterpenoid compounds. We selected five typical ursane and oleanane-type compounds, namely ursolic acid, ursonic acid, tormentic acid, trihydroxyursan, and moronic acid, as marker compounds to assess their expression abundance in different tissues. As depicted in the K-means plot (Supplementary Figure S4), the expression abundance of most markers (Supplementary Table S2; class1: tormentic acid, class2: trihydroxyursan, class6: ursolic acid) in spikes exceeded that in other parts, lending objective support to the traditional Chinese medicine theory of utilizing the spikes of *P. vulgaris* for medicinal purposes. A minor subset of markers (class9: ursolic acid, moronic acid) exhibited higher expression in the roots, indicating potential specific accumulation of certain triterpenoid compounds in roots.

### **Identification of candidate genes involved in *P. vulgaris* triterpenoid saponins biosynthesis**

To identify enzyme-encoding genes associated with the final three biochemical reactions of triterpenoid saponin biosynthesis in *P. vulgaris*, a comprehensive approach was undertaken using the assembled reference genome (Figure 2).

### *Genome-wide gene family identification*

Initially, a genome-wide identification of three gene families (OSC, CYP450 and UGT) was conducted by homologous blast and HMM model search and verified by Pfam domain recognition. A total of 11 OSC, 364 CYP450 and 172 UGT genes were identified (Figures 2A, 2B and 2C). Considering the impracticality of functionally characterizing all these genes, a streamlined methodology was developed to efficiently screen unknown enzyme-encoding genes from this vast dataset.

### *Phylogenetic screening*

We initially sought clues based on evolutionary principles. All functionally characterized enzyme-encoding genes involved in the three steps were compiled from previous studies (30 OSCs, ten CYP450s and 22 UGTs, Supplementary Table S3), and served as crucial reference in phylogenetic analysis. Subsequently, we identified and extracted all genes associated with OSC, CYP450 and UGT families in 18 selected plant genomes (Supplementary Table S4) using the comprehensive genome-wide identification approach described in the **Methods** section. The phylogenies of the three gene families were constructed by aggregating all identified genes. The OSC gene phylogeny revealed four distinct monophyletic branches, clearly separating the functionally differentiated  $\beta$ -amyrin synthase ( $\beta$ -AS), mixed-AS, lupeol synthase (LUS) and dammarenediol synthase (DS) into four branches (Figure 2D). This alignment between functional roles and phylogenetic grouping underscores the evolutionary constraints on the functional divergence of the OSC gene family, offering valuable insights for our candidate gene screening. For the OSC phylogeny, all characterized OSCs that generate  $\alpha$ - and  $\beta$ -amyrins fall into either the  $\beta$ -AS or the mixed-AS branches, with *P. vulgaris* exhibiting six genes distributed into these two branches (three in each branch, Supplementary Figure S5). Consequently, the six *P. vulgaris* OSCs were selected as the candidate enzyme-encoding genes in the screening process. Similarly, for the CYP450 phylogeny, all ten characterized enzymes of ursolic and oleanolic acids from different angiosperm lineages clustered in the CYP716 subfamily (Figure 2E), in which *P. vulgaris* possesses seven genes (Supplementary Figure S6). Thus, the seven CYP716s were the screened candidates that potentially function in this step. Finally, our

observation for the UGT phylogeny revealed that the characterized UGTs utilizing ursolic and oleanolic acid as substrates belonged to the UGT73 subfamily (Figure 2F and Supplementary Figure S7). Accordingly, 12 *P. vulgaris* UGT73 genes were preliminarily identified as candidates for further investigation.

#### *Expression assessment of candidate genes*

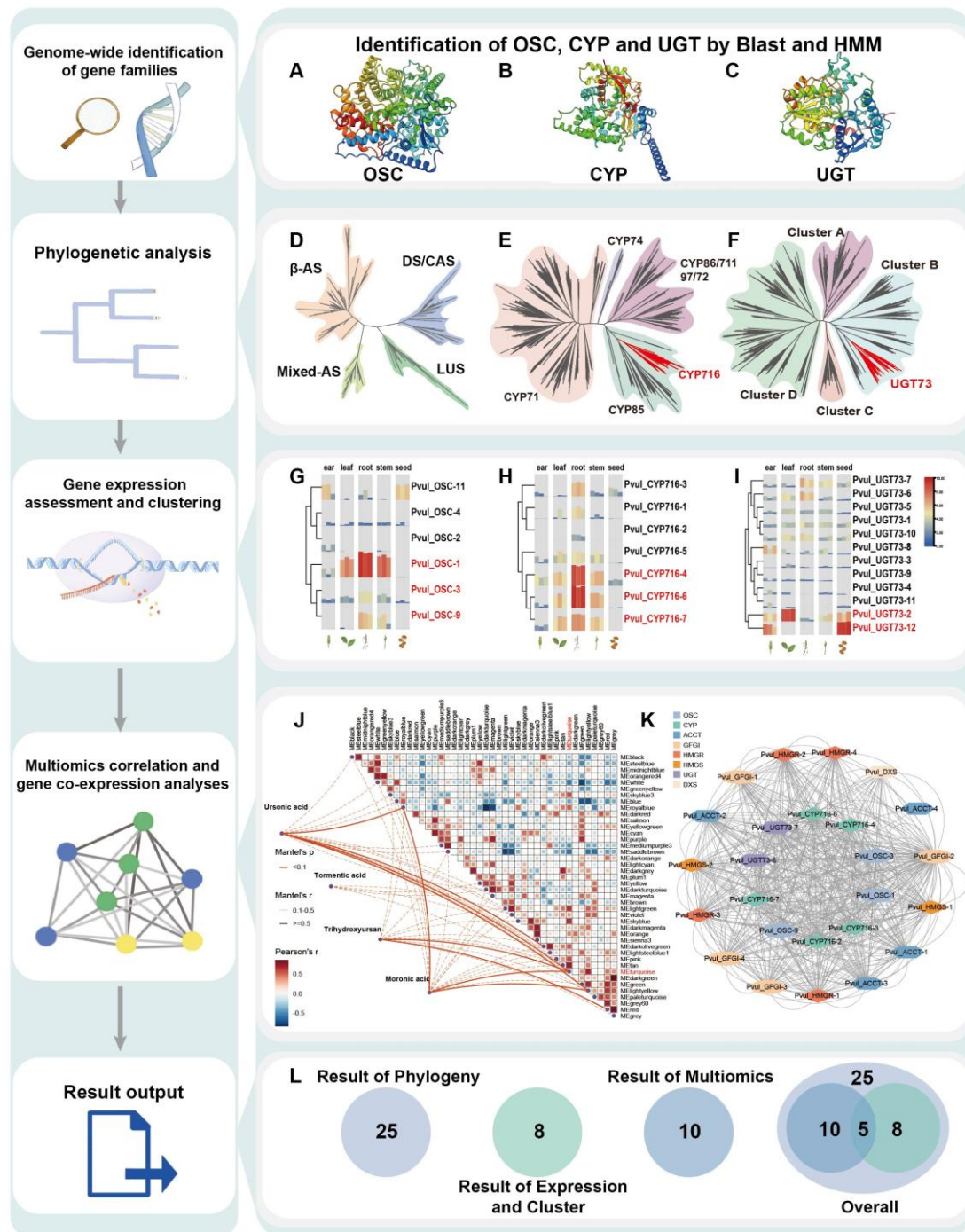
Following the evolutionary screening, we continued to examine gene expression profiles to gather additional clues, aiding in the refinement of potential candidate genes involved in triterpene saponins biosynthesis in *P. vulgaris*. Transcriptomic data revealed notable expression levels for three OSCs (Pvul\_OSC-1, Pvul\_OSC-3 and Pvul\_OSC-9, labeled in red in Figure 2G), three CYP716s (Pvul\_CYP-4, Pvul\_CYP-6 and Pvul\_CYP-7, Figure 2H) and two UGT73s (Pvul\_UGT73-2, Pvul\_UGT73-12, Figure 2I) in leaves, roots or seeds, indicating their heightened activity in these organs. Hence, these genes are more likely to be the enzyme-encoding genes involved in triterpenoid saponins biosynthesis in *P. vulgaris*.

#### *Metabolome-transcriptome correlation and co-expression analyses*

Our metabolomic analysis of five organs of *P. vulgaris* revealed significant accumulation of ursolic and tormentic acid, trihydroxyursan and moronic acid in leaves and/or roots (Supplementary Table S2). To identify genes potentially associated with these compounds, we conducted correlation analyses between their contents and gene expression patterns across different organs. All genes were classified into 42 modules (Supplementary Table S6), among which the turquoise, saddlebrown, royalblue, darkolivegreen and yellow modules exhibited significant correlation with ursolic acid, tormentic acid, trihydroxyursan and moronic acid (Figure 2J). Among the five modules, we observed a grouping of three OSCs, five CYP716s and two UGT73s into the turquoise module (Figure 2K and Supplementary Table S5). This module also included most of the upstream enzyme-encoding genes in the MVA and MEP pathways (HMGS, ACCT, FPPs, GPPs and IDS), indicating a shared expression pattern. Consistent expression patterns are typically observed among genes participating in the up- and downstream steps of the same pathway. Consequently, the three OSCs, five CYP716s and two UGT73s in this module are likely to play a downstream role to the MVA and

MAP pathways. In contrast, other modules contain few OSCs, CYP716s, UGT73s and other upstream enzyme-encoding genes ([Supplementary Table S5](#)). Notably, candidates identified through metabolome-transcriptome correlation and co-expression analyses largely overlapped with those identified through phylogenetic clues and expression assessment, except for only two UGT73s. This underscores the utility of metabolome-transcriptome correlation and the co-expression analyses in providing valuable clues for the identification of unknown genes.

In summary, our screening strategy, integrating phylogenetic analysis, gene expression assessment, metabolome-transcriptome correlation and co-expression analyses effectively identified 12 candidates (three OSCs, five CYP716s and four UGT73s, [Figure 2L](#)) from a gene pool comprising hundreds of members as the most promising candidate genes encoding enzymes involved in triterpene saponins biosynthesis. Two OSCs and three CYP716s were selected by multiple approaches, indicating their higher degree of confidence. However, for the UGTs involved in the last catalyzing step, four candidates were the union of different screening results, because UGT73\_2 and UGT 73\_12 with the highest expression levels among all UGT73s were not screened by the co-expression network, which, instead, suggested two other genes (UGT73\_6 and UGT73\_7) with lower expression levels. We speculate that UGTs have multiple roles participating not only in the triterpene saponins biosynthesis but also in other pathways. To validate the effectiveness of our screening strategy, two OSCs, three CYP716s and four UGT73s candidates were subject to functional validation for their enzymatic activities.



**Figure 2. Flow chart and results of screening enzyme-encoding genes involved in the biosynthesis of ursolic and oleanolic acid saponins in *P. vulgaris*.**

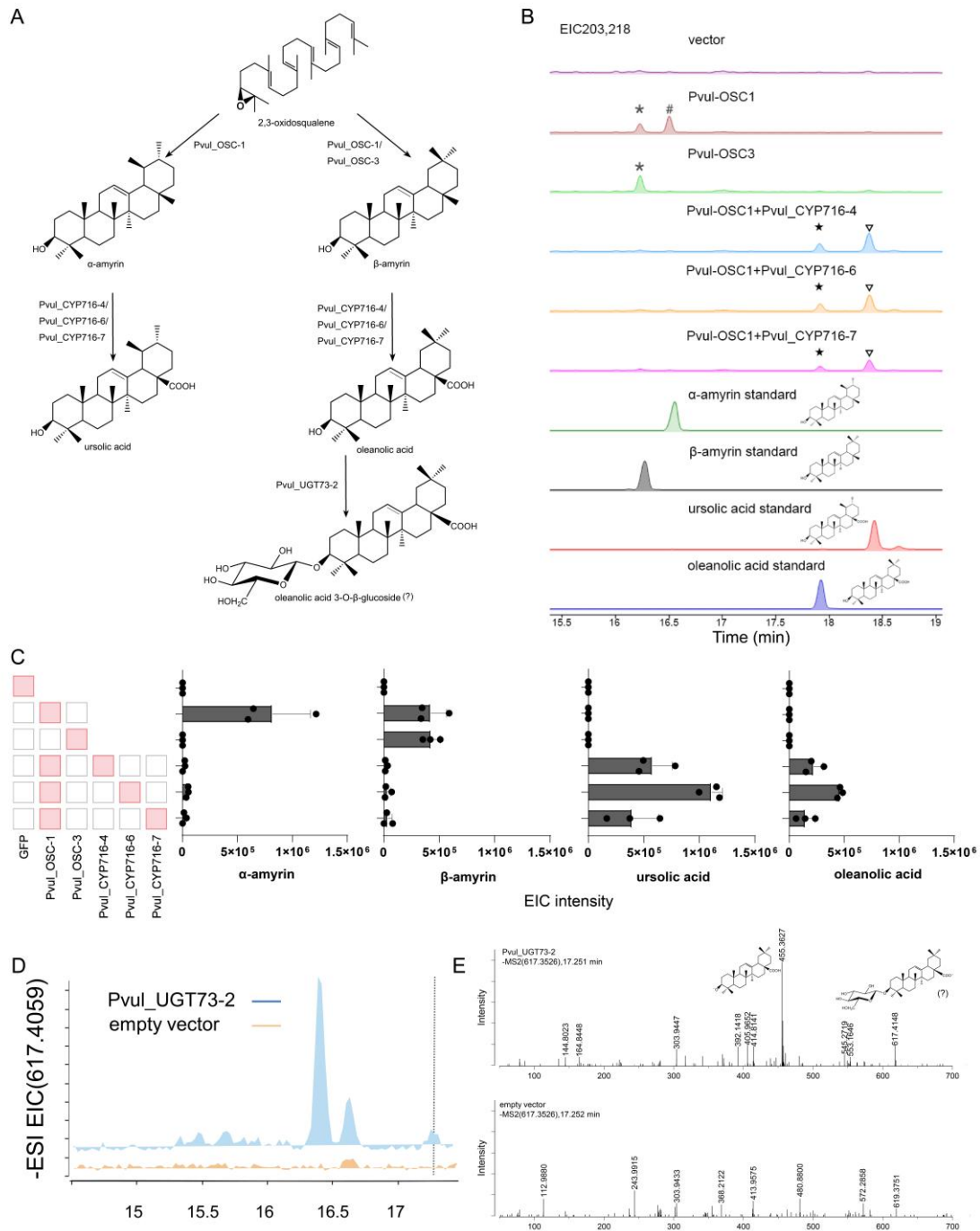
The left column indicates five steps of our integrated protocol for screening unknown genes, while the right part displays the results generated at each screening step. Genome-wide identification of OSC (A), CYP450 (B) and UGT (C) genes were performed first by blast and HMM search from the genomes of *P. vulgaris* and other 18 species. Subsequently, phylogenetic analyses of OSC (D), CYP450 (E) and UGT (F) genes were performed using the sequences of all identified genes and characterized enzymes collected from published literatures. Subfamily branches are distinguished by unique colors and shaded bands. The detailed phylogenies and IDs for the candidate genes are provided in [Supplementary Figures 5-7](#). The screened candidate genes by phylogenetic analyses continued to be applied to the assessment of their expression in different organs of *P. vulgaris*. Genes were clustered according to their differential expression levels and highly expressed genes were selected, resulting in three OSCs (G), three CYP716s (H) and two UGT73s (I). Metabolome-

transcriptome correlation analysis was conducted to associate metabolic products with genes according to the content of metabolites and gene expression patterns (**J**). Connections between modules and metabolites indicate a P-value less than 0.1, with the thickness of the lines representing the degree of correlation. Genes were classified into different modules by distinct co-expression patterns by WGCNA ([Supplementary Figure 8](#)). Among the modules showing significant correlation with ursolic acid and oleanolic acid, the turquoise module contained three OSCs, five CYP716s and two UGT73s ([Supplementary Table S5](#)), as well as most enzyme-encoding genes in the upstream process of triterpene saponins biosynthesis (**K**). Hence, this step produced 10 candidates. At last, the screening pipeline generated a total of 36 candidate genes potentially participating in the triterpene saponins biosynthesis in *P. vulgaris*, and five of them were repeatedly screened by multiple analyses, thus were considered as the most promising candidates (**L**).

### Functional validation of candidate genes

Two OSCs (Pvul\_OSC-1 and Pvul\_OSC-3) and three CYP716s (Pvul\_CYP-4, Pvul\_CYP-6 and Pvul\_CYP-7) candidates were successfully cloned from cDNA prepared from the leaves of *P. vulgaris*, using gene-specific primers ([Supplementary Table S7](#)). We then investigated the function of OSCs/CYP716s by *Agrobacterium*-mediated transient expression in the leaves of *Nicotiana benthamiana*. Gas chromatography-mass spectrometry (GC-MS) analysis of leaf extracts showed a peak with the same retention time and mass spectrum as an authentic  $\beta$ -amyrin standard ([Figure 3A](#)), confirming that Pvul\_OSC-3 is indeed a  $\beta$ -amyrin synthase ([Figure 3B](#) and [Supplementary Figure S9](#)). In addition, Pvul\_OSC-1 was proven able to produce both  $\alpha$ -amyrin and  $\beta$ -amyrin ([Figure 3A](#) and [3B](#), [Supplementary Figure S10](#)). Consequently, we first identified the two enzyme-encoding genes responsible for triterpenoids precursor skeleton.

Next, we focused on the candidates involved in the oxidation of  $\alpha$ -amyrin and  $\beta$ -amyrin. The members of the CYP716 family were mostly characterized as triterpene oxidases that catalyze three-step C-28 oxidation of  $\alpha$ -amyrin,  $\beta$ -amyrin and lupeol (Malhotra and Franke, 2022). Of the five CYP716s selected, three cloned CYPs (Pvul\_CYP716-4, Pvul\_CYP716-6, or Pvul\_CYP716-7) were shown to play a functional role as expected. Transient expression of each of the three CYP716s coupled with Pvul\_OSC-1 in *N. benthamiana* resulted in nearly total conversion of  $\alpha$ -amyrin to ursolic acid, or  $\beta$ -amyrin to oleanolic acid, exhibiting a higher capacity to produce ursolic acid ([Figure 3C](#)).



**Figure 3. Functional characterization of *P. vulgaris* candidate genes involved in triterpene biosynthesis.**

**(A)** Biosynthetic route leading to the production of  $\alpha$ -amyrin,  $\beta$ -amyrin, ursolic acid, and oleanolic acid, as well as the putative glucosylation product of oleanolic acid (oleanolic acid 3-O- $\beta$ -glucoside) in *P. vulgaris*. PvuI\_OSC-1/-3, oxidosqualene cyclase; PvuI\_CYP716-4/-6/-7, cytochrome P450 proteins; PvuI\_UGT73-2, UDP-dependent glycosyltransferase. **(B)** GC-MS extracted ion chromatograms (EIC) for leaf extracts of *N. benthamiana* after expression of each OSC or co-expression of the PvuI\_OSC-1 with the different PvuI\_CYP716. All intermediates were validated by comparison to synthetic authentic standards. **(C)** The extracted ion abundance for the exact ion mass of products produced in *N. benthamiana* after expression of each PvuI\_OSC or coexpression of the PvuI\_OSC-1 with the different PvuI\_CYP716. Data are means  $\pm$  sd.;  $n = 3$  biological replicates. **(D)** In vitro assay of PvuI\_UGT73-2 protein expressed in *E. coli*. Shown are HPLC-MS

chromatograms representing oleanolic acid (5), as well as one mass feather ( $m/z$  617.4636) that might be as putative product of Pvul\_UGT73-2. (E) MS<sup>2</sup> spectra of the new compound produced by Pvul\_UGT73-2, along with predicated ion fragment structure.

Subsequently, we determined the enzymes responsible for the addition of a sugar unit at the C-3/C-28 position of the oleanolic acid scaffold. The UGTs are typically responsible for glycosylation of plant natural products by using UDP-activated sugar donors to transfer sugar units onto small molecules (Rahimi *et al.*, 2019). Accordingly, we successfully cloned three UGT73s (Pvul\_UGT73-2, Pvul\_UGT73-7 and Pvul\_UGT73-12) from *P. vulgaris*. To determine if the candidate UGTs have catalytic activities towards oleanolic acid, they were heterologously expressed in *E. coli*. The corresponding crude protein extracts were assayed as putative sugar acceptors and UDP-D-glucose as sugar donor. HPLC-MS results showed that Pvul\_UGT73-2 transfers a glucose moiety from UDP-D-glucose onto the oleanolic acid (Figure 3D). In addition, tandem mass spectrometry (MS<sup>2</sup>) analysis supported the glucosylation of oleanolic acid (Figure 3E). However, further characterization is required to determine whether Pvul\_UGT73-2 catalyzes 3-O-glucosylation of the oleanolic acid.

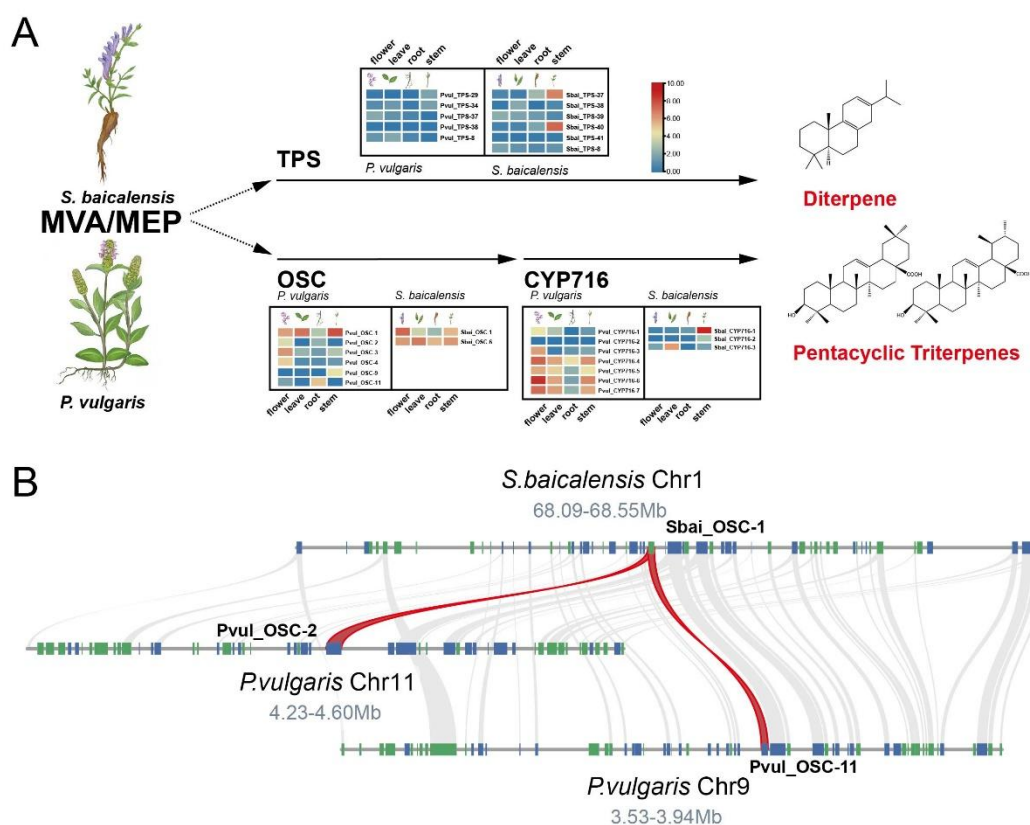
Following a series of functional experiments, all enzymes for the three missing steps of triterpene saponins biosynthesis were finally identified and characterized, thereby successfully unraveling the complete biosynthetic pathway in *P. vulgaris*. The screening methodology is also proved to be of high accuracy, and likely of broad application in exploring unknown genes in metabolic pathways.

### **Independent WGDs and differential gene retention account for distinct terpenoids accumulation among Lamiaceae plants**

Lamiaceae is famous for a rich diversity of terpenoid compounds, characterized by diverse structures and high content. However, the composition and abundance of terpenoids vary significantly among different species within the family. For instance, *P. vulgaris* produces rich triterpenoid compounds, while *S. baicalensis* mainly generates diterpenoid compounds. This variability can be explained by several factors such as copy number variation and differences in expression of the genes involved in the triterpenoid and diterpenoid biosynthetic pathways.

In triterpenoid biosynthesis, *P. vulgaris* possesses six OSCs and seven CYP716s

serving as potential enzymes, while *S. baicalensis* has only two OSCs and three CYP716s (Figure 4A). This notable difference in gene copy numbers may contribute to the abundance of pentacyclic triterpenes in *P. vulgaris*. The underlying mechanism likely involves species-specific gene duplications that led to the expansion of gene (sub)families. Further examination revealed well preserved collinearity between the orthologous genes encoding the three enzymes in the *P. vulgaris* and *S. baicalensis* genomes, and showed a primary 2:1 ratio (Figure 4B and Supplementary Figure S11), suggesting that the *P. vulgaris* enzymes are derived from a lineage-specific WGD (Supplementary Figure S12). While *S. baicalensis* also underwent an independent WGD, many duplicates of enzyme-encoding genes have been lost over time. In contrast, retained duplicates in *P. vulgaris* evolved enzymatic activities for triterpenoids biosynthesis.



**Figure 4. Comparative analyses of triterpene and diterpene biosynthetic pathways in *P. vulgaris* and *S. baicalensis*.**

**(A)** The gene copy number and expression heatmap of different organs (left column to right column: flower, leaf, root and stem) involved in the triterpene and diterpene biosynthetic pathways in *P. vulgaris* and *S. baicalensis*. TPS: Terpene Synthase. **(B)** Collinearity of OSCs in *P. vulgaris* and *S. baicalensis*.

Beyond the copy number variation, differential expression levels could also contribute to variations in terpenoids accumulation among different Lamiaceae plants. The expression levels of CYP716s are also relatively higher in different organs of *P. vulgaris* than in those of *S. baicalensis* (Figure 4A). In diterpenoid biosynthesis, the key diterpene skeleton-synthesizing enzyme, TPS, has six copies in *S. baicalensis*, whereas *P. vulgaris* has five copies. However, there is an expressional discrepancy, with all *P. vulgaris* TPSs exhibiting extremely low expression level in all organs, while two copies of *S. baicalensis* TPSs are highly expressed in stems (Figure 4A), and the two TPSs were proved to play a functional role in the *S. baicalensis* diterpenoid biosynthesis (Li *et al.*, 2023a). Therefore, the low expression of *P. vulgaris* TPSs is in line with the extremely low diterpenoids content in this plant.

## Discussion

### Multi-omics approaches and its potential in studying plant metabolism

The above-ground part of mature *P. vulgaris* has long been used as traditional Chinese medicine, as recorded in several ancient works on traditional Chinese medicine, including ‘Sheng Nong’s herbal classic’ (over 2,000 years ago) and ‘Compendium of Materia Medica’. Nowadays, it is a primary ingredient in herbal teas popular in South China. This work assembled a high-quality reference genome of *P. vulgaris*, and provided rich transcriptomic and metabolic data generated from different organs. These data will certainly provide valuable information for the cultivation, breeding and further utilization of this plant.

Compared with a previously published *P. vulgaris* genome (Zhang *et al.*, 2024), our genome assembly has a bigger genome size and contig N50 value, indicating enhanced completeness and consistency of genomic assembly (Supplementary Table S1). The high-quality assembly, together with a rich transcriptome dataset, allowed high-quality genome annotation, while the combined metabolomic data provided a reliable foundation for elucidating the triterpene saponin biosynthetic pathway. The discovery and characterization of the enzyme-encoding genes in *P. vulgaris* helps to better understand the evolution of the triterpene saponins metabolism in *Prunella*. These enzymes and the completely deciphered pathway are a valuable genetic source and could serve as the molecular basis of potential cell engineering and an alternative

choice for large-scale industrial production.

### **An optimized hierarchical pipeline of studying biosynthetic pathways**

To unravel the biosynthetic pathways of the two pentacyclic triterpene saponins in *P. vulgaris*, we propose a methodological protocol that leverages multi-omics data and integrates various approaches (phylogenetic analysis, gene expression assessment, metabolome-transcriptome correlation, and co-expression analyses) to efficiently screen for unknown enzyme-encoding genes. Functional characterization experiments confirmed the high accuracy of the analytical results, amongst which our identified OSCs align with the results of a previous study (Zhang *et al.*, 2024), validating the effectiveness of our screening strategy. Importantly, our integrated protocol for gene screening may extend beyond triterpene saponin biosynthesis or even plant metabolism, showing potential for broad applications. Our strategy is grounded on the evolutionary theory that orthologous or closely related genes are likely to have the same/similar functions. Expression assessment evaluates the transcription level of genes, typically correlating with phenotypes traits, for instance, gene expression levels and metabolite abundance in this study. Modern genetics and molecular biology also suggest that most genes, instead of working alone, function together with other genes in certain pathways or complex regulatory networks that mediate the organismal growth, development, reproduction, metabolism, physiology and response to stimuli (Li *et al.*, 2023c). So different genes must cooperate or interact with others to accomplish their missions. In fact, WGCNA (Langfelder and Horvath, 2008) was developed based on gene expressional correlations to build so-called transcriptional modules, and genes working in the same pathway are often classified into the same module.

Although the above-mentioned approaches were usually employed individually or separately in studies for the search for unknown genes, they are not used together as a combined protocol as in this study. Therefore, we developed this integrated screening protocol that combines different methods and approaches, and they together will generate results of high confidence. Admittedly, such screening strategy does require a larger volume of data and increased costs. However, for studies involving smaller gene families or studies where the screening objectives are more

clearly defined, certain steps can be reduced or skipped. Conversely, for studies involving larger gene families or with fewer defined targets, we are convinced the entire process remains a worthwhile and effective approach, leading to more precise outcomes, and this protocol should suit the identification of genes that participate in various pathways.

### **Evolutionary insights into CYP450 and UGT gene families: implications for triterpene saponin biosynthesis**

Previous research has shown that, although other CYP450 families (CYP51, CYP71, CYP72, CYP85, CYP87, CYP88, and CYP93) can modify the skeletons of oleanolic and ursolic acid-type saponins, CYP716 primarily catalyzes the oxidation at the C-28 position (Malhotra and Franke, 2022). This finding provides a clear direction for our screening efforts within the vast CYP450 superfamily. Such directional specificity is dictated by evolutionary principles, and the variations in enzyme activity align with the patterns of molecular evolution. Similarly, the UGT family is also extensive, with the UGT73 subfamily being selected based on these evolutionary clues. Utilizing evolutionary principles allows for efficient and rapid screening of target genes and may elucidate the mechanism underlying biosynthetic pathways (Huang *et al.*, 2024, Li *et al.*, 2024). This study represents a preliminary validation and investigation of the UGT73 subfamily, without an in-depth analysis of the glycosylation reactions in triterpene saponins specifically named Pruvuloside A and B in *P. vulgaris* ([Supplementary Table S2](#)). Based on the constructed gene co-expression networks, future research is expected to identify additional enzyme-encoding genes and incorporate potential regulatory transcription factors, thereby constructing a comprehensive regulatory network for the biosynthesis of triterpene saponins in *P. vulgaris*.

The distinct OSC and CYP716 copy numbers and expression levels between *P. vulgaris* and *S. baicalensis* provide insights for the discrepancy in major terpenoid metabolites between these two species. Although both genomes have experienced a recent but independent WGD, the subsequent retention patterns differed largely within their genomes, with genes involved in triterpenoids biosynthesis much better preserved in *P. vulgaris* than in *S. baicalensis*, reflecting a synchronous gene

retention/loss pattern in the same pathway/network (Shi *et al.*, 2020). Therefore, both gene copy number variation and expressional discrepancy might have led to the different terpenoids accumulation in the two plants, and possibly have contributed to the diversity of terpenoid compounds.

## **Materials and methods**

### **Sample collection, genome sequencing and assembly**

Fresh leaves and buds of *P. vulgaris* cultivated at the Institute of Botany, Jiangsu Province and the Chinese Academy of Sciences, Nanjing, were collected for genome sequencing. Short-read libraries with a 350-bp insert size were constructed and sequenced on the Illumina NovaSeq 6000 platform (Illumina, San Diego, CA, USA), while long-read libraries were constructed and sequenced on the Oxford Nanopore platform (Oxford Nanopore Technologies, Oxford, UK). Quality control involved the use of FastQC (V0.20.1) (Andrews, 2014) and Oxford Nanopore GUPPY (V4.0.2) (Krishnakumar *et al.*, 2018) to filter low-quality reads. Genome size evaluation employed GCE (V1.0.2) (Liu *et al.*, 2013) and K-mer frequency depth distribution was obtained using Jellyfish. The preliminary genome assembly was generated by NextDenovo (V2.4.0) (Jiang *et al.*, 2023) using 61.5 Gb of Nanopore reads. Subsequently, Oxford Nanopore long reads and 30.7 Gb of Illumina short reads were used to correct assembly error with Racon (V1.4.11) (Vaser *et al.*, 2017) and Pilon (V1.23) (Walker *et al.*, 2014). Purge\_Haplotigs (V1.0.4) (Roach *et al.*, 2018) was employed to de-hybridize the corrected genome, yielding the final assembly (Xue *et al.*, 2023a, Xue *et al.*, 2023b).

Fastp (V0.21.0) facilitated the filtering of raw data, followed by comparison with the reference genome and additional filtering. Sorting contigs and orienting scaffolding were accomplished by ALLHIC (V0.9.12) (Zhang *et al.*, 2019) and JUICER (V1.5) (Durand *et al.*, 2016) using 70.08 Gb Hi-C data. The completeness of the genome assembly was assessed using Benchmarking Universal Single-Copy Orthologs (BUSCO) (Manni *et al.*, 2021) (embryophyta\_odb10).

### **Genome annotation**

RepeatModeler (V2.0.2) (Bao and Eddy, 2002) was utilized to construct a *de novo*

repeat sequence database for the *P. vulgaris* genome, and then RepeatMasker (V4.1.0) (Tarailo-Graovac and Chen, 2009) was employed to search for repetitive sequences in the database. Gene structures were annotated using three methods: (I), *de novo* predictions using STAR (V2.7.10a) (Dobin *et al.*, 2012) and Braker (V2.1.6) (Brůna *et al.*, 2021). (II), homologous annotation using protein sequences of *Arabidopsis thaliana*, *Oryza sativa* and *Pogostemon cablin* as the reference with GenomeThreader (V1.7.1) (Gremme *et al.*, 2005). (III), Transcriptome-based annotation employing RNA-sequencing results with Hisat2 (Kim *et al.*, 2019), StringTie (Pertea *et al.*, 2015) and TransDecoder (V5.5.0) (Haas *et al.*, 2013). The integration of annotation results was performed using EvidenceModeler (Haas *et al.*, 2008). The assessment of genome annotation results employed BUSCO (V5.2.2) (Manni *et al.*, 2021).

### **Genome-wide identification and phylogenetic analysis of gene families**

For the evolutionary analyses of OSC, CYP450, and UGT gene families, functionally characterized genes were retrieved from the UniProt database (<https://www.uniprot.org/>). Genome-wide identification of the members within the three gene families began with a BLAST (V2.13.0) (Camacho *et al.*, 2009) search using an e-value threshold of 1e-5 within *P. vulgaris* and 17 additional angiosperm genomes. HMMER (V3.3.2) (Eddy, 2011) was subsequently used to validate the presence of characteristic structural domains within the BLAST-identified members (OSC: PF13243 and PF13249; CYP716: PF00067; UGT73: PF00201), and those lacking these domains were excluded. Amino acid sequences of the three families were aligned using ClustalW with default parameters (Thompson *et al.*, 1994) as integrated in MEGA (Tamura *et al.*, 2007). Phylogenetic analysis was conducted using the maximum likelihood program fasttree (V2.1.11) (Price *et al.*, 2010).

### **Transcriptome analysis and weighted co expression network analysis (WGCNA)**

The analysis of transcriptome data began with the utilization of Fastp (V0.23.2) (Chen *et al.*, 2018) to filter raw data from 15 samples corresponding to five plant organs (roots, stems, leaves, seeds and spikes). Subsequently, Hisat2 (Kim *et al.*, 2019) was employed to map all transcriptomic clean data to the reference genome using default parameters. Samtools (V1.15.1) (Danecek *et al.*, 2021) was used to sort and compress

the sam file to obtain the bam file. Rsubread (V2.8.2) (Liao *et al.*, 2019) facilitated the quantification of the TMM standardized expression of each sample's bam files, consolidating the count files of fifteen samples into a total expression matrix file. The R package WGCNA was applied for obtaining a weighted co-expression network of the expression matrix, employing a soft threshold of 9 and Pearson correlation coefficient algorithm. The identification of the most relevant module to triterpene saponin biosynthesis involved examining the distribution of identified candidate genes by phylogenetic analyses along with other enzyme-encoding genes in the same pathway within the WGCNA modules.

### ***Agrobacterium-mediated transient expression in Nicotiana benthamiana***

Total RNA was extracted from the leaves, stems and roots of *P. vulgaris* using the RNAprep Pure Plant Kit (Tiangen Biotech, China). Reverse transcription to cDNA was performed using the PrimeScript™ RT-PCR Kit (Takara Bio, Japan). Candidate sequences of OSCs, CYP716s and UGT73s were PCR-amplified from cDNA of *P. vulgaris* using Phanta Super-Fidelity DNA Polymerase (Vazyme, China), subcloned into the pClone007 Blunt Simple Vector (Tsingke Biotech, China) and sequenced (Sangon Biotech, China). Afterwards, PCR amplification with primers containing appropriate overhangs was applied, followed by gel purification of PCR products, which were then inserted into digested (XbaI/KpnI) pCAMBIA1300-GFP plasmid using ClonExpress® II One Step Cloning Kit (Vazyme, China). Assembled plasmid reactions were transformed into *Escherichia coli* DH5α cells and plated on selective LB agar plates (containing 50 µg ml<sup>-1</sup> kanamycin) for overnight growth at 37 °C. Colonies were screened using PCR and the sequences of PCR products were confirmed using Sanger sequencing. Positive transformants were then used to inoculate 2 ml of liquid LB cultures, which were then shaken overnight at 37 °C. Plasmids containing genes of interest were transformed into *Agrobacterium tumefaciens* GV3101 using the freeze-thaw method, plated onto selective LB agar plates (50 µg ml<sup>-1</sup> of kanamycin and 30 µg ml<sup>-1</sup> of gentamycin) and grown for 2 days at 30 °C. Positive transformants were verified through colony PCR and these were then inoculated into 2 ml of liquid LB cultures, which were shaken for 2 days at 30 °C, after which 25% glycerol stocks were prepared and stored at -80 °C for future use.

Functional characterization of candidate genes through *Agrobacterium*-mediated transformation in *N. benthamiana* was performed as described previously (Sainsbury et al., 2012). Briefly, *Agrobacterium* strains harbouring plasmid constructs of interest were grown at 30 °C at 200 rpm for 16 h in LB medium with kanamycin (50 µg ml<sup>-1</sup>), rifampicillin (25 µg ml<sup>-1</sup>) and gentamycin (25 µg ml<sup>-1</sup>). Cells were pelleted through centrifugation for 5 min at 8000 g and then suspended in *Agrobacterium* induction buffer (pH 5.6) containing 10 mM MES (2-[N-morpholino]-ethanesulfonic acid), 10 mM MgCl<sub>2</sub> and 150 µM acetosyringone and allowed to incubate at room temperature for 3 h. The concentrations of cell resuspensions were measured by taking their optical density OD<sub>600</sub> and combinations of strains of interest were then combined at a final OD<sub>600</sub> of 0.4 for each strain. Using a needleless syringe, these strain mixtures were infiltrated into the abaxial side of *N. benthamiana* leaves from 4–5-week-old plants, which were germinated and grown in soil in a growth chamber with a 16-h light and 8-h dark photo-cycle at 25 °C. Following infiltration, plants were grown for another 4 days under the same condition, after which leaves were excised for subsequent metabolite extraction. For a typical experiment, three leaves from three different plants were used for each strain mixture.

### **GC–MS analysis**

Two hundred mg aliquots of lyophilized *N. benthamiana* leaves were ground to a fine powder, and extracted three times with 1 ml methanol and 20 min sonication cycles at room temperature. They were then centrifuged for 10 min at 12,000 x g and 50 µL of supernatant was transferred to 1.5 mL sample vials and dried at 50°C. Samples were derivatized using 250 µL of Tri-Sil Z reagent (N-methyl N-(trimethylsilyl) trifluoroacetamide) prior to analysis. GC-MS analysis was performed using an Agilent 8860 fitted with a HP-5MS column (30 cm × 250 µm × 0.25 µm thickness) coupled to an Agilent 5977B mass selective detector. Ultra helium was used as the carrier gas at a flow rate 1.0 ml/min. The GC temperature program was set to 80°C for 1 min, followed by a gradient to 320°C at the increment of 20°C per minute and held at 320°C for an additional 17 min. The mass spectrometer was set to scan from 20 to 750 mass units with an initial 8-minute solvent delay. Data analysis was performed using MassHunter Qualitative Software (Agilent).

### **LC–MS analysis**

Samples were analyzed on an Agilent 1290 Infinity II UHPLC paired with a coupled Agilent 6546 Q-TOF mass spectrometer (6546 LC–MS). All samples were analyzed using electrospray ionization (ESI) in positive ionization mode. The instrument also had an in-line diode array detector (DAD) for routine analysis of UV active compounds (Agilent 1290 Infinity II DAD for 6546 LC–MS). UV data were typically collected at wavelengths of 210, 230, 254 and 280 nm (4 nm bandwidth for each) with reference to 360 nm (100 nm bandwidth). Reversed-phase (C18) analysis was predominantly performed on the 6546 LC–MS using a Poroshell 120 EC-C18 column (Agilent, 2.7  $\mu\text{m}$ , 4.6  $\times$  100 mm) with a flow rate of 0.4 mL min<sup>-1</sup> injecting 10  $\mu\text{L}$  per run. The mobile phase consisted of water + 0.1% formic acid (solvent A) and acetonitrile (solvent B) and began at 25% [B] for 5 min, followed by a gradient from 25-55% [B] from 5 to 10.0 min, 55-70% [B] from 10 to 15.0 min, 70-90% [B] from 15.0 to 20.0 min, 90-100% [B] from 2.0 to 30.0 min, 100-75% [B] from 30.0 to 35.0 min, 75-25% [B] from 35.0 to 37.0 min, and held at 25% [B] until 40.0 min. Data dependent MS<sup>2</sup> analysis was carried out from 10.0 to 14.0 min to detect the sugar nucleotides such as oleanolic acid sugar nucleotide.

### **Heterologous expression of UGT73 candidates in *E. coli*, protein preparation and *in vitro* enzyme assays**

The coding sequence of UGT73s (Pvul\_UGT73-2, Pvul\_UGT73-7 and Pvul\_UGT73-12) was amplified by PCR and annealed into the pET29a-MBP plasmid with hexa His-tag at the N-terminus. This plasmid construct was transformed into BL21 (DE3) and positive transformants were selected on LB medium plates containing kanamycin (50  $\mu\text{g ml}^{-1}$ ) through growth at 37 °C for 16 hours. Presence of the plasmid constructs was confirmed by colony PCR. A single, positive colony was picked and grown in 2 ml of culture of liquid LB medium at 37 °C. Following overnight of growth, 1 ml of the culture was used to inoculate 100 ml of LB medium and was grown at 37 °C until reaching an optical density at 600 nm (OD<sub>600</sub>) of 0.6. Then, a final concentration of 1 mM IPTG was added, and the culture was grown at 16°C for another 16 h. After centrifuging at 8,000 g for 20 min, the harvested cell pellets were resuspended in 20 ml of lysis buffer (50

mM Tris-HCl, pH 8.0). The samples were cooled in ice and the ultrasonication repeated at 100 W with the disruption period of 10 s pulse /5 s intervals for a duration of 30 min. The supernatant crude enzyme protein was snap frozen in liquid nitrogen and stored at  $-80^{\circ}\text{C}$  or used immediately.

Enzyme reactions with Pvul\_UGT73-2 crude protein were performed in Tris-HCl buffer (25 mM, containing 1 mM dithiothreitol, 1 mM  $\text{CaCl}_2$ , 5 mM  $\text{MgCl}_2$ , pH 8.6) and typically contained 700  $\mu\text{g}$  of total protein, 750  $\mu\text{M}$  UDP-D-glucose and 100  $\mu\text{M}$  of ursolic acid or oleanolic acid substrate (dissolved in DMSO) in a total reaction volume of 1000  $\mu\text{l}$ . Control reactions were performed by using vector proteins. Following addition of all components, reactions were incubated at  $30^{\circ}\text{C}$  for 1 h. The reaction was added with 500  $\mu\text{l}$  of MeOH to quench the reaction. Quenched reactions were then filtered and transferred into LC-MS/MS vials, as previously described.

### **Data availability**

The processed Illumina, NanoPore reads, the genome assembly, along with the gene models have been deposited at China National GeneBank DataBase (CNCBdb, <https://db.cngb.org/>) with BioProject ID: PRJCA024193.

### **Acknowledgments**

The authors thank Dr. Zhen Li at Ghent University for discussion. J.Y.X. acknowledges funding from the High-level Key Discipline Construction Project for Traditional Chinese Medicine—Resources Science of Chinese Medicinal Materials from National Administration of Traditional Chinese Medicine, Y.V.d.P. acknowledges funding from the Fundamental Research Funds for the Central Universities, the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 833522) and from Ghent University (Methusalem funding, BOF.MET.2021.0005.01), S.X. acknowledges the Open Fund of Jiangsu Key Laboratory for the Research and Utilization of Plant Resources (No. JSPKLB202201).

### **Author contributions**

J.Y.X., Y.V.d.P., S.X. and L.S. conceived the study. S.J.L. collected samples. S.J.L., B.Y.S. and T.L. assembled and annotated the genome, conducted phylogenetic and WGD analyses,

and performed analyses of candidate genes screening. S.X., Z.L. and X.Z. conducted experiments for the functional characterization of candidate genes. J.Y.X., S.X. and S.J.L. drafted the manuscript. Y.V.d.P., R.W. and L.S. participated in the revision of the manuscript. All authors read and approved the final manuscript.

**Conflict of interest statement**

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References:

- Andrews, S. (2014) FastQC A Quality Control tool for High Throughput Sequence Data.
- Bao, Z. and Eddy, S.R. (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res*, 12, 1269-1276.
- Brůna, T., Hoff, K.J., Lomsadze, A., Stanke, M. and Borodovsky, M. (2021) BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics*, 3.
- Bryson, A.E., Lanier, E.R., Lau, K.H., Hamilton, J.P., Vaillancourt, B., Mathieu, D., Yocca, A.E., Miller, G.P., Edger, P.P., Buell, C.R. and Hamberger, B. (2023) Uncovering a miltiradiene biosynthetic gene cluster in the Lamiaceae reveals a dynamic evolutionary trajectory. *Nat Commun*, 14, 343.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, 10, 421.
- Chen, S., Zhou, Y., Chen, Y. and Gu, J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34, i884-i890.
- China (2010) *Pharmacopoeia of the People's Republic of China* Beijing: China Medical Science Press.
- Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M. and Li, H. (2021) Twelve years of SAMtools and BCFtools. *GigaScience*, 10.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. (2012) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29, 15-21.
- Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S. and Aiden, E.L. (2016) Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems*, 3, 95-98.
- Eddy, S.R. (2011) Accelerated Profile HMM Searches. *PLoS computational biology*, 7, e1002195.
- Erthmann, P., Agerbirk, N. and Bak, S. (2018) A tandem array of UDP-glycosyltransferases from the UGT73C subfamily glycosylate sapogenins, forming a spectrum of mono- and bisdesmosidic saponins. *Plant molecular biology*, 97, 37-55.
- Gremme, G., Brendel, V., Sparks, M.E., Kurtz, S.J.I. and Technology, S. (2005) Engineering a software tool for gene structure prediction in higher organisms. 47, 965-978.
- Group, T.A.P., Chase, M.W., Christenhusz, M.J.M., Fay, M.F., Byng, J.W., Judd, W.S., Soltis, D.E., Mabberley, D.J., Sennikov, A.N., Soltis, P.S. and Stevens, P.F. (2016) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society*, 181, 1-20.
- Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., LeDuc, R.D., Friedman, N. and Regev, A. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols*, 8, 1494-1512.
- Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R. and Wortman, J.R. (2008) Automated eukaryotic gene structure annotation using EVIDENCEModeler and the program to assemble spliced alignments. *GENOME BIOLOGY*, 9.
- Hanlin, Z., Zhu, Y., Yanjun, C., Xiaoyun, L., Yubing, W., Zhengquan, H. and Chao, Z. (2022) Study on the Genome of *Prunella vulgaris* L. Combining High-throughput Sequencing and Karyotyping. *Molecular Plant Breeding*, 1-9.
- Huang, X.C., Tang, H., Wei, X., He, Y., Hu, S., Wu, J.Y., Xu, D., Qiao, F., Xue, J.Y. and Zhao, Y. (2024) The

gradual establishment of complex coumarin biosynthetic pathway in Apiaceae. *Nat Commun*, 15, 6864.

Jiang, H., Zhuo, W., Zongyi, S., Benxia, H., Adeola Oluwakemi, A., Fan, L., Jingjing, L., José, R.S., David, N.C., Kai, Y., Jue, R., Chuan-Le, X., De-Peng, W., Dong-Dong, W. and Sheng, W. (2023) An efficient error correction and accurate assembly tool for noisy long reads. *bioRxiv*, 2023.2003.2009.531669.

Kim, D., Paggi, J.M., Park, C., Bennett, C. and Salzberg, S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*, 37, 907-915.

Krishnakumar, R., Sinha, A., Bird, S.W., Jayamohan, H., Edwards, H.S., Schoeniger, J.S., Patel, K.D., Branda, S.S. and Bartsch, M.S. (2018) Systematic and stochastic influences on the performance of the MinION nanopore sequencer across a range of nucleotide bias. *Sci Rep*, 8, 3159.

Krokida, A., Delis, C., Geisler, K., Garagounis, C., Tsikou, D., Peña-Rodríguez, L.M., Katsarou, D., Field, B., Osbourn, A.E. and Papadopoulou, K.K. (2013) A metabolic gene cluster in *Lotus japonicus* discloses novel enzyme functions and products in triterpene biosynthesis. *The New phytologist*, 200, 675-690.

Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC BIOINFORMATICS*, 9.

Li, H., Wu, S., Lin, R., Xiao, Y., Malaco Morotti, A.L., Wang, Y., Galilee, M., Qin, H., Huang, T., Zhao, Y., Zhou, X., Yang, J., Zhao, Q., Kanellis, A.K., Martin, C. and Tatsis, E.C. (2023a) The genomes of medicinal skullcaps reveal the polyphyletic origins of clerodane diterpene biosynthesis in the family Lamiaceae. *Molecular Plant*, 16, 549-570.

Li, P., Lv, X., Wang, J., Zhang, C., Zhao, J. and Yang, Y. (2023b) Research on the anti-ageing mechanism of *Prunella vulgaris* L. *Sci Rep*, 13, 12398.

Li, Q., Dai, Y., Huang, X.C., Sun, L., Wang, K., Guo, X., Xu, D., Wan, D., An, L., Wang, Z., Tang, H., Qi, Q., Zeng, H., Qin, M., Xue, J.Y. and Zhao, Y. (2024) The chromosome-scale assembly of the *Notopterygium incisum* genome provides insight into the structural diversity of coumarins. *Acta pharmaceutica Sinica B*, 14, 3760-3773.

Li, S., Nakayama, H. and Sinha, N.R. (2023c) How to utilize comparative transcriptomics to dissect morphological diversity in plants. *Current opinion in plant biology*, 76, 102474.

Liao, Y., Smyth, G.K. and Shi, W. (2019) The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res*, 47, e47.

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., Li, Z., Chen, Y., Mu, D. and Fan, W.J.Q.B. (2013) Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. 35, 62-67.

Liu, C., Smit, S.J., Dang, J., Zhou, P., Godden, G.T., Jiang, Z., Liu, W., Liu, L., Lin, W., Duan, J., Wu, Q. and Lichman, B.R. (2023a) A chromosome-level genome assembly reveals that a bipartite gene cluster formed via an inverted duplication controls monoterpenoid biosynthesis in *Schizonepeta tenuifolia*. *Molecular Plant*, 16, 533-548.

Liu, C., Smit, S.J., Dang, J., Zhou, P., Godden, G.T., Jiang, Z., Liu, W., Liu, L., Lin, W., Duan, J., Wu, Q. and Lichman, B.R. (2023b) A chromosome-level genome assembly reveals that a bipartite gene cluster formed via an inverted duplication controls monoterpenoid biosynthesis in *Schizonepeta tenuifolia*. *Mol Plant*, 16, 533-548.

Ma, Y., Cui, G., Chen, T., Ma, X., Wang, R., Jin, B., Yang, J., Kang, L., Tang, J., Lai, C., Wang, Y., Zhao, Y., Shen, Y., Zeng, W., Peters, R.J., Qi, X., Guo, J. and Huang, L. (2021) Expansion within the CYP71D subfamily drives the heterocyclization of tanshinones synthesis in *Salvia miltiorrhiza*. *Nat Commun*, 12, 685.

Malhotra, K. and Franke, J. (2022) Cytochrome P450 monooxygenase-mediated tailoring of triterpenoids and steroids in plants. *Beilstein journal of organic chemistry*, 18, 1289-1310.

Manni, M., Berkeley, M.R., Seppey, M. and Zdobnov, E.M. (2021) BUSCO: Assessing Genomic Data Quality and Beyond. *Current protocols*, 1, e323.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T. and Salzberg, S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *NATURE BIOTECHNOLOGY*, 33, 290-+.

Price, M.N., Dehal, P.S. and Arkin, A.P. (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One*, 5, e9490.

Rahimi, S., Kim, J., Mijakovic, I., Jung, K.H., Choi, G., Kim, S.C. and Kim, Y.J. (2019) Triterpenoid-biosynthetic UDP-glycosyltransferases from plants. *Biotechnology advances*, 37, 107394.

Roach, M.J., Schmidt, S.A. and Borneman, A.R. (2018) Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, 19, 460.

Shi, T., Rahmani, R.S., Gugger, P.F., Wang, M., Li, H., Zhang, Y., Li, Z., Wang, Q., Van de Peer, Y., Marchal, K. and Chen, J. (2020) Distinct Expression and Methylation Patterns for Genes with Different Fates following a Single Whole-Genome Duplication in Flowering Plants. *Molecular biology and evolution*, 37, 2394-2413.

Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. *Molecular biology and evolution*, 24, 1596-1599.

Tarailo-Graovac, M. and Chen, N. (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics*, Chapter 4, Unit 4.10.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22, 4673-4680.

Vaser, R., Sović, I., Nagarajan, N. and Research, M.Š.J.G. (2017) Fast and accurate de novo genome assembly from long uncorrected reads.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K. and Earl, A.M. (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. In *PLoS One*, pp. e112963.

Wang, L., Lee, M., Sun, F., Song, Z., Yang, Z. and Yue, G.H. (2022) A chromosome-level genome assembly of chia provides insights into high omega-3 content and coat color variation of its seeds. *Plant Communications*, 3.

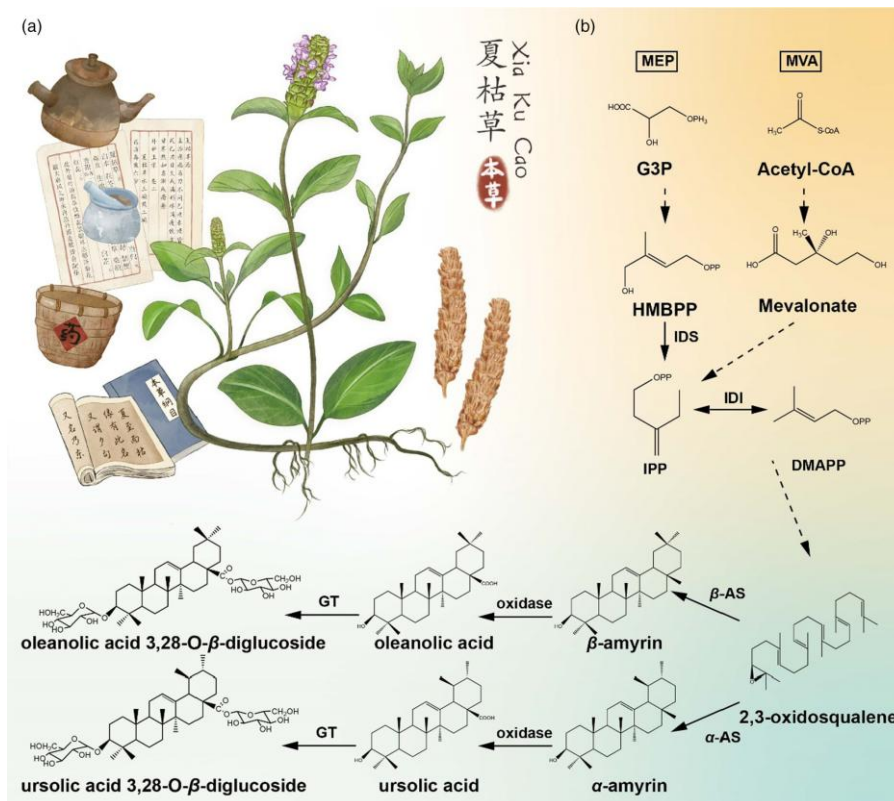
Wang, S.J., Wang, X.H., Dai, Y.Y., Ma, M.H., Rahman, K., Nian, H. and Zhang, H. (2019) *Prunella vulgaris*: A Comprehensive Review of Chemical Constituents, Pharmacological Effects and Clinical Applications. *Current pharmaceutical design*, 25, 359-369.

Wu, Z., Raven, P.H., Hong, D. and Missouri Botanical, G. (1900) *Flora of China* Beijing, St. Louis: Science Press ; Missouri Botanical Garden Press.

Xue, J.-Y., Fan, H.-Y., Zeng, Z., Zhou, Y.-H., Hu, S.-Y., Li, S.-X., Cheng, Y.-J., Meng, X.-R., Chen, F., Shao, Z.-Q. and Van de Peer, Y. (2023a) Comprehensive regulatory networks for tomato organ development based on the genome and RNAome of MicroTom tomato. *Horticulture Research*, 10.

Xue, J.Y., Li, Z., Hu, S.Y., Kao, S.M., Zhao, T., Wang, J.Y., Wang, Y., Chen, M., Qiu, Y., Fan, H.Y., Liu, Y., Shao, Z.Q. and Van de Peer, Y. (2023b) The *Saururus chinensis* genome provides insights into the evolution of pollination strategies and herbaceousness in magnoliids. *The Plant journal : for cell and molecular biology*, 113, 1021-1034.

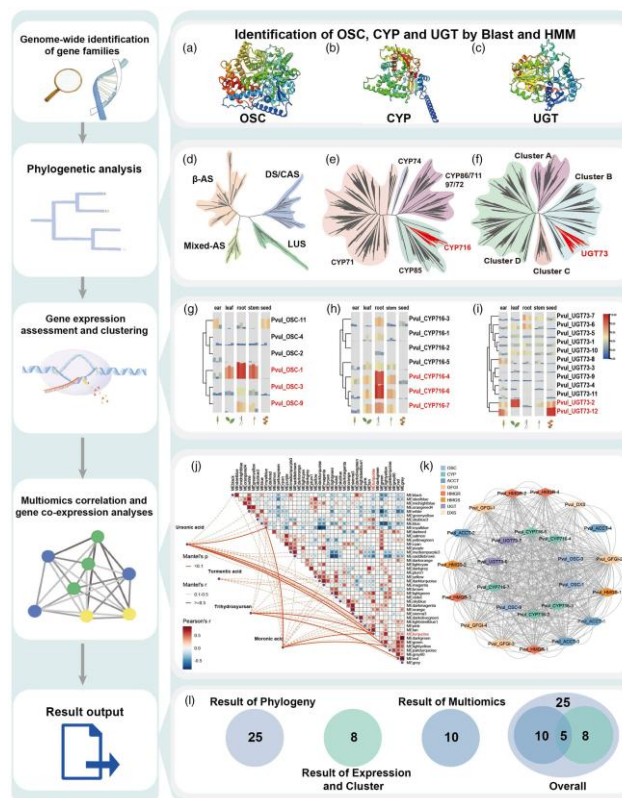
- Yang, W. (2015) A Study on the Reproduction and Growth Characteristics of *Prunella vulgaris* L.: Nanjing Agricultural University.
- Zhang, S., Meng, F., Pan, X., Qiu, X., Li, C. and Lu, S. (2024) Chromosome-level genome assembly of *Prunella vulgaris* L. provides insights into pentacyclic triterpenoid biosynthesis. *118*, 731-752.
- Zhang, X., Zhang, S., Zhao, Q., Ming, R. and Tang, H. (2019) Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat Plants*, *5*, 833-845.
- Zhao, Q., Yang, J., Cui, M.Y., Liu, J., Fang, Y., Yan, M., Qiu, W., Shang, H., Xu, Z., Yidiresi, R., Weng, J.K., Pluskal, T., Vigouroux, M., Steuernagel, B., Wei, Y., Yang, L., Hu, Y., Chen, X.Y. and Martin, C. (2019) The Reference Genome Sequence of *Scutellaria baicalensis* Provides Insights into the Evolution of Wogonin Biosynthesis. *Mol Plant*, *12*, 935-950.
- Zhou, J., Hu, T., Gao, L., Su, P., Zhang, Y., Zhao, Y., Chen, S., Tu, L., Song, Y., Wang, X., Huang, L. and Gao, W. (2019) Friedelane-type triterpene cyclase in celastrol biosynthesis from *Tripterygium wilfordii* and its application for triterpenes biosynthesis in yeast. *The New phytologist*, *223*, 722-735.



**Figure 1.** Morphology of *Prunella vulgaris* and its hypothetical pentacyclic saponins biosynthetic pathway.

(a) Morphology of *P. vulgaris*.

(b) The proposed pentacyclic saponins biosynthetic pathway in plants. Dashed lines represent multiple steps, solid lines represent a single step. G3P, glyceraldehyde 3-phosphate; GT, glycosyltransferase; HMBPP, dimethylallyl pyrophosphate; IDI, isopentenyl diphosphate isomerase; IDS, isoprenyl diphosphate synthase; MEP, methylerythritol phosphate pathway; MVA, mevalonate pathway;  $\alpha$ -AS,  $\alpha$ -amyrin synthase;  $\beta$ -AS:  $\beta$ -amyrin synthase.



**Figure 2.** Flow chart and results of screening enzyme-encoding genes involved in the biosynthesis of ursolic and oleanolic acid saponins in *Prunella vulgaris*.

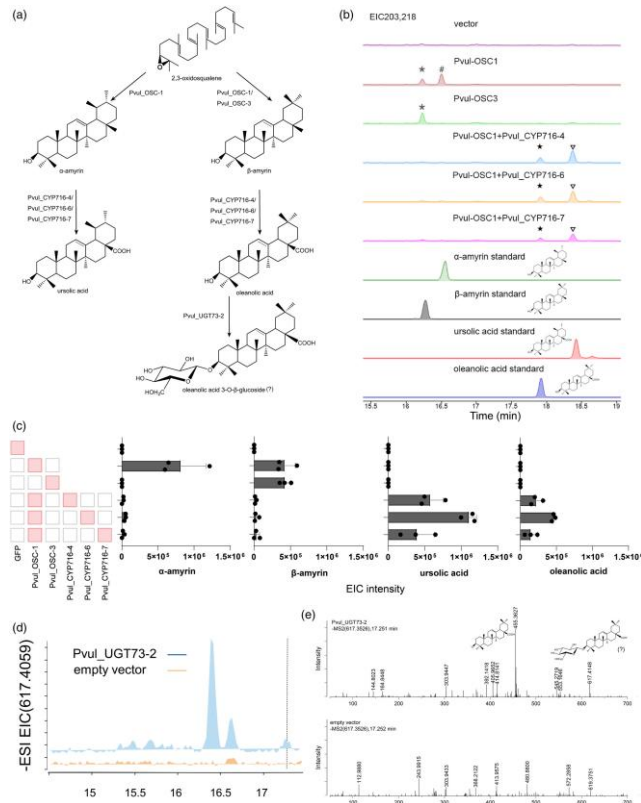
The left column indicates five steps of our integrated protocol for screening target genes, while the right part displays the results generated at each screening step. Genome-wide identification of OSC (a), CYP450 (b), and UGT (c) genes was performed first by the blast and HMM searches from the genomes of *P. vulgaris* and other 18 species. Subsequently, phylogenetic analyses of OSC (d), CYP450 (e), and UGT (f) genes were performed using the sequences of all identified genes and characterized enzymes collected from published literatures. Subfamily branches are distinguished by unique colors and shaded bands. The detailed phylogenies and IDs for the candidate genes are provided in Figures S5–S7. The screened candidate genes by phylogenetic analyses continued to be applied to the assessment of their expression in different organs of *P. vulgaris*.

Genes were clustered according to their differential expression levels and highly expressed genes were selected, resulting in three OSCs (g), three CYP716s (h), and two UGT73s (i).

Metabolome–transcriptome correlation analysis was conducted to associate metabolic products with genes according to the content of metabolites and gene expression patterns (j).

Connections between modules and metabolites indicate a *P*-value less than 0.1, with the thickness of the lines representing the degree of correlation. Genes were classified into different modules by distinct co-expression patterns by WGCNA (Figure S8). Among the modules showing a significant correlation with ursolic acid and oleanolic acid, the turquoise module contained three OSCs, five CYP716s, and two UGT73s (Table S5), as well as most enzyme-encoding genes in the upstream process of triterpene saponins biosynthesis (k). Hence, this step produced 10 candidates.

At last, the screening pipeline generated a total of 36 candidate genes potentially participating in the triterpene saponins biosynthesis in *P. vulgaris*, and five of them were repeatedly screened by multiple analyses, thus were considered as the most promising candidates (l).



**Figure 3.** Functional characterization of *Prunella vulgaris* candidate genes involved in triterpene biosynthesis.

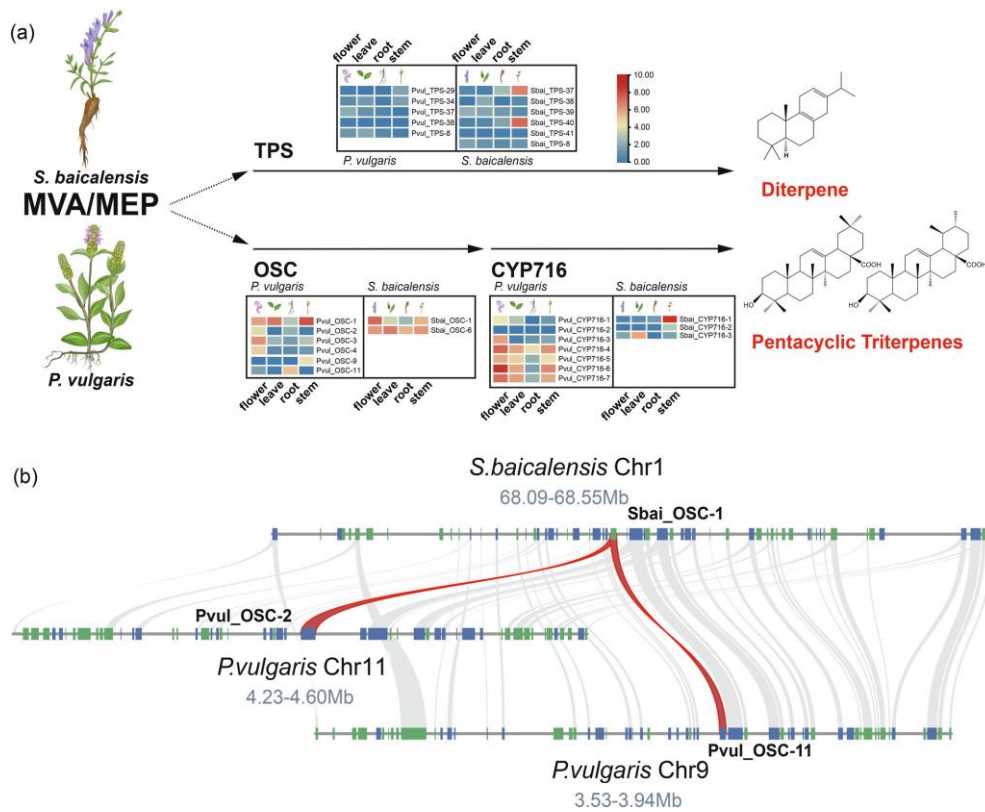
(a) Biosynthetic route leading to the production of  $\alpha$ -amyrin,  $\beta$ -amyrin, ursolic acid, and oleanolic acid, as well as the putative glycosylation product of oleanolic acid (oleanolic acid 3-O- $\beta$ -glucoside) in *P. vulgaris*. *PvuL\_OSC-1/-3*, oxidosqualene cyclase; *PvuL\_CYP716-4/-6/-7*, cytochrome P450 proteins; *PvuL\_UGT73-2*, UDP-dependent glycosyltransferase.

(b) GC-MS extracted ion chromatograms (EIC) for leaf extracts of *Nicotiana benthamiana* after expression of each OSC or co-expression of the *PvuL\_OSC-1* with the different *PvuL\_CYP716*. All intermediates were validated by comparison to synthetic authentic standards.

(c) The extracted ion abundance for the exact ion mass of products produced in *N. benthamiana* after expression of each *PvuL\_OSC* or co-expression of the *PvuL\_OSC-1* with the different *PvuL\_CYP716*. Data are means  $\pm$  SD;  $n = 3$  biological replicates.

(d) *In vitro* assay of *PvuL\_UGT73-2* protein expressed in *Escherichia coli*. Shown are HPLC-MS chromatograms representing oleanolic acid (5), as well as one mass feather ( $m/z$  617.4636) that might be as putative product of *PvuL\_UGT73-2*.

(e) MS<sup>2</sup> spectra of the new compound produced by *PvuL\_UGT73-2*, along with predicated ion fragment structure.



**Figure 4.** Comparative analyses of triterpene and diterpene biosynthetic pathways in *Prunella vulgaris* and *Scutellaria baicalensis*.

(a) The gene copy number and expression heatmap of different organs (left column to right column: flower, leaf, root, and stem) involved in the triterpene and diterpene biosynthetic pathways in *P. vulgaris* and *S. baicalensis*. TPS, terpene synthase.

(b) Collinearity of OSCs in *P. vulgaris* and *S. baicalensis*.

## Supporting information

Supplementary Figure S1. Genomic survey of *P. vulgaris*.

Supplementary Figure S2. Circos diagram of the *P. vulgaris* genome.

Supplementary Figure S3. Cluster diagram of overall metabolites.

Supplementary Figure S4. K-Means diagram of differential metabolites.

Supplementary Figure S5. Phylogeny of OSC genes extracted from 18 plant genomes and functionally characterized enzyme-encoding genes.

Supplementary Figure S6. Detailed Phylogeny of CYP716 genes.

Supplementary Figure S7. Detailed phylogeny of UGT73 genes.

Supplementary Figure S8. Gene co-expression module classification by WGCNA using transcriptomic data of roots, stems, leaves, seeds and spikes (three replicates for each organ) of *P. vulgaris*.

Supplementary Figure S9. Functional characterization of Pvul\_OSC-3.

Supplementary Figure S10. Functional characterization of Pvul\_OSC\_1 and Pvul\_CYP716s.

Supplementary Figure S11. Synteny of characterized OSCs and CYP716s in *P. vulgaris* with their orthologs in *S. baicalensis*.

Supplementary Figure S12. Whole-genome duplication (WGD) events analysis in *P. vulgaris* and three other plants.

Supplementary Table S1. Statistics for *P. vulgaris* genome assembly in Zhang et al. (2024) and this study.

Supplementary Table S2. Triterpenoids detected in *P. vulgaris* by metabolome.

Supplementary Table S3. IDs and renames of OSC, CYP716, UGT73 genes, and corresponding functional genes in 18 species.

Supplementary Table S4. Plant genome used in this study.

Supplementary Table S5. Distribution of potential *P. vulgaris* enzyme-encoding genes involved in the triterpenoid saponin biosynthesis in different co-expression modules.

Supplementary Table S6. Gene expression profile of *P. vulgaris*.

Supplementary Table S7. PCR Primers used in this study.