Plant Communications
Research article

CellPress
Partner Journal

A near-complete assembly of the *Houttuynia cordata* genome provides insights into the regulatory mechanism of flavonoid biosynthesis in Yuxingcao

Zhengting Yang[1,4,*], Fayin He[1,4], Yingxiao Mai[2,4], Sixian Fan[1], Yin An[1], Kun Li[1], Fengqi Wu[2], Ming Tang[1], Hui Yu[1], Jian-Xiang Liu[3,*] and Rui Xia[2,*]

[1]Key Laboratory of State Forestry Administration on Biodiversity Conservation in Karst Mountainous Areas of Southwestern China, School of Life Sciences, Guizhou Normal University, Guiyang, Guizhou 550025, China

[2]State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, College of Horticulture, South China Agricultural University, Guangzhou, Guangdong 510640, China

[3]Lishui Innovation Center for Life and Health, Zhejiang University, Hangzhou 310027, China

[4]These authors contributed equally to this article.

***Correspondence: Zhengting Yang (zhengtingyang@gznu.edu.cn), Jian-Xiang Liu (jianxiangliu@zju.edu.cn), Rui Xia (rxia@scau.edu.cn)**

https://doi.org/10.1016/j.xplc.2024.101075

ABSTRACT

***Houttuynia cordata*, also known as Yuxingcao in Chinese, is a perennial herb in the Saururaceae family. It is highly regarded for its medicinal properties, particularly in treating respiratory infections and inflammatory conditions, as well as boosting the human immune system. However, a lack of genomic information has hindered research on the functional genomics and potential improvements of *H. cordata*. In this study, we present a near-complete assembly of *H. cordata* genome and investigate the biosynthetic pathway of flavonoids, specifically quercetin, using genomics, transcriptomics, and metabolomics analyses. The genome of *H. cordata* diverged from that of *Saururus chinensis* around 33.4 million years ago; it consists of 2.24 Gb with 76 chromosomes (4*n* = 76) and has undergone three whole-genome duplication (WGD) events. These WGDs played a crucial role in shaping the *H. cordata* genome and influencing the gene families associated with its medicinal properties. Through metabolomics and transcriptomics analyses, we identified key genes involved in the β-oxidation process for biosynthesis of houttuynin, one of the volatile oils responsible for the plant's fishy smell. In addition, using the reference genome, we identified genes involved in flavonoid biosynthesis, particularly quercetin metabolism, in *H. cordata*. This discovery has important implications for understanding the regulatory mechanisms that underlie production of active pharmaceutical ingredients in traditional Chinese medicine. Overall, the high-quality genome assembly of *H. cordata* serves as a valuable resource for future functional genomics research and provides a solid foundation for genetic improvement of *H. cordata* for the benefit of human health.**

Key words: *Houttuynia cordata*, flavonoid biosynthesis, genome assembly, houttuynin, quercetin, whole-genome duplication

Yang Z., He F., Mai Y., Fan S., An Y., Li K., Wu F., Tang M., Yu H., Liu J.-X., and Xia R. (2024). A near-complete assembly of the *Houttuynia cordata* genome provides insights into the regulatory mechanism of flavonoid biosynthesis in Yuxingcao. Plant Comm. **5**, 101075.

INTRODUCTION

Houttuynia cordata, commonly known as Yuxingcao or fishy-smelling herb in China, is a perennial herbaceous plant from the Saururaceae family (Bahadur Gurung et al., 2021; Yin et al., 2023). It is the only species in the *Houttuynia* genus of the Saururaceae family (Han, 1995) and is found in various regions, including China, Japan, Korea, northeastern India, and Southeast Asia (Xu et al., 2021; Luo et al., 2022; Wei et al.,

Published by the Plant Communications Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.

2024). It typically grows in moist to soggy soils on shady hillsides, waysides, and ridges at altitudes ranging from 300 to 2600 m (Luo et al., 2022). *H. cordata* has a unique reproductive method, relying primarily on underground rhizome formation and parthenogenesis instead of traditional sexual reproduction (Laldinsangi, 2022). It holds great value as both a culinary ingredient and a medicinal herb, with numerous studies showing its notable antiviral activity (Yuan et al., 2022; Jiu et al., 2023).

The medicinal properties of *H. cordata* are largely attributed to its volatile oils and flavonoid compounds, which have been extensively studied for their antibacterial, anti-inflammatory, antiviral, and antioxidant effects (Laldinsangi, 2022; Rafiq et al., 2022; Pradhan et al., 2023; Wei et al., 2024). The primary components of the volatile oil in *H. cordata* include houttuynin, β-pinene, 2-undecanone (methyl nonyl ketone), ethyl caprylate, and α-pinene (Yang et al., 2019; Lin et al., 2022). Houttuynin is particularly noteworthy for its unique fishy smell and potent antibacterial properties. Its derivative, sodium houttuyfonate (SH), is water stable and exhibits a broad range of antibacterial, anti-inflammatory, antioxidant, and antitumor activities against various pathogens (Liu et al., 2021b; Shen et al., 2021; Cheng et al., 2023; He et al., 2023). In addition, SH has been found to modulate the immune system and intestinal flora (Zhang et al., 2020). During the process of steam distillation, SH undergoes a transformation into 2-undecanone, which further enhances its therapeutic potential by activating the Nrf2/HO-1/NQO-1 signaling pathway (Lou et al., 2019). Despite the significant medical benefits of SH, its biosynthetic pathway remains unknown, leading to the need for artificial synthesis.

Flavonoids are vital plant secondary metabolites synthesized via the phenylpropanoid pathway by enzymes such as chalcone synthase (CHS) and flavonoid 3-hydroxylase (F3H) (Nabavi et al., 2020; Dong and Lin, 2021). These compounds, which include flavonols, flavones, and anthocyanins, exhibit antioxidant, antibacterial, and antiviral properties (Dias et al., 2021; Ekalu and Habila, 2020; Liu et al., 2021a). Research has demonstrated that flavonoids from *H. cordata* can alleviate lung inflammation and mitigate H1N1-induced lung injury in mice by inhibiting influenza virus and Toll-like receptor signaling (Hung et al., 2015; Lee et al., 2015; Zhou et al., 2022). Quercetin, a flavonol found in *H. cordata*, is efficiently absorbed in the human digestive system, particularly in the small intestine and stomach (Michala and Pritsa, 2022). This compound forms glycosides such as quercitrin, isoquercitrin, baimaside, hyperoside, rutin, and isohyperoside, which have demonstrated remarkable medicinal properties. Recent studies have highlighted quercetin's protective effects against UVB-induced skin damage and lung and liver injuries, as well as its potential to inhibit tumor growth (Mapoung et al., 2021; Wang et al., 2022a; Pradhan et al., 2023).

Advances in sequencing technology have made it possible and affordable to sequence the genomes of medicinal plants with large genome sizes. This breakthrough has greatly enhanced research on medicinal plants at the molecular and genetic levels (Cheng et al., 2021b). In line with this progress, the "1K Medicinal Plant Genome Project" has been proposed, aiming to complete the collection and sequencing of 1000 important medicinal plant genomes within 3–5 years (Su et al., 2022). The 1K Medicinal Plant Genome Database (1K-MPGD) is now available at http://www.herbgenome.com/. Within the Saururaceae family are six species spread across four relictual genera: *Saururus*, *Gymnotheca*, *Anemopsis*, and *Houttuynia* (Han, 1995). Many of these species are herbaceous plants. However, only the diploid genome of *Saururus chinensis* ($2n = 22$) has been reported to date (Xue et al., 2023). *H. cordata* is a perennial polyploid herb, existing in diploid or tetraploid form, and features a diverse chromosome count, ranging from 18 to 108 (Luo et al., 2022). Despite its broad genetic diversity, the precise chromosome number and ploidy level of *H. cordata* remain undefined at the genomic level. This absence of detailed genomic data limits the effective utilization of this medicinal plant. Therefore, it is crucial to decode the complete genome of *H. cordata* to gain a comprehensive understanding of its genome structure. This information will also facilitate the dissection of biosynthetic pathways for bioactive compounds, aiding genomics-assisted breeding and herbal synthetic biology.

In the current study, we sequenced and assembled a high-quality genome of *H. cordata*, achieving an exceptional chromosome-level assembly in its field. We then examined the phylogenetic relationships and evolution of *H. cordata*, explored whole-genome duplications (WGDs) in this species, and identified crucial components of the flavonoid biosynthetic pathway. Leveraging the assembled genome, together with transcriptomic and metabolomic analyses, we identified key genes responsible for houttuynin biosynthesis and quercetin metabolism in *H. cordata*. This study not only broadens our understanding of the genetic and metabolic complexity of *H. cordata* but also establishes a foundational genome resource that can catalyze future functional genomics studies and enhance the pharmacological exploitation of this important herbaceous plant.

RESULTS

Sequencing, assembly, and annotation of the genome

H. cordata consists of underground white rhizomes, ovate-shaped leaves above ground, four white bracts, and a pale-yellow spike inflorescence without petals (Figure 1A). According to the genome survey, the identification of multiple coverage peaks in the k-mer depth distribution indicated the potential polyploidy of *H. cordata* (Supplemental Figure 1A). Moreover, the coverage of the AAAB (0.32) type was greater than that of the AABB (0.07) type in Smudgeplot, providing strong evidence for autotetraploidy (Supplemental Figure 1B). The haplotype size (n) was estimated as approximately 600 Mb by k-mer analysis (Supplemental Table 1), which was consistent with the total genome size ($4n$) estimated by flow cytometry (1.9–2.5 Gb) (Supplemental Figure 2).

Using a multi-platform approach, we sequenced and *de novo* assembled the genome of *H. cordata* (Figure 1B). The selected plant was identified as an autotetraploid with a predicted heterozygosity rate exceeding 2% and an estimated genome size of 2.4 Gb (Supplemental Figure 1), consistent with the range of 1.9–2.5 Gb estimated by flow cytometry (Supplemental Figure 2). We obtained 65 Gb of PacBio high-fidelity (HiFi) reads by optimizing circular consensus sequencing

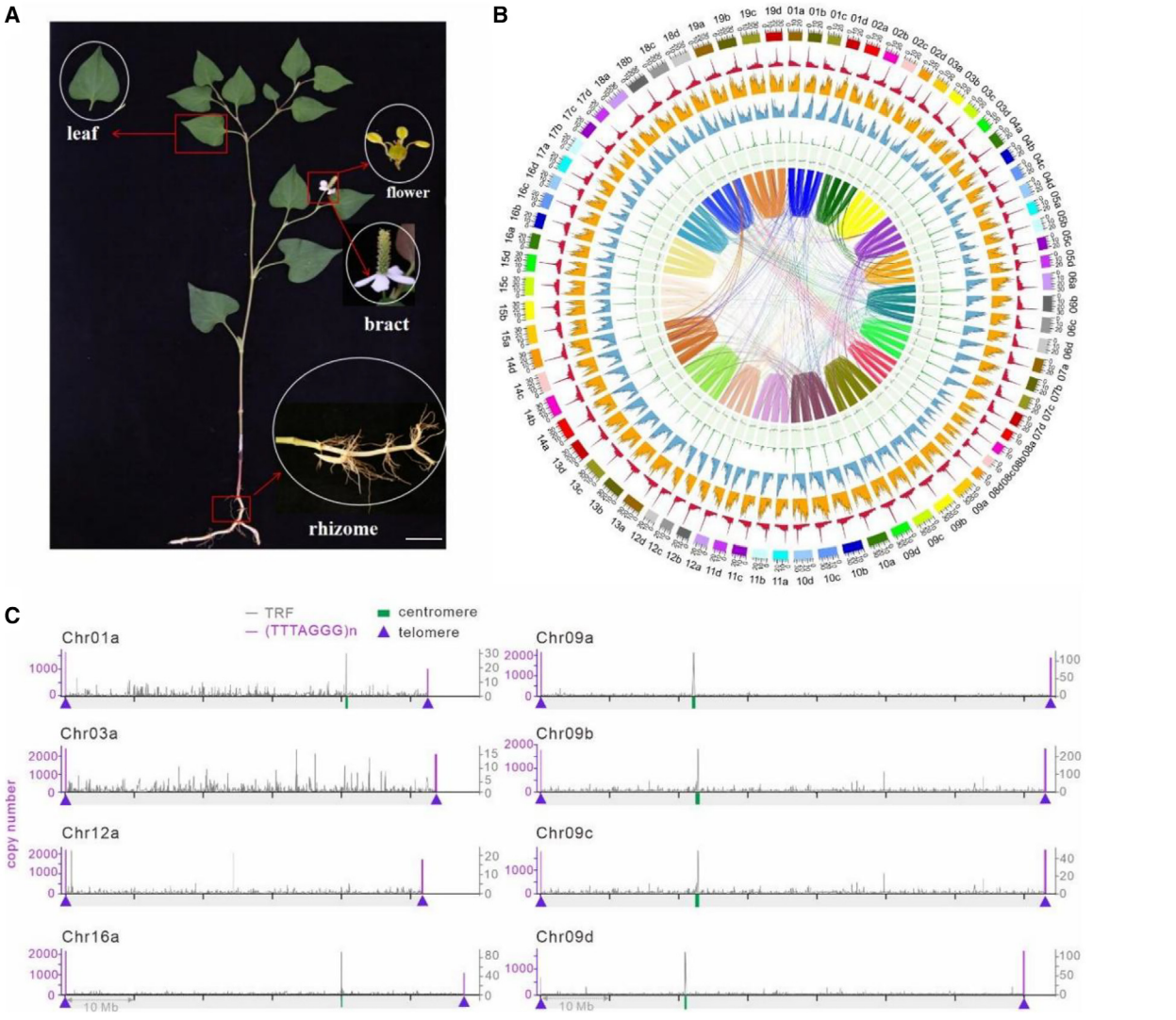

Figure 1. Morphological and genomic features of *H. cordata*.

(A and B) (A) Overview of the entire *H. cordata* plant, with enlarged images of the leaf, inflorescences, bract, and rhizome. Scale bar, 10 cm. **(B)** Features of the assembled *H. cordata* genome. The outer to inner circles represent chromosome-scale pseudochromosomes (chr01–19[a–d]), class I transposable element (TE) density, class II TE density, coding gene density, GC content, collinear blocks, and window size. Each linking line in the center of the plot represents a pair of homologous genes. TE density refers to the density of TEs in a specific genomic region, and coding gene density refers to the density of protein-coding genes. The proportion of tandemly repeated DNA sequences in a given genomic region is represented by the tandem repeat proportion. GC content indicates the proportion of guanine (G) and cytosine (C) nucleotides in a specific genomic region. Collinear blocks, with a minimum length of 100 kb, represent collinear genomic segments that share homologous DNA sequences. The genomic region being analyzed is divided into windows of 500 kb each, represented by the window size. The four sets of haplotype chromosomes are denoted by a–d.

(C) Locations of telomere and centromere regions in selected chromosomes. The purple line in each subgraph represents the distribution of "TTTAGGG" copies on each chromosome, with the positions of telomeres marked by purple triangles. The gray line in each subgraph represents the copy number of tandem repeats (50–1000 bp monomer) on each chromosome, with possible centromere positions marked by green rectangles. TRF, Telomere Repeats Finder.

(CCS) technology, which enabled us to generate highly accurate (99.94%) long HiFi reads with an average length of 16 kb. These assembled HiFi reads were then grouped, ordered, and oriented into pseudomolecules (hereafter referred to as Hi-C pseudomolecules) This initial assembly yielded a draft genome of 2.24 Gb, consisting of 86 contigs with a contig N50 length of 28.73 Mb. Subsequently, we used 111 Gb (~50×) of chromosome conformation capture (Hi-C) data to scaffold these contigs into 78 pseudomolecules, representing 76 chromosomes, one mitochondrial genome, and one chloroplast genome (Supplemental Table 1 and Figure 1B). The resulting scaffolded assembly was found to be consistent with the karyotype analysis (Supplemental Figure 3), confirming the presence of chr01 [abcd] to chr19 [abcd] ($4n$ = 76) (Supplemental Figures 4–6). At this stage, eight gaps were identified within the pseudomolecules. To enhance the assembly quality, we polished the pseudochromosome sequences using Illumina short reads and attempted gap filling using HiFi long reads. Ultimately, this process produced a 2.24-Gb genome of *H. cordata*, consisting of 78 pseudomolecules with a pseudomolecule N50 length of 29.19 Mb and a GC content

of 39.57% (Supplemental Table 1). Comparison of the haplotype sequences of the autotetraploid revealed a high level of comparability among the homologous chromosomes, with an alignment rate of 84%–86%, and abundant structural variations such as inversions, translocation, and duplications were also detected (Supplemental Table 1 and Supplemental Figure 7).

To predict gene models, we used a combination of *de novo* prediction, homology-based searches, and RNA-sequencing data, resulting in the prediction of 126 864 protein-coding genes in *H. cordata*. The primary gene models exhibited a mean length of 1629 bp, with an average exon length of 293 bp and an average intron length of 589 bp (Supplemental Table 1). On the basis of a Benchmarking Universal Single-Copy Orthologs (BUSCO) assessment, the completeness of the gene annotations was approximately 99.1% (Supplemental Table 2). We functionally annotated the protein-coding genes using various databases, including Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), Swiss-Prot, and others, an over 98.66% of the genes could be mapped to corresponding functional items, with only 0.82% remaining unannotated (Supplemental Table 3). These analyses indicate a relatively comprehensive structural annotation of the reference genome. RepeatMasker was used to identify repetitive sequences, which accounted for 53.60% of the genome. Long terminal repeat (LTR) retrotransposons were the predominant type of repetitive element, comprising 33.93% of the genome, including 4.12% Copia, 22.64% Gypsy, and 7.18% unknown. The LTR assembly index (LAI) was estimated as 15.49, consistent with the criteria for a reference genome (Ou et al., 2018) (Supplemental Table 4).

Telomeres and centromeres play essential roles in maintaining genome stability. Telomeres, located at the ends of chromosomes, consist of repeated sequences (TTTAGGG) and help to protect the integrity of chromosomes. We identified telomere sequences at both ends of 72 out of the 76 chromosomes. This enabled us to assemble the four sets of haplotype genomes, providing further evidence for the presence of intact telomeres in the genome (Figure 1C, Supplemental Figure 8, and Supplemental Table 5). However, four chromosomes—chr7b, chr7c, chr7d, and chr8d—displayed telomere sequences at only one terminus. This unusual pattern suggests potential structural variations, such as telomere loss or rearrangements. This finding emphasizes the need for further investigation of the genomic stability and integrity of these specific regions (Supplemental Figure 8 and Supplemental Table 5). Centromeres consist of tandem repeat sequences of various lengths that can vary among closely related species. Using a tool called Telomere Repeats Finder (TRF), we identified potential centromere positions based on the distribution of tandem repeat sequences. We identified candidate centromere regions in 35 of the 76 chromosomes (Figure 1C, Supplemental Figure 8, and Supplemental Table 5). Notably, some chromosomes exhibited distinctive patterns of highly enriched tandem repeats, whereas others did not show clear clusters of enriched sequences (Figure 1C). The centromere regions in *H. cordata* were relatively short, ranging from 2 to 71 kb in length. In addition, there were variations in the lengths of the repeating monomers within these regions, which ranged from 66 to 754 bp (Supplemental Table 5). For instance, the centromeres of chr18 [a, c, d] were the shortest, consisting of a repeating monomer length of 94 bp that was repeated approximately 29.5 times (Supplemental Figure 8 and Supplemental Table 5). By contrast, the centromere of chr14c had a larger monomer length of 754 bp that was repeated 45.9 times (Supplemental Figure 8 and Supplemental Table 5). These variations suggest potential differences in the organization and structure of centromeres among different chromosomes. The identification and characterization of centromere regions in *H. cordata* provides valuable insights into the genome dynamics and stability of this plant species.

To assess the completeness of the genome assembly, we examined the presence of ribosomal RNA (rRNA) sequences, specifically the 18S and 5S rRNA genes. These sequences are highly repetitive and are useful indicators of genome quality. We identified the distribution of the 18S rRNA gene primarily at the ends of chr06 [a–d] and chr07 [a–d], confirming the presence of these rRNA sequences in the assembled genome (Supplemental Figure 9A). The 5S rRNA gene arrays were predominantly located on chr09 [a–d] (Supplemental Figure 9B). The presence of these rRNA sequences further supports the completeness and reliability of the assembled genome.

Evolution and whole-genome duplication of the *H. cordata* genome

The genome of *H. cordata* provides a valuable resource for studying the evolution of the magnoliid branch (Soltis et al., 2009). To examine the evolutionary position of *H. cordata*, we performed a comparative analysis of 21 angiosperm species, including representatives from Magnoliaceae, eudicots, monocots, and basal angiosperms such as *Amborella trichopoda* (Amborella Genome Project, 2013) and *Nymphaea colorata* (Dong et al., 2018). We identified a total of 19 455 gene families, 25 of which contained 86 genes specific to *H. cordata* (Supplemental Table 6). To establish phylogenetic relationships and estimate divergence times, we selected 150 single-copy orthologous gene families to construct a phylogenetic tree (Figure 2A). *H. cordata* clustered together with other magnoliid species, including *Aristolochia fimbriata*, *Aristolochia contorta*, *Piper nigrum*, and *Saururus chinensis*. The divergence between *H. cordata* and *Aristolochia* species occurred approximately 106.8 million years ago (MYA), whereas the divergence between *H. cordata* and *S. chinensis* took place approximately 33.7 MYA (Figure 2A and Supplemental Table 7).

We next analyzed gene family expansion and contraction across the selected species. We found that 74.0% of the gene families (4974) had experienced expansion in *H. cordata*, and 26.0% (1752) of the gene families had undergone contraction (Supplemental Table 8). The expanded gene families were enriched in GO terms associated with transcriptional regulation and the cell cycle (Supplemental Table 9), indicating their involvement in gene expression and cell division. In the KEGG pathway enrichment analysis, the expanded gene families were primarily associated with plant hormone signal transduction and thiamine, cysteine, and methionine metabolism (Supplemental Table 10). Among the expanded gene families, particular emphasis should be placed on those rapidly expanded gene families that clustered in the GO enrichment analysis categories of DNA-binding transcription factor activity, secondary metabolic processes, defense response to other organisms, and

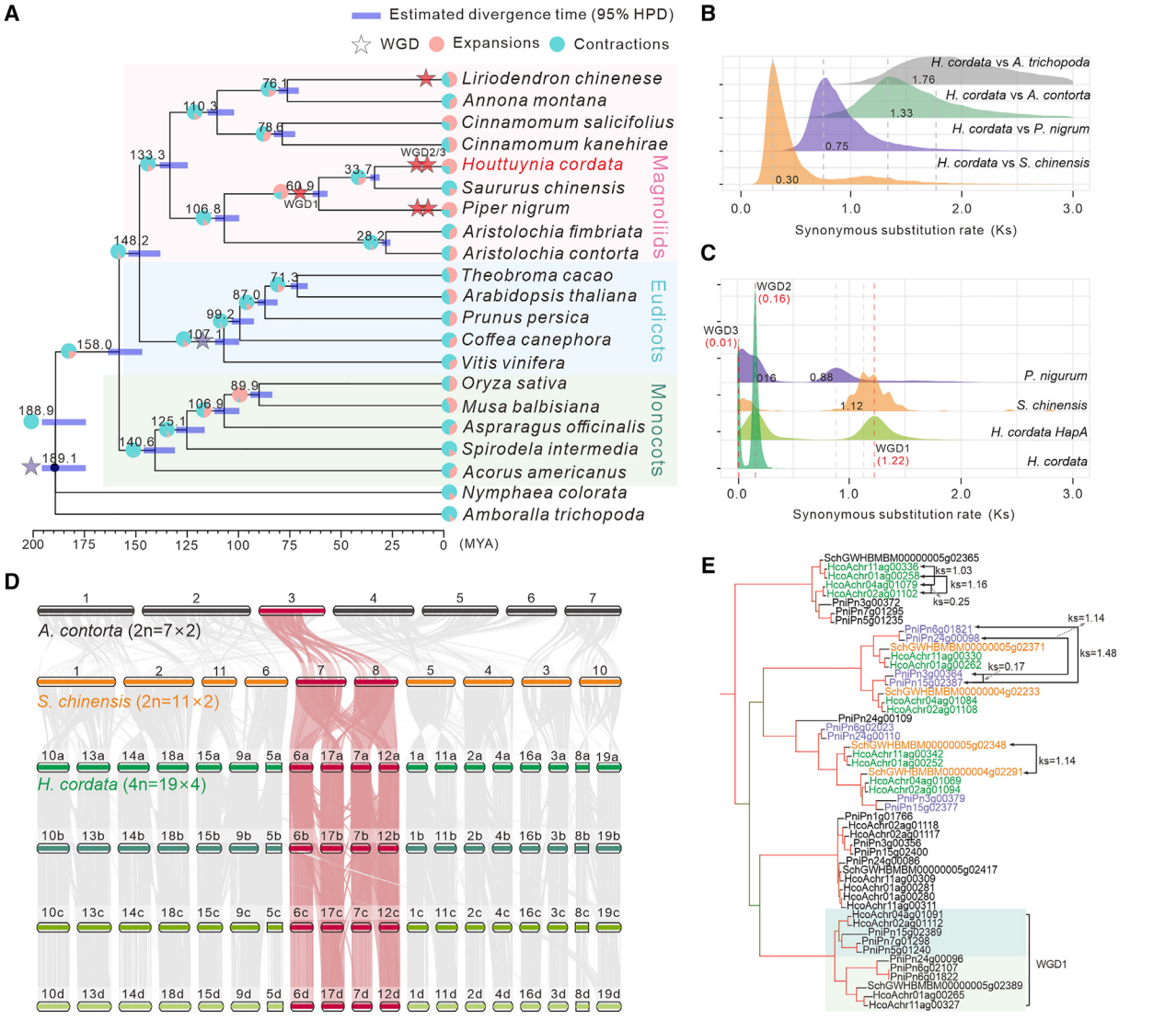

Figure 2. Phylogenetic analysis and whole-genome duplications of the *H. cordata* genome.

(A) Phylogenetic tree depicting the relationships among *H. cordata* and 20 other angiosperm species. The pie charts show the proportions of gene families with expansions (pink) and contractions (light blue) at each node or species. The estimated divergence time (in million years ago, MYA) is indicated at each node, with bars representing 95% confidence. Nodes that have undergone whole-genome duplications (WGDs) are marked with a star. Purple stars indicate well-known ancient WGD events, and red stars represent WGD events specific to certain species.

(B and C) Synonymous substitution rate (K_s) distributions. The K_s distributions of syntenic blocks for orthologous genes **(B)** and paralogous genes **(C)** are shown for *H. cordata*, haplotype A (HapA) of *H. cordata*, *Amborella trichopoda*, *Aristolochia contorta*, *Piper nigrum*, and *Schisandra chinensis*.

(D) Collinear relationships among *A. contorta*, *S. chinensis*, and *H. cordata* chromosomes. The gray lines in the background indicate syntenic blocks between the genomes that span more than 15 genes, with some representative blocks highlighted in red.

(E) Phylogenetic tree of homologous genes from *S. chinensis*, *P. nigrum*, and *H. cordata*. The genes from *H. cordata*, *S. chinensis*, and *P. nigrum* are highlighted in green, orange, and purple, respectively. The branch color indicates the bootstrap support value, decreasing from red to green.

UDP-glucosyltransferase activity (Supplemental Table 11). In the KEGG enrichment analysis, these rapidly expanded gene families were involved in phenylpropanoid and tryptophan metabolism, glucosinolate biosynthesis, flavonoid biosynthesis, and diterpenoid biosynthesis (Supplemental Table 12).

WGDs and the amplification of repetitive sequences are recognized as key drivers of species evolution. To explore the potential occurrence of WGDs in *H. cordata*, we analyzed the synonymous substitution rates (K_s) of orthologous gene pairs between *H. cordata* and four other species. The K_s peak distributions between *H. cordata* and *A. trichopoda*, between *H. cordata* and *A. fimbriata*, and between *H. cordata* and *S. chinensis* generally aligned with their predicted evolutionary relationships (Figure 2B). We identified WGDs in *P. nigrum*, *S. chinensis*, and *H. cordata* by analyzing the density distribution of K_s of paralogous gene pairs (Figure 3C). In the *H. cordata* genome, we detected three distinct K_s peaks of 1.22, 0.16, and 0.01 (Figure 2C), indicating the occurrence of three WGDs in *H. cordata*. The most recent WGD, WGD3, as a polyploidization event, resulted in the formation of a tetraploid species of *H. cordata* from a diploid species (Figure 2D). However, the collinear gene pairs between

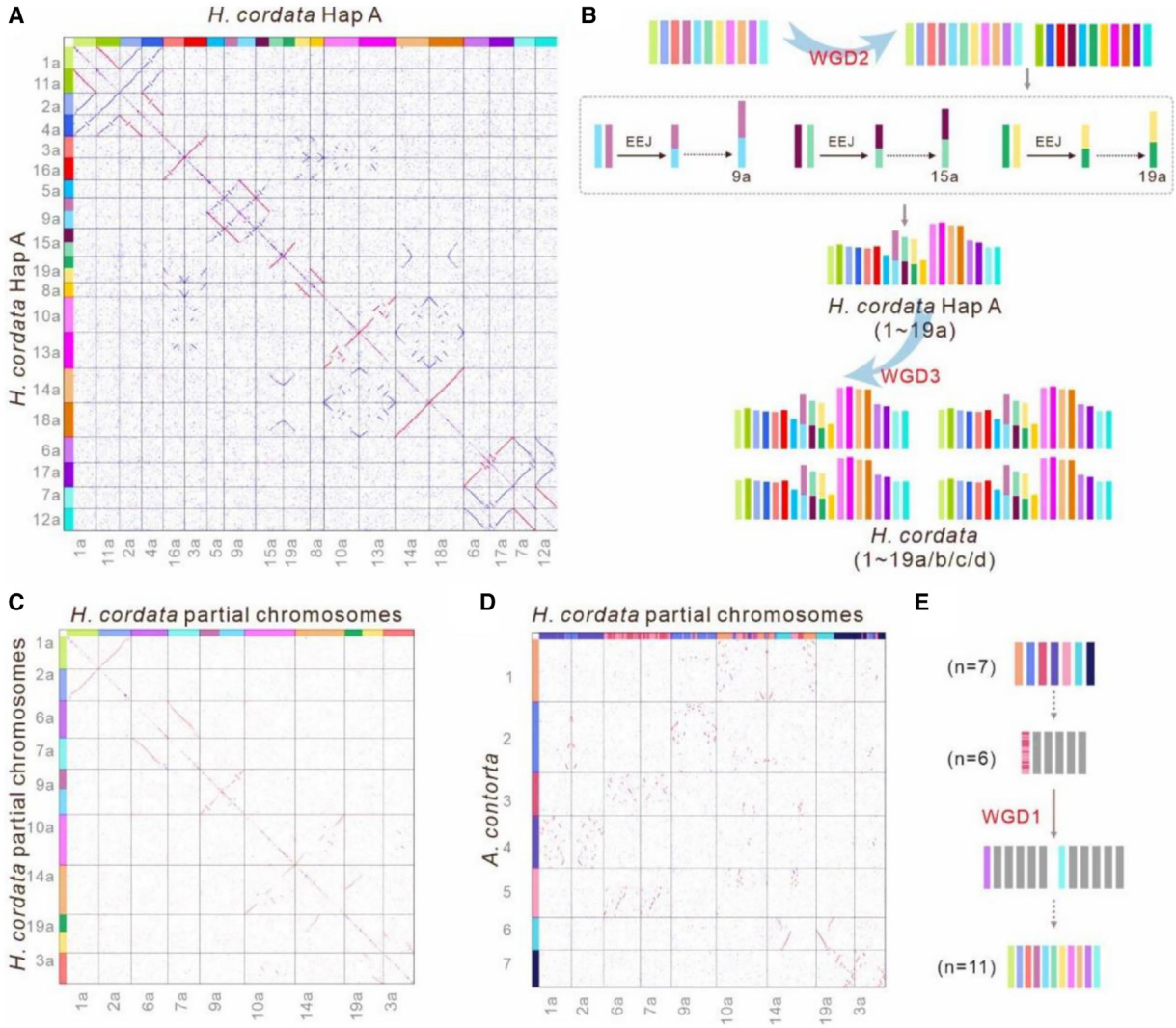

Figure 3. Inference of karyotype evolution in *H. cordata*.

(A) The dot plot displays collinear blocks within haplotype A (HapA) of *H. cordata*. Chromosomes have been sorted and colored on the basis of homology, with pairs of similar colors indicating highly homologous chromosomes resulting from whole-genome duplication (WGD).

(B) Speculative processes of chromosome duplications and three end-to-end joining (EEJ) events following WGD2 and WGD3.

(C) Dot plot of collinear blocks among nine chromosomes in *H. cordata*, representing the ancestral karyotype prior to WGD2. Chromosome colors correspond to those in **(A)** and **(B)**.

(D) Dot plot of collinear blocks among nine chromosomes of *H. cordata* and *A. contorta*. Chromosomes are color-coded according to *A. contorta*.

(E) Speculative models for chromosomal changes in the WGD1 event.

haplotype A of *H. cordata* (HcoHapA) and *A. contorta*, which did not experience independent WGD events, exhibited a 4:1 ratio (Supplemental Figure 10A and 10B), reflecting the doubling of WGD1 and WGD2. The detected K_s peak was approximately 1.12 in *S. chinensis*, and a notable number of gene pairs exhibited K_s values concentrated below 0.25, suggesting that *S. chinensis* has undergone only one WGD, consistent with previous findings (Xue et al., 2023) (Figure 2C). The syntenic depth between *S. chinensis* and HcoHapA showed a 2:4 pattern (Supplemental Figure 11A and 11B), confirming an additional WGD event in *H. cordata*. By contrast, two distinct peaks, measuring around 0.88 and 0.16, were observed in the *P. nigrum* genome (Figure 2C). In the collinearity analysis between HcoHapA and *P. nigrum*, the syntenic depths showed a 4:2 pattern. However, some of the HcoHapA blocks exhibited alignments with five to eight corresponding *P. nigrum* blocks (Supplemental Figure 12A and 12B); for example, chr01a, chr02a, chr04a, and chr11a aligned to Pn1, Pn3, Pn5, Pn6, Pn7, Pn13, Pn15, and Pn24 (Supplemental Figure 12A and 12B). This indicates that *P. nigrum* has undergone at least two WGDs. However, relying solely on evaluation of the density distribution of K_s, it is challenging to determine whether *P. nigrum*, *S. chinensis*, and *H. cordata* have experienced a common WGD. Alternatively, these data suggest a shared WGD1 event between *H. cordata* and *P. nigrum*. Our analysis of paralogous genes further supports this observation, as we detected paralogous genes in *P. nigrum* that originated from WGD1 (block $K_s \approx 1.0$–2.0; Supplemental Figures 13–15). On the basis of genes within the identified blocks and orthologous genes of *H. cordata*, we deduced the evolutionary relationships (Figure 2D, Supplemental Figure 16,

and Supplemental Table 13). Notably, orthologous gene pairs tended to cluster together, rather than paralogous gene pairs (Figure 2A and 2E), further supporting the occurrence of WGD1 in the ancestor of *H. cordata* and *P. nigrum*.

Inference of karyotype evolution in the *H. cordata* genome

Advances in genome sequencing in multiple organisms have opened up new possibilities for understanding the evolution of ancestral karyotypes (Murat et al., 2017). After multiple WGD events, the *H. cordata* genome still exhibits traces of duplication, suggesting the preservation of ancient genomic information. By analyzing conserved blocks in HcoHapA, we identified duplications and inferred chromosome rearrangement events following WGD2 (Figure 3A). Notably, Chr1a and Chr11a displayed high homology, indicating their duplication through WGD2. Three evident chromosome rearrangement events occurred in *H. cordata* after WGD2. Prominent homology between the lower half of Chr9a and Chr5a, as well as the upper half of Chr9a and Chr15a, suggested end-to-end joining (EEJ) following duplications from two ancestral chromosomes (Figure 3A and 3B). Similarly, Chr19a exhibited high homology with the lower half of Chr15a and Chr8a, indicating its gradual evolution through EEJ from two ancestral chromosomes (Figure 3A and 3B). On the basis of this inference, the chromosome count initially started at $n = 11$. It then doubled through WGD2, resulting in a count of $n = 22$. Subsequently, three EEJ events occurred, leading to a final count of $n = 19$. Thereafter, the *H. cordata* species underwent WGD3 (polyploidization), leading to the formation of a tetraploid genome. ($4n = 19 \times 4$) (Figure 3B).

To detect duplications caused by WGD1, we selected nine chromosomes (Chr1a/2a/3a/6a/7a/9a/10a/14a/19a) to avoid interference from duplications that resulted from WGD2. Collinearity analysis revealed clear duplications between Chr1a and Chr2a as well as between Chr6a and Chr7a (Figure 3C). In addition, Chr9a was formed through EEJ of an ancestral chromosome following duplication (Figure 3B). Incomplete reciprocal homologous fragments were also detected among Chr10a, Chr14a, Chr19a, and Chr3a (Figure 3C), indicating subsequent changes to WGD1. Insertion of a fragment from Chr19a into the middle of Chr14a (Figure 3C) resulted in chromosomal reduction, and fragment exchanges between Chr10a and Chr3a were observed. On the basis of these findings, the chromosome count of the common ancestor species of *H. cordata* and *P. nigrum* was estimated to be 6 before WGD1.

A. contorta ($n = 7$), an outgroup species of *H. cordata* and *P. nigrum*, did not undergo independent WGDs. This suggests that chromosome fusions occurred in the ancestral species, resulting in a change in chromosome number from 7 to 6. Collinearity analysis between *H. cordata* and *A. contorta* revealed that the collinearity of almost all chromosomes was not well maintained, likely owing to the distant evolutionary relationship between *A. contorta* and *H. cordata* (Figure 3D). However, it is evident that Chr6a and Chr7a in *H. cordata* evolved through the fusion of Ac3 and Ac5 in *A. contorta* (Figure 3D). In addition, the fusion of Ac3 and Ac5 was also detected in the collinearity between *A. contorta* and *P. nigrum* (Supplemental Figure 17A and 17B). Overall, a fusion event occurred in the ancestral species of *H. cordata* and *P. nigrum*, resulting in a shift from 7 to 6 chromosomes (Figure 3E). Following WGD1 in the ancestral species ($n = 6$), a replica of a Chr19a fragment was inserted into Chr14a (Figure 3C), resulting in a species with 11 chromosomes (Figure 3E).

Sodium houttuyfonate biosynthesis in *H. cordata*

Houttuynin, also known as decanoyl acetaldehyde, is an important compound found in *H. cordata* that is responsible for its distinct fishy odor (Laldinsangi, 2022). The water-soluble derivative, SH, is highly stable and exhibits important medicinal properties (Liu et al., 2021b). As a result, it is widely used in various medical applications. An initial ultra-high performance liquid chromatography with quadrupole time-of-flight mass spectrometry (UPLC-Q-TOF-MS) analysis successfully detected the presence of SH in both the rhizome (Supplemental Figure 18) and leaves (Supplemental Figure 19) of *H. cordata*. This was determined based on retention time (11.89) and mass spectral data, which included a parent ion at *m/z* 302.30493 and fragment ions at *m/z* 295.28714, 297.28287, and so forth. These results were consistent with those of the SH standard sample (Supplemental Figure 20). A noteworthy finding was a significantly higher ($p < 0.05$) concentration of SH in rhizomes (79.59 ppm) compared with leaves (2.40 ppm) (Figure 4A). This disparity aligns with the stronger fishy odor evident in the rhizomes, suggesting a potential biosynthetic pathway for SH within the plant.

Houttuynin differs from typical secondary metabolites such as terpenes with isoprene units, flavonoids with phenolic structures, or alkaloids with nitrogen atoms. Its unique characteristics, including the presence of an aldehyde group and decanoyl group, suggest that it is derived from fatty acid oxidation. Therefore, we propose that houttuynin is produced through the β-oxidation pathway of fatty acid metabolism. In this pathway, fatty acids abundant in *H. cordata*, such as palmitic acid (C16), are first converted into palmitoyl-coenzyme A (CoA) through acyl-CoA synthesis. This compound then undergoes sequential removal of two-carbon units through β-oxidation. After three cycles of β-oxidation, decanoyl-CoA (C10-CoA) is formed. We hypothesize that thioesterase enzyme converts decanoyl-CoA into decanoic acid ($C_{10}H_{20}O_2$), which is further transformed into decanal ($C_{10}H_{20}O$). Subsequently, aldehyde dehydrogenase likely catalyzes the conversion to decanoyl acetaldehyde (houttuynin, $C_{12}H_{22}O_2$), ultimately leading to the synthesis of SH ($C_{12}H_{23}NaO_4S$). Houttuynin is prone to instability and can readily transform into another compound known as 2-undecanone (methyl nonyl ketone, $C_{11}H_{22}O$) during production (Chen et al., 2014) (Figure 4B). Our hypothesis is supported by the identification of decanoic acid and SH in this study and palmitic acid, decanal, and 2-undecanone in previous research through UPLC-Q-TOF-MS analysis (Persson et al., 2010).

To gain further understanding of the differences in houttuynin synthesis between rhizome and leaf tissues, we performed transcriptome profiling of enzyme-coding candidate genes involved in the fatty acid biosynthetic pathway in *H. cordata*. By comparing rhizome and leaf tissues (Rz vs. Lf), we identified 49 differentially expressed genes (DEGs) that may play a role in this process (Figure 4C). These genes can be classified into seven categories based on their functions within the fatty acid pathway: fatty

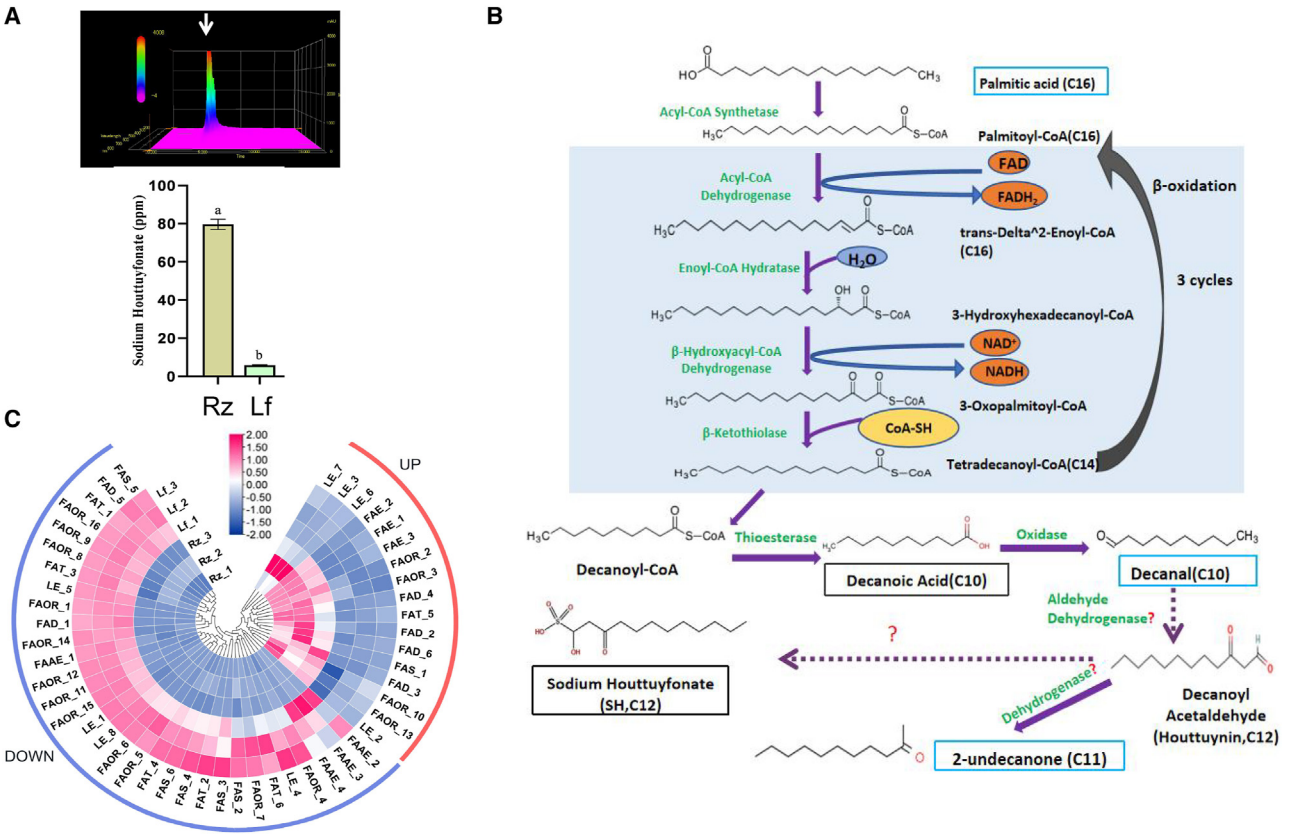

Figure 4. Proposed sodium houttuyfonate biosynthetic pathway in *H. cordata*.

(A) Analysis of sodium houttuyfonate (SH) in different tissues using high-performance liquid chromatography. The retention time of the peak associated with SH is indicated by a white arrow. The accompanying bar graph provides a quantitative representation of the SH content in various tissues. The letters a and b indicate significant differences between groups based on ANOVA analysis, followed by Tukey's HSD test for post hoc analysis.

(B) Proposed biosynthetic pathway of SH. Houttuynin originates from fatty acid metabolism, specifically palmitic acid ($C_{16}H_{32}O_2$, molecular weight [MW] 256.42). Palmitic acid is converted to palmitoyl-CoA (C16) by acetyl-CoA synthetase, followed by mitochondrial fatty acid β-oxidation. This cycle continues until decanoyl-CoA is formed, which is then converted to decanoic acid ($C_{10}H_{20}O_2$, MW 172.26) by thioesterase. Decanoic acid is further oxidized to decanoylacetaldehyde (houttuynin, $C_{12}H_{22}O_2$, MW 198.16). Subsequently, the aldehyde group is transformed into a hydroxyl group, reacting with sodium sulfate to form water-soluble sodium houttuyfonate (SH, $C_{12}H_{23}NaO_5S$, MW 302.36). In addition, houttuynin is prone to instability and can readily transform into 2-undecanone (methyl nonyl ketone, $C_{11}H_{22}O$) during production. Dashed lines represent potential intermediate steps, and question marks indicate uncertain enzyme involvement. Metabolites identified by UPLC-Q-TOF-MS in this study are marked with black boxes, and those reported in previous studies are enclosed in cyan boxes.

(C) Clustered heatmap of differentially expressed genes (DEGs) in rhizome vs. leaf tissues (Rz vs. Lf) of *H. cordata*. Values from three biological replicates are shown. The DEGs are categorized into seven functional classes based on their roles in the fatty acid metabolic pathway. These classes include fatty acid synthases (FAS), fatty acid activation enzymes (FAAE), fatty acid oxidoreductases (FAOR), fatty acid transferases (FAT), fatty acid esterases (FAE), lipid enzymes (LE), and fatty acid dehydrogenases (FAD). Gene-expression differences are quantified as $\log_2$ fold changes, with statistical significance set at $p < 0.001$. "Up" indicates a $\log_2$ fold change > 1 (upregulation in rhizome relative to leaf), and "Down" indicates a $\log_2$ fold change < −1 (downregulation in rhizome relative to leaf).

acid synthases, fatty acid activation enzymes, fatty acid oxidoreductases, fatty acid transferases, fatty acid esterases, lipid enzymes, and fatty acid dehydrogenases. Seventeen of these genes were upregulated, and 32 were downregulated (Supplemental Table 14 and 15). This comparative analysis sheds light on the possible enzymatic processes involved in houttuynin biosynthesis, providing valuable insights into the unique fatty acid pathway for SH accumulation in *H. cordata*.

Flavonoid biosynthesis and quercetin metabolism in *H. cordata*

H. cordata is widely recognized for its remarkable flavonoid production and diverse array of beneficial properties (Ekalu and Habila, 2020). This makes it an ideal subject for studying the metabolic pathways involved in flavonoid biosynthesis, which can be greatly influenced by the unique characteristics of the source organism. We carried out a thorough analysis of total flavonoids in various tissues of *H. cordata*, including rhizomes (Rz), leaves (Lf), bracts (Br), and the inflorescence (If). We found that rhizomes had the lowest total flavonoid content (2.38 mg/g), whereas leaves (5.80 mg/g), the inflorescence (5.59 mg/g), and bracts (9.49 mg/g) exhibited higher levels (Figure 5A). To further investigate the distribution of flavonoids within *H. cordata*, we used a comprehensive targeted metabolomics approach, which identified 212 enriched flavonoid metabolites (Figure 5B). Among these compounds, flavonols and flavones were the most prominent, making up 32% and 31% of the total identified

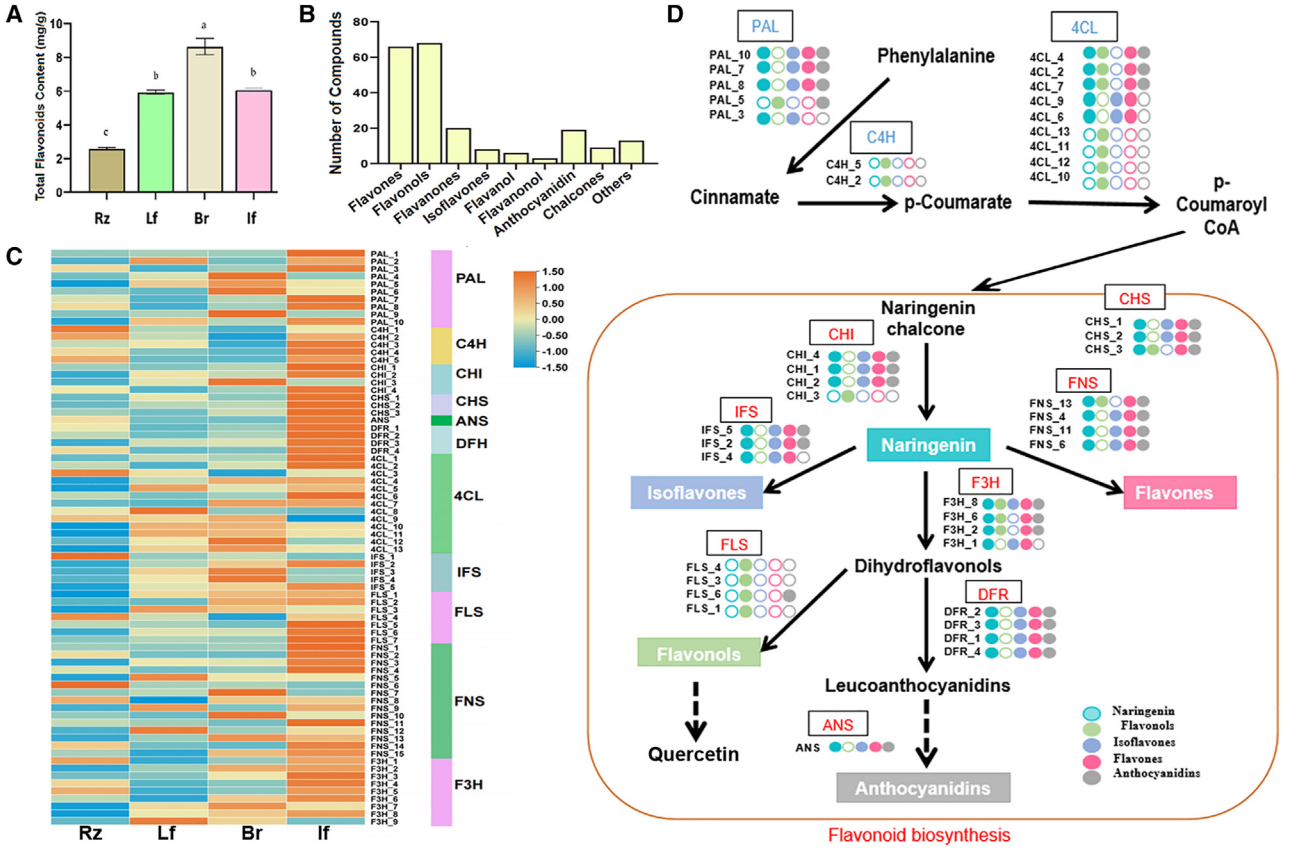

Figure 5. Expression analysis of enzyme-coding genes involved in flavonoid biosynthesis in *H. cordata*.

(A) Analysis of total flavonoid content in different tissues. Lf, leaves; Rz, rhizomes; Br, bracts; If, inflorescences. The letters represent significant differences between groups determined by ANOVA, followed by Tukey's HSD test for post hoc analysis.

(B) Number of flavonoid metabolites in *H. cordata*.

(C) Heatmap illustrating the expression levels of enzyme-coding genes involved in flavonoid biosynthesis across tissues of *H. cordata*.

(D) Simplified flavonoid biosynthetic pathway in *H. cordata*. Colored circles represent specific metabolites: cyan for naringenin, pink for flavones, blue for isoflavone, green for flavonol, and gray for anthocyanidins.

flavonoids, respectively (Figure 5B). This accumulation pattern highlights the diverse range of flavonoids present in *H. cordata* and emphasizes the notable presence of flavonols and flavones in this medicinal plant.

To gain insight into flavonoid biosynthesis in *H. cordata*, we performed transcriptomic profiling of various tissues. This comprehensive analysis revealed a set of 76 conserved enzyme-coding genes that are closely associated with the flavonoid biosynthetic pathway. These identified genes play a crucial role in orchestrating critical steps in the pathway and encode key enzymes such as ammonia-lyase (PAL), 4-coumarate CoA ligase (4CL), cinnamate 4-hydroxylase (C4H), chalcone isomerase (CHI), CHS, flavonol synthase (FLS), flavanone 3-hydroxylase (F3H), anthocyanidin synthase (ANS), dihydroflavonol 4-reductase (DFR), isoflavone synthase (IFS), and flavone synthase (FNS) (Figure 5C and Supplemental Table 16). These findings provide a comprehensive understanding of the genetic basis underlying flavonoid biosynthesis in *H. cordata*.

To further enhance our candidate selection process, we performed a comprehensive analysis of the relationships among gene-expression levels and the top five flavonoid metabolites (Figure 5D). These metabolites included flavanone (naringenin, cyan), quercitrin (flavonols, green), luteolin (flavones, pink), delphinidin-3-*O*-rutinoside (anthocyanidins, gray), and glycitein (isoflavones, blue). Our findings revealed significant correlations ($|r| \geq 0.6$, $p < 0.05$) between 44 out of the 76 enzyme-coding candidate genes and at least one of the tested metabolites (Figure 5D). For example, three PAL genes (PAL_10/7/8) were significantly correlated with all flavonoid types except for flavonols. PAL_5 showed notable correlations with flavonols (green) and anthocyanidins (gray), whereas PAL_3 exhibited significant correlations with naringenin (cyan) and isoflavones (blue). Overall, our analysis identified key genes, including four PAL (PAL_10/7/8/3), five 4CL (4CL_4/2/7/9/6), three CHS (CHS_1/2/3), and three CHI (CHI_4/1/2) genes that were all significantly correlated with naringenin (cyan), highlighting their crucial roles in naringenin biosynthesis (Figure 5D). We also discovered a distinct gene set associated with flavonol biosynthesis (green), comprising one PAL (PAL_5), seven 4CL (4CL_4/2/7/13/11/12/10), one CHI (CHI_3), one CHS (CHS_3), one FNS (FNS_13), three F3H (F3H_8/6/2/1), and four FLS (FLS_4/3/6/1) genes. We observed a specific correlation with isoflavone (blue) biosynthesis involving four PAL (PAL_10/7/8/3),

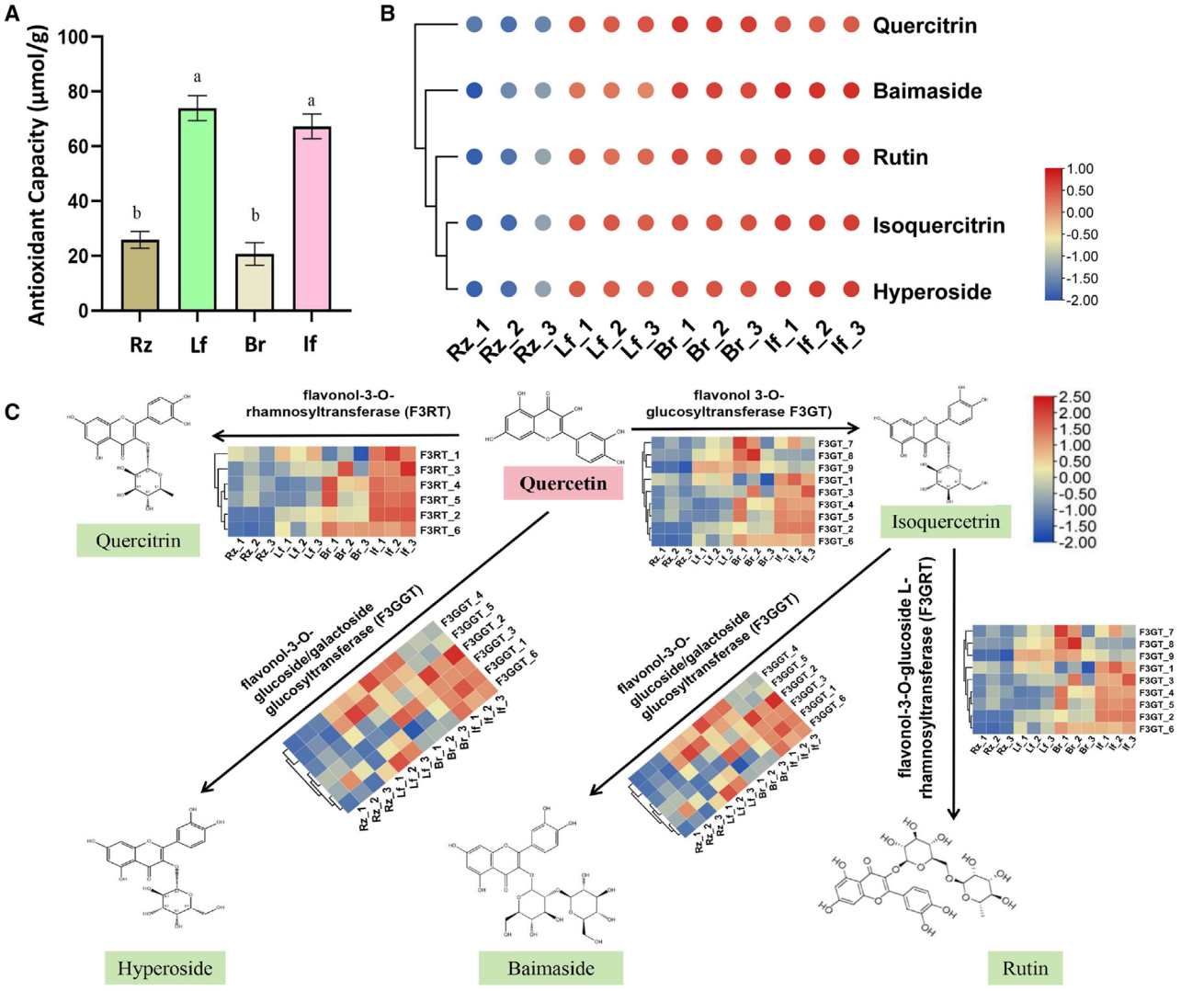

Figure 6. Quercetin metabolic pathway in *H. cordata*.

(A) Total antioxidant capacity in various tissues of *H. cordata*. Lf, leaves; Rz, rhizomes; Br, bracts; If, inflorescences. The letters a and b represent significant differences between groups determined by ANOVA, followed by Tukey's HSD test for post hoc analysis.

(B) Abundance of quercetin glycosides in different *H. cordata* tissues.

(C) Overview of the quercetin metabolic pathway. Heatmaps accompanying the text display the expression level (TPM) of enzyme-coding genes involved in this pathway. Values from three biological replicates are shown.

five 4CL (4CL_9/6), three CHS (CHS_1/2), three CHI (CHI_4/1/2), three FNS (FNS_4/11/16), three IFS (IFS_5/2), two F3H (F3H_8/1), four DFR (DFR_2/3/1/4), and one ANS gene. Similarly, we found a significant correlation with flavone (pink) biosynthesis that involved three PAL (PAL_10/7/8), five 4CL (4CL_4/2/7/9/6), three CHS (CHS_1/2), three CHI (CHI_4/1/2), four FNS (FNS_13/4/11/16), three IFS (IFS_5/2/4), four F3H (F3H_8/6/2/1), four DFR (DFR_2/3/1/4), and one ANS gene. Finally, we observed a significant correlation with anthocyanidins (gray) for four PAL (PAL_10/7/8/5), three 4CL (4CL_4/2/7), three CHS (CHS_1/2), three CHI (CHI_4/1/2), four FNS (FNS_13/4/11/16), two IFS (IFS_5/2), three F3H (F3H_8/6/2), four DFR (DFR_2/3/1/4), and one ANS gene (Figure 5D and Supplemental Table 16). These findings highlight specific gene-expression networks involved in flavonoid biosynthesis within *H. cordata*.

Quercetin, one of the most well-known flavonoids, is known for its powerful antioxidant properties and other health benefits in traditional Chinese medicine (Wang et al., 2016). Our study revealed significant variations ($p < 0.05$) in antioxidant capacity among different tissues of *H. cordata*. Notably, leaves (73.9 μmol/g) and inflorescences (67.3 μmol/g) exhibited considerably higher ($p < 0.05$) antioxidant capacity compared with rhizomes (25 μmol/g) and bracts (20 μmol/g) (Figure 6A). Furthermore, glycosylated forms of quercetin, such as isoquercitrin, rutin, hyperoside, baimaside, and quercitrin, which play a crucial role in *in vivo* antioxidant activity, were markedly enriched ($p < 0.05$) in leaves, inflorescences, and bracts. These tissues displayed high concentrations (indicated by red color) of glycosylated quercetin forms in the UPLC-Q-TOF-MS metabolomics analysis, whereas rhizomes showed

the lowest concentrations (indicated by blue color) (Figure 6B). This distribution pattern (with the exception of bracts) is consistent with the observed antioxidant capacities, highlighting the distinct roles of different plant parts in accumulating quercetin glycosides.

Further analysis focused on the expression levels of homologous glucosyltransferase (UGT) enzyme genes (F3RT, F3GT, F3GGT, and F3GRT) that catalyze the formation of these five quercetin glycosides in *H. cordata*. The expression trends of these UGT genes, i.e., three F3RT (F3RT_2/3/6), five F3GT (F3GT_7/9/3/2/6), four F3GGT (F3GGT_4/5/2/6), and three F3GRT (F3GRT_5/1/7) genes, mirrored the observed patterns of quercetin metabolite accumulation, with lower expression in the rhizome than in the other organs (Figure 6C). This finding suggests a notable relationship between gene expression and quercetin metabolism, providing valuable insight into the regulation of quercetin biosynthesis and its distribution across various plant tissues.

Gene co-expression analysis of flavonoid biosynthesis genes in *H. cordata*

To gain a deeper understanding of the regulatory mechanisms that control flavonoid biosynthesis in *H. cordata*, we used WGCNA (weighted gene co-expression network analysis), a powerful computational tool that helped to reveal gene relationships and regulatory patterns within our dataset. Through this analysis, we organized all genes into 32 modules, distinguished by different colors such as MEblack, MEblue, and MEbrown, based on their similar expression patterns across samples (Figure 7A and Supplemental Table 17). Among these modules, the MEblue module stood out, exhibiting notably strong and statistically significant positive correlations ($r > 0.8$, $p < 0.01$) with flavones, isoflavones, anthocyanidins, and naringenin (excluding flavonols, Supplemental Table 17). This finding suggests that the genes in the MEblue module play a crucial role in biosynthesis of these specific flavonoid compounds.

Delving further into the MEblue module, we focused on predominantly significant correlations ($r > 0.91$, $p < 1e{-}200$) which had an extremely low probability of occurring by random chance, leading us to identify seven enzyme-coding candidate genes (Supplemental Figure 21). These included one 4CL (Hocor15aG0005700), one FNS (Hocor16aG0024200), two DFRs (Hocor12aG0138000, Hocor01aG0150600), one ANS (Hocor14aG0093300), and two CHIs (Hocor10aG0148000, Hocor16aG0092100), which shared similar expression patterns across samples (Figure 7B). We performed RT–qPCR to validate the expression of these seven candidate genes identified from the MEblue module. They were notably highly expressed in inflorescences, corroborating the expression patterns observed in the heatmaps (Supplemental Figure 22). The strong positive correlations allowed us to perform an extensive analysis of species-specific networks within the blue module. This analysis led to the identification of key transcription factors (TFs) involved in regulation of the seven candidate genes associated with flavonoid biosynthesis (Figure 7B). These 25 TFs belonged to 12 families, including well-known families such as bZIP, MYB, ARF, AP2, and GRAS, which have been extensively studied for their critical roles in secondary metabolism (Yao et al., 2018; Cao et al., 2020; Li et al., 2020; Han et al., 2023) (Supplemental Table 18).

To further support the hypothesis of co-regulation, we performed TF binding site predictions, revealing the presence of known *cis* elements in the upstream regions of the candidate genes. Specifically, we identified G-box binding sites for bZIP TFs, MYB binding sites for MYB TFs, TGA elements and the AUXRR core for ARF TFs, HD-ZIP binding sites for HD-ZIP TFs, and the W-box for WRKY TFs (Figure 7C and Supplemental Table 19). These findings provide compelling evidence for the potential hierarchical regulatory relationships between the TFs and the enzyme-coding genes involved in flavonoid biosynthesis.

Importantly, our analysis also revealed higher expression levels of these TFs in inflorescences, consistent with the expression patterns of the enzyme-coding genes and the contents of flavonoid metabolites (Figure 5A and 5C). We performed subcellular localization (Figure 7E) and RT–qPCR assays (Figure 7F) for two of the TFs, MYB_2 and AP2_2, confirming their nuclear localization and significant expression ($p < 0.05$) in inflorescences. These observations suggests that MYB_2 and AP2_2 may have specific roles in the regulation of flavonoid biosynthesis during inflorescence development. Their higher expression levels in inflorescences emphasize the importance of these TFs for shaping the flavonoid profiles of *H. cordata* (Figure 7D and Supplemental Table 20).

Together, our analyses identify co-expressed genes involved in flavonoid biosynthesis in *H. cordata* and highlight the potential applications of TF-mediated regulation of flavonoid biosynthesis in *H. cordata* in future studies.

DISCUSSION

The Saururaceae family consists of four genera: *Anemopsis*, *Gymnotheca*, *Houttuynia*, and *Saururus*. Members of this family typically have heart-shaped leaves and spike-like flowers (Smith and Stockey, 2007). Among these genera, only the genome of *S. chinensis* has been reported previously. This genome is 537 Mb and diploid with 11 chromosome pairs ($2n = 22$) (Xue et al., 2023). *H. cordata*, a species from the Surrogacies family of the magnoliid group, has been a topic of debate regarding its evolutionary position within the magnoliid branch and the relationships among its species (Datta and Dasgupta 1977; Soltis et al., 2009). In the present study, we expanded our sequencing efforts to include *H. cordata* and investigated the evolutionary relationships within the Saururaceae family. In addition, we identified candidate genes involved in the production of bioactive compounds in *H. cordata*.

To overcome the challenges posed by the large genome size and high heterozygosity of *H. cordata*, we produced a chromosome-level and partially telomere-to-telomere high-precision genome assembly. We identified telomeres in 72 out of 76 chromosomes, as well as candidate centromere regions, with minimal gaps, providing insight into the organization and stability of these genomic elements. Previous studies have identified independent WGD events in species such as *Liriodendron chinense* and *P. nigrum*, whereas the *Aristolochia* species have not experienced such events (Qin et al., 2021). Our phylogenetic analysis revealed a close evolutionary relationship between *H. cordata* and other magnoliid species such as *S. chinensis*, *P. nigrum*, *A. fimbriata*, and *A. contorta*. *H. cordata* diverged from the *Aristolochia* genus

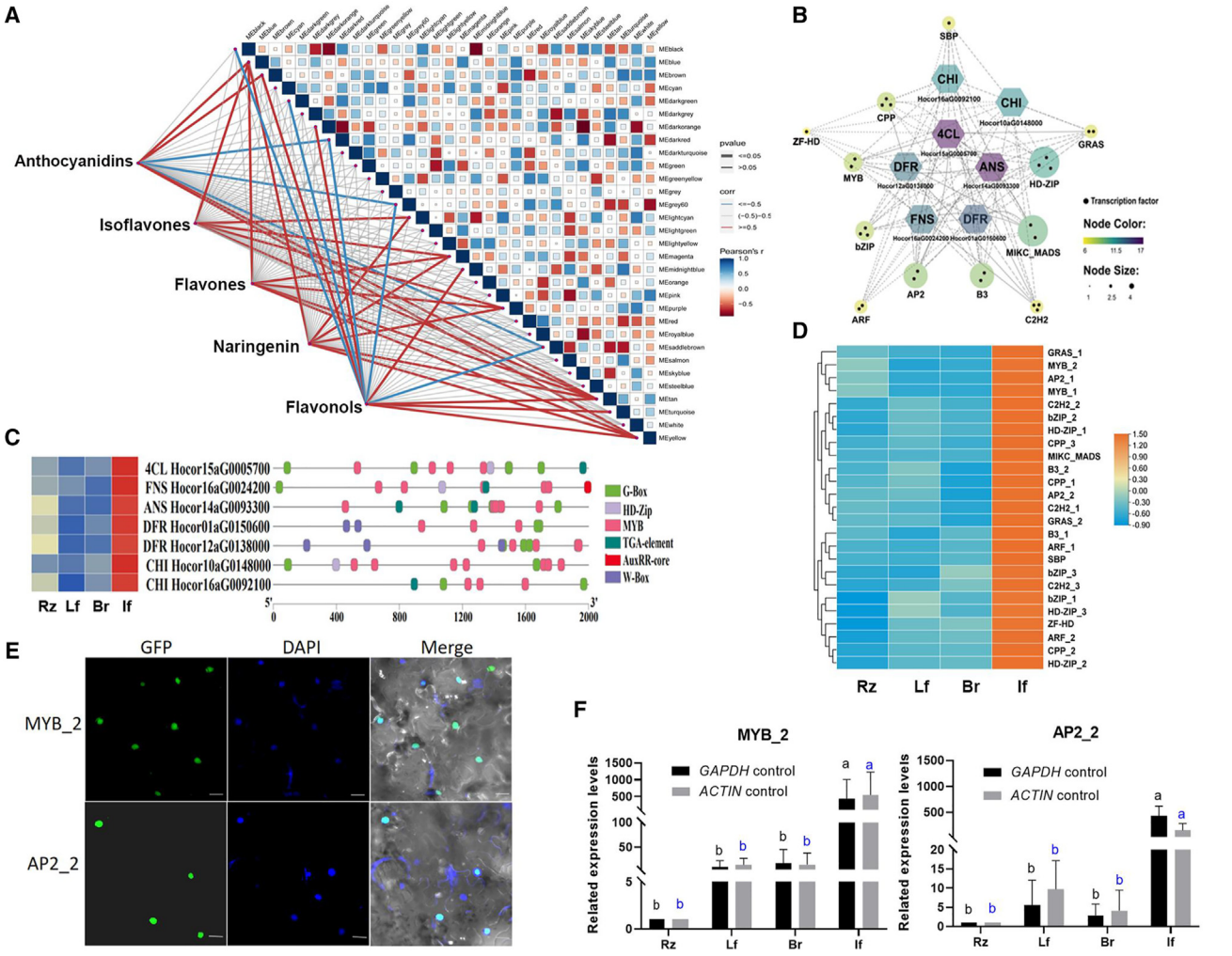

Figure 7. Gene co-expression network analysis in the flavonoid biosynthesis pathway.

(A) Weighted gene co-expression network analysis between all genes and five metabolites in *H. cordata*. The network diagram visually depicts the co-expression patterns among these genes and metabolites. The heatmap provides a comprehensive overview of metabolite profiles in different tissues of *H. cordata* during the blooming stage, with lines indicating correlations.

(B) Positive regulatory network illustrating the interactions between transcription factors (TFs) and pathway genes. Enzyme-coding genes are represented as hexagons, and TFs are depicted as circles. The lines indicate positive correlations between the TFs and enzyme-coding genes.

(C) Identification of TF binding sites in the 2-kb upstream sequences of seven candidate genes and their expression levels in various tissues.

(D) Expression levels (TPM) of 25 TFs in different tissues.

(E) Subcellular localization of two *H. cordata* TFs fused to the GFP tag and expressed in *Nicotiana tabacum* cells. Nuclei were stained with DAPI. The merged images show co-localization of GFP and DAPI signals within the nucleus, confirming the nuclear localization of the TFs. Scale bars, 20 μm.

(F) RT–qPCR analysis of two TFs in *H. cordata*. Real-time qPCR analysis of two TFs across various tissues of *H. cordata*, with expression levels normalized to those of the housekeeping genes *GAPDH* and *ACTIN*. Data are presented as mean ± SEM, and each bar represents the average expression from three biological replicates. Statistical significance was determined using ANOVA, with different letters indicating significant differences ($p < 0.05$) between tissues. Tissues examined included rhizomes, leaves, bracts, and inflorescences during the blooming period.

approximately 108.4 MYA, from *P. nigrum* 61.9 MYA, and from *S. chinensis* more recently, about 33.4 MYA. Our data support the occurrence of a WGD event in the common ancestor of *H. cordata* and *P. nigrum* following their divergence from *Aristolochia*. Subsequently, *H. cordata* and *P. nigrum* underwent species differentiation. The ancestral species of *H. cordata* experienced two additional WGD events, resulting in the formation of a homologous tetraploid. In the case of *P. nigrum*, its ancestral species underwent two independent WGD events and chromosome fusions, leading to the formation of the contemporary *P. nigrum* genome. The presence of a WGD before the divergence of *H. cordata* and *P. nigrum* suggests a shared evolutionary event between these two species. These findings align with previous observations of a single WGD event in *S. chinensis* (Xue et al., 2023), indicating a shared WGD event in the *Saururus* genome after the divergence of *Saururus* and *Liriodendron*. Chromosome duplications and structural changes resulting from WGDs have influenced their genetic makeup and have potentially contributed to their divergence and speciation. The observation of fusion events in the common ancestor of *H. cordata* and *P.*

nigrum, leading to a reduction in the chromosome basis, highlights the significance of chromosomal rearrangements during their evolutionary history. These findings provide important insights into the evolution of *H. cordata* and its species within the Saururaceae, shedding light on the genetic mechanisms that underlie their adaptation and diversification.

The prevalence of gene family expansion suggests that it played a vital role in the adaptive evolution of *H. cordata*, potentially facilitating its adaptation to specific environmental conditions or ecological niches. By contrast, the contraction of gene families implies possible gene loss or reduction in certain functional categories. This pattern suggests that rapidly expanding gene families are particularly crucial for defense mechanisms, growth, and developmental processes in *H. cordata*.

H. cordata, which emits a distinct fish-like odor when its leaves are crushed, has valuable medical applications (Luo et al., 2022). One specific component of *H. cordata*, houttuynin (decanoyl acetaldehyde, $C_{12}H_{22}O_2$), is responsible for causing "fishy burps." However, houttuynin is unstable and easily converts into 2-undecanone (methyl nonyl ketone, $C_{11}H_{22}O$) during production (Chen et al., 2014). To address this instability, SH ($C_{12}H_{23}O_5SNa$), the water-stable form of houttuynin, which has pharmaceutical activity similar to that of houttuynin, is synthesized (Chen et al., 2014). Currently, SH is obtained through a synthetic process starting from lauric acid, acetic acid, and slaked lime. However, this process has limitations in terms of environmental impact, yield, and purity (Liu et al., 2021b). Therefore, it is important to understand the *in vivo* biosynthetic pathway of SH in *H. cordata*. In this study, high accumulation of SH was found in the rhizome of *H. cordata*, and a biosynthetic pathway for SH production in *H. cordata* was proposed. This pathway is related to the fatty acid oxidation pathway. Unlike other plant secondary metabolites, houttuynin is structurally similar to primary metabolic products derived from the fatty acid oxidation pathway. It is likely formed through the breakdown of long-chain fatty acids into acetyl-CoA units. Finally, houttuynin is converted to SH, possibly by aldehyde dehydrogenase. The *H. cordata* genome provides an opportunity to further study fatty acid metabolism and better understand the *in vivo* production of SH, as well as providing support for *in vitro* synthetic biology research.

H. cordata is well known for its exceptional accumulation of flavonoids, beneficial compounds found in many traditional Chinese medicinal plants (Bai et al., 2019; Zhu et al., 2021). However, the biosynthetic pathways of these compounds in *H. cordata* are not fully understood, limiting advances in its molecular breeding. By performing transcriptomic and metabolic profiling analysis, we have identified 76 conserved enzyme-coding genes that play a crucial role in flavonoid biosynthesis in *H. cordata*. These genes, which are differentially expressed across various tissues, contribute to our understanding of the molecular mechanisms that underlie flavonoid biosynthesis in medicinal plants.

Quercetin, a type of flavanol, possesses a wide range of biological activities, including antioxidant, anti-inflammatory, and anticancer effects (Di Petrillo et al., 2022; Wang et al., 2022a). We found that quercetin and quercetin glycosides (isoquercitrin, rutin, hyperoside, baimaside, and quercitrin) were present at high levels in the leaves, inflorescences, and bracts of *H. cordata*. The process of glycosylation, which involves the addition of sugar moieties to flavonoid compounds, is facilitated by an enzyme called glucosyltransferase (UGT). This enzymatic activity results in the formation of various flavonoid glycosides, including flavonol aglycones that are known for their high bioavailability and bioactivity (Ross et al., 2001). Our current study identified several UGT genes that are differentially expressed in different tissues of *H. cordata*, suggesting their involvement in quercetin glycoside formation. Further investigation of their functions will provide valuable insights into the quercitin biosynthetic pathway in plants and can support future synthetic biology.

In summary, we assembled a near-complete genome of *H. cordata* with 76 chromosomes ($4n = 76$) and identified key genes involved in the biosynthesis of houttuynin and flavonoids. Our findings establish a genomic foundation that enhances our understanding of *H. cordata*'s medicinal properties and supports future research and genetic improvements to maximize its therapeutic potential.

METHODS

Sample preparation and genome sequencing

In June 2022, we obtained botanical samples of *H. cordata* from Kaili (107°58′ E, 26°05′–27°38′ N) in Guizhou Province, China. To ensure sample quality, we carefully selected a limited number of healthy and pathogen-free plants. High-molecular-weight genomic DNA for genome sequencing was isolated from young leaves using a standard CTAB (hexadecyltrimethylammonium bromide) protocol, and DNA libraries were constructed for the PacBio HiFi reads. The young leaves were used for cell culture and crosslinking of chromatin to construct the Hi-C library.

Genome survey and *de novo* genome assembly

Data-quality assessment and pre-processing were carried out using fastp v0.20.1 (Chen, 2023b) with default options. We used clean data for k-mer analysis with jellyfish 1.1.6 (Marçais and Kingsford, 2011) using a k-mer size of 17 bp. The k-mers from sequencing reads were visualized for Smudgeplot (v0.2.5) (Ranallo-Benavidez et al., 2020), inferring the reference genome as autotetraploid (4n; Supplemental Figure 1B). We then used genomescope 2.0 (Ranallo-Benavidez et al., 2020) to estimate the genome size, heterozygosity, and repeat content in a polyploid model.

The *H. cordata* genome was assembled using HiFi and Hi-C sequencing data. SMRTbell libraries were sequenced on a PacBio Sequel II system, and consensus reads (HiFi reads) were generated using CCS software (https://github.com/pacificbiosciences/ccs) with default parameters. Hifiasm v0.166-r375 was used to construct the primary genome and haplotype draft contig genomes from the long and highly accurate HiFi and Hi-C reads (Cheng et al., 2021a; Han et al., 2022). The Hi-C reads were quality controlled and mapped to the *H. cordata* contig assembly using Juicer (Durand et al., 2016). The 3D-DNA pipeline (v180419) was then used to generate a candidate chromosome-level assembly, correct mis-joins, and order and orient the contigs. Manual inspection and refinement of the draft assembly were performed using Juicebox Assembly Tools (Seppey et al., 2019). Gap filling was carried out using LR_Gapcloser software (v1.0) with HiFi reads. The chloroplast and mitochondrial genomes were separately assembled using GetOrganelle software (v1.7.5) using the PacBio CCS data (Xu et al., 2019; Baeza et al., 2023). Two rounds of polishing were performed on the short-read data using nextpolish (Chang et al., 2023). The completeness of the chromosome-level genome was evaluated using BUSCO software (v4.0.6) (Vanholme et al., 2019) with the embryophyta_odb10 ortholog set.

Genome annotation

To identify transposable elements in the *H. cordata* genome, we used the EDTA pipeline v1.8 (Ou et al., 2019). This pipeline enabled us to identify LTR, TIR, and non-TIR elements. To further analyze the genome and annotate the TEs, we used RepeatMasker (v4.1.6). In addition, the MAKER2 pipeline (Stanke et al., 2008) was used to predict the structures of coding genes. This pipeline incorporated three main approaches: *ab initio* predictions, protein homology, and transcriptome data. Prior to gene prediction, the assembled genome underwent hard and soft masking using RepeatMasker (v4.1.6). *Ab initio* gene prediction was performed using Augustus v3.1 (Stanke et al., 2006), and homology-based gene prediction was carried out using Exonerate v2.2.2 (Slater and Birney, 2005). To predict gene models, EVidenceModeler v2.1.0 was used to integrate the results from all three prediction methods (Haas et al., 2008).

The functions of protein-coding genes were determined using three methods. First, eggNOG-mapper v2.1.12 (Huerta-Cepas et al., 2017) annotation was used to compare genes with the eggNOG homologous gene database, enabling the annotation of gene functions, including GO and KEGG annotations. Second, a sequence similarity search was performed using Diamond v2.1.3 (Buchfink et al., 2015) to compare protein sequences with protein databases such as Swiss-Prot and NR, enabling the characterization of protein sequences. Finally, a structural domain similarity search was performed using InterProScan 5.68-100.0 to compare subdatabases in InterPro, thereby obtaining conserved amino acid sequences, motifs, and domains of the proteins (Jones et al., 2014). tRNAScan-SE 2.0 (Lowe and Eddy, 1997) was used with default parameters to identify tRNA genes, and Barrnap 0.8 (Lagesen et al., 2007; Nawrocki and Eddy, 2013) was used to predict rRNA sequences. Infernal 2.5.2 (Nawrocki and Eddy, 2013) was used to search for non-coding RNAs in the Rfam database.

The original sequence data of the *H. cordata* genome produced in this study have been deposited in the Genome Sequence Archive (Genomics, Proteomics and Bioinformatics, 2021) of the National Center for Biological Information of China/National Genomics Data Center of Beijing Institute of Genomics, Chinese Academy of Sciences (Nucleic Acid Research Center, 2022) (GSA: CRA010237). Open access to the data is available at https://ngdc.cncb.ac.cn/gsa.

Haplotype comparison

We constructed pairwise alignments of the sequences of the four haplotypes using minimap2 (2.24-r1122) (Li, 2021) with the -asm5 option, then used SyRI (1.6.3) (Goel et al., 2019) to convert the corresponding BAM files into variant information files. We then used plotsr (0.5.4) (Goel and Schneeberger, 2022) to visualize all variant information files. The variant information files between haplotypes were also uploaded to cloud storage: https://drive.google.com/drive/folders/1seLh1gEZZyf2E2KrVr3hkZK3fZ6BAcZI?usp=sharing.

Phylogenetic analysis

Gene-evolution analysis was performed by identifying paralogs and orthologs among 21 plant species. The orthologous gene groups were clustered using OrthoFinder2 with default parameters (Emms and Kelly, 2019). The amino acid sequences were aligned using MAFFT v7 (Katoh and Standley, 2013) and trimmed with trimAl v2.0 (Capella-Gutiérrez et al., 2009). A maximum-likelihood phylogenetic tree was constructed using RAxML v7.2.8 with 1000 bootstrap replicates based on the PROTGAMMAJTT model (Stamatakis, 2014). *Amborella trichopoda* was used as the outgroup. The species tree was then used to estimate divergence times using the MCMCTree program in the PAML package v4.10.6 (Yang, 2007). Multiple fossil times were obtained from TimeTree (http://www.timetree.org/) and used for time calibrations. CAFE5 (Mendes et al., 2021) was used to infer the expansion and contraction of gene families on the basis of the chronogram of 21 plant species mentioned above.

Synteny analysis and gene-duplication identification

Syntenic blocks within and between species were identified using MCScanX-h v1.1.11 (Wang et al., 2012) with homologous gene sets and BLASTP. To predict WGD events in one species and estimate the divergence time between two species, we analyzed the distribution of synonymous substitution rates (K_s) (Huang et al., 2023). Gene families were subjected to GO and KEGG analyses using the R package *clusterProfiler* v3.19 for further investigation (Wu et al., 2021).

Synteny network analyses and phylogenetic profiling

Twenty-one plant species were analyzed, including seven Magnoliaceae, five monocots, six eudicots, and two basal angiosperms. Protein sequences and GFF/GFF3 attachments from fully sequenced genomes were obtained from various databases, such as Ensembl (https://plants.ensembl.org/index.html), National Omics Center (https://www.nstda.or.th/noc/), CGDB (http://bio2db.com/), Phytozome (https://phytozome-next.jgi.doe.gov/), NCBI (https://www.ncbi.nlm.nih.gov/), and CPBD (http://citrus.hzau.edu.cn/download.php). Multiple protein sequence alignments were created using MUSCLE v3.8.1551 (Edgar, 2004), and phylogenetic trees were built using the maximum-likelihood technique with IQ-TREE v2.1.4 (-m MFP -bb 1000) (Nguyen et al., 2015). The resulting trees were annotated and visualized using iTOL v6 (Letunic and Bork, 2016) and were grouped on the basis of gene functions and species relatedness. To identify collinear genes, BLASTP v2.10.0 was used to compare protein sequences from each of the 21 plant genomes, with an E-value cutoff of 1e−5. Genomic collinearity was detected and a library dataset of all possible syntenic gene pairs among the 21 plant genomes was constructed using MCScanX (Wang et al., 2012) with default parameters. Protein sequences were compared using MUSCLE v5.1 (Edgar, 2004) and then transformed into codon comparisons using PAL2NAL (Suyama et al., 2006). Finally, the K_a and K_s values between homologous gene pairs were calculated using the YN model (Yang and Nielsen, 2000).

Karyotype analysis of *H. cordata*

We used karyotype analysis to identify and analyze the chromosomes of *H. cordata* (Wang et al., 2022b). Root tips from *H. cordata* plants were collected and treated with a colchicine solution, then fixed using Kanoy solution to ensure that the root-tip cells were in metaphase. The fixed root-tip samples were then hydrolyzed and stained with Giemsa dye. Finally, the root-tip samples were prepared using the extrusion method. The chromosomes of *Houttuynia* were captured and measured using a high-resolution optical microscope. The chromosomes were classified on the basis of their size, banding pattern, and centromeric location.

To clarify chromosomal alterations in *H. cordata*, we used the core eudicot karyotype (ACEK; $n = 21$) (Wang et al., 2022) as a reference to perform a comparative genomic analysis with *Houttuynia* and *A. contorta* using WGDI (version 0.6.2) (Sun et al., 2022).

RNA extraction, library construction, and sequencing

We collected samples from four different tissue segments during the blooming period: leaves, rhizomes, inflorescences, and bracts. To maintain sterility and preserve the samples, we followed a rigorous protocol. Total RNA was extracted using TRIzol reagent according to the manufacturer's instructions (Yang et al., 2020). RNA libraries were prepared using the NEBNext Ultra RNA Library Prep Kit for Illumina and sequenced on the Illumina NovaSeq 6000 platform to capture comprehensive transcriptome profiles (Mildrum et al., 2020). The quality of the cDNA libraries was assessed using an Agilent Bioanalyzer 2100 instrument, and three replicates were performed for each sample (Grissom et al., 2005).

Identification of genes involved in flavonoid biosynthesis

We performed transcriptomic and metabolic profiling of *H. cordata* leaves, rhizomes, bracts, and inflorescences, focusing on flavonoid biosynthetic genes in 11 gene families, namely PAL, 4CL, C4H, CHI, CHS, FLS, F3H, ANS, DFR, IFS, and FNS. The data were visualized using phylogenetic trees, sequence alignments, gene structure diagrams, and heatmaps (Chen et al., 2023a). These graph panels were combined or arranged using TBtools v2.102 (Chen et al., 2023a).

Transcriptomic analysis

To ensure data integrity, our data pre-processing included rigorous quality assessment and pre-processing using FastQC v0.12.1 and fastp v0.20.1 (Zarour et al., 2021) with default options. These stringent filtering steps were designed to generate a high-quality dataset for subsequent analysis, and metrics such as Q20, Q30, and GC content were calculated. The filtered reads were aligned to the reference genome using HISAT2 v2.2.1 (Kim et al., 2019) in paired-end mode. The results were processed with SAMtools v1.10 (Li et al., 2009) to sort and index the alignment files. Read counting was performed using featureCounts v2.0.1 (Liao et al., 2014) to quantify gene counts, and the gene counts were normalized into transcripts per million (TPM) expression levels. To identify DEGs among different plant tissues, stringent criteria were applied (Shi and Gu, 2020). A $\log_2$ fold change of ≥ 1 and a significance level (p_{adj}) of $\leq$ 0.05 were required.

Transcriptional regulation of flavonoid biosynthesis

We performed a series of gene-expression and co-expression network analyses using the entire transcriptome dataset to uncover transcriptional regulatory networks involving the flavonoid biosynthetic genes and TFs. Genes with TPM < 1 in all tissues were filtered out, and the remaining genes were used to construct a co-expression network using WGCNA (Chen and Ma, 2021). The co-expression network modules were obtained using the blockwise modules function with the following parameters: soft threshold power = 14; TOMtype = signed; merge-CutHeight = 0.25; and minModuleSize = 50. PlantTFDB (Guo et al., 2008) was then used with default parameters to identify TFs in the *H. cordata* genome. Finally, the networks between the genes and TFs were visualized in Cytoscape v3.7.1 (Chen et al., 2023c).

Metabolomic analysis

Initial preparation of plant samples for metabolomic analysis included lyophilization and pulverization of individual plant parts, i.e. leaves, rhizomes, inflorescences, and bracts. Metabolites were then extracted from 100 mg of each pulverized sample using a well-established method (Meng et al., 2022). We performed liquid chromatography–mass spectrometry (LC–MS) using the high-sensitivity SCIEX QTRAP 6500+ system for targeted metabolomics studies. This platform enabled us to perform a class-targeted metabolomics analysis, providing detailed insights into the metabolite profiles relevant to our study. Metabolites that met specific criteria, including a VIP (variable importance in projection) score of ≥ 1 and fold change of ≥ 2 or ≤ 0.5, were classified as differentially abundant metabolites. The analysis was performed rigorously, with three independent replicates for each plant tissue. To determine variations in metabolite composition among different components of *H. cordata*, advanced statistical methodologies were used. The identified metabolites were then annotated using the KEGG compound database and mapped to the KEGG pathway database (Altman et al., 2013). This systematic approach enabled the identification of significantly enriched metabolic pathways by assessing the statistical significance of the overlap between identified metabolites and known metabolic pathways using a hypergeometric test.

Statistical analysis of metabolite content and antioxidant capacity

Metabolite contents and antioxidant capacity were quantitatively analyzed using standard curves of known concentrations. All statistical analyses were performed using R software (v4.0.3) with the *stats* package (v4.0.3) for ANOVA and the *agricolae* package (v1.3–3) for Tukey's HSD test.

Determination of sodium houttuyfonate content

To precisely quantify SH in *H. cordata*, various plant parts were heat blanched at 105°C for 2–10 min, then dried to a constant weight at 60°C. The dried samples were ground and sieved (40–60 mesh). For the extraction, an ultrasonic extraction method was used (Cui et al., 2022). A mixture of 0.02 M K_2HPO_4, 0.02 M KH_2PO_4, and methanol was prepared in a 10:10:80 ratio, with the pH adjusted to 9. The extraction was carried out at 60°C for 120 min. After extraction, centrifugation was performed at 25°C at 2000 rpm for 10 min. The supernatant was filtered through a 0.22-μm microporous filter for chromatographic analysis. The chromatographic analysis used a C18 column (250 cm × 4.6 mm) with the same mobile phase. An isocratic elution mode was used at a flow rate of 0.5 ml/min, and the column temperature was maintained at room temperature (Schellinger and Carr, 2006). The detection wavelength was set to 286 nm, and the total detection duration was 18 min.

Mass spectrometric identification of sodium houttuyfonate

We used an Agilent 1290 ultra-high performance LC system coupled with a Thermo Fisher Q-Exactive Orbitrap Plus high-resolution mass spectrometer to accurately determine the concentration of SH (Liu et al., 2019). Chromatographic separation was achieved using a Waters T3 column (21 mm × 50 mm, 1.8 μm), with precise temperature control at 50°C. The mobile phases consisted of 0.1% formic acid in water (phase A) and acetonitrile (phase B), delivered at a flow rate of 0.3 ml/min. We carefully designed a gradient program, starting with 2% B for the first minute, followed by a linear increase to 100% B over 1–18 min, maintenance at 100% B until 22 min, then a rapid reduction to 2% B over 0.1 min and maintenance at 2% B for the final 2.9 min. The injection volume was 5 μl. For MS analysis, we set dynamic data acquisition to cover a range of 100–1000 *m/z*. We fine-tuned the ESI source parameters, including an auxiliary gas flow rate of 16 Arb, full MS resolution at 70 000, and a collision energy (NCE) of 25 eV for MS/MS mode with a resolution of 17 500. The spray voltage was set to −3.0 kV for negative mode and 3.0 kV for positive mode, with a dynamic exclusion duration of 4 s and energy steps in NCE of 10, 30, and 50. Data acquisition ranged from *m/z* 100 to 1000. We performed data interpretation and analysis using Xcalibur 4.3 and MS-DIAL 5.0.3 workstations (Tsugawa et al., 2015). To validate the identification of compounds, we cross-referenced them with the standard SH (CAS 83766-73-8), searched through local databases, and confirmed diagnostic ions. In addition, we analyzed primary quasi-molecular ions and secondary MS fragmentation patterns to ensure a high level of accuracy in compound identification.

Plasmid construction

To generate the constitutive expression constructs, YFP sequences were initially amplified using KOD-Plus DNA polymerase (TOYOBO, www.toyobo-global.com). The amplified sequences were then cloned into the pSK34 vector (Banno et al., 2001) using the SpeI and NotI restriction sites to create the pSKY36 vector. Sequences encoding full-length AP2_2 (Hocor06aG0037700) and MYB_2 (Hocor18aG0011700) (amino acids 1–356) were amplified and inserted into pSKY36 at the AscI and SpeI sites as described in Yang et al. (2014b), yielding the constructs 35S::AP2_2-YFP and 35S::MYB_2-YFP. The details of all primers used are provided in Supplemental Table 22.

Protein subcellular localization and gene-expression analysis

Subcellular localization of AP2_2 and MYB_2 proteins was assessed by transient expression experiments in tobacco (*Nicotiana benthamiana*) as detailed in Yang et al. (2014a). Various constructs were transformed into *Agrobacterium tumefaciens* strain GV3101, which was subsequently suspended in infiltration medium consisting of 1 mM MES (2-(*N-morpholino)ethanesulfonic acid;* pH 5.6), 10 mM magnesium chloride,

and 200 μM acetosyringone. The bacterial cultures were adjusted to an OD_{600} of approximately 0.6. The constructs were transiently expressed in tobacco leaf epidermal cells via agroinfiltration. Fluorescence was observed 3 days post infiltration using a Zeiss LSM 710 confocal microscope (www.zeiss.com). For nuclear staining, leaf samples were stained with 1 μg/ml DAPI (4′,6-diamidino-2-phenylindole; Sigma-Aldrich) for 10–30 min. Gene-expression analysis was performed using total RNA extracted from treated leaves. RT–qPCR and RT–PCR were performed as described previously (Liu et al., 2007; Yang et al., 2023). Two micrograms of total RNA was used for first-strand cDNA synthesis using M-MLV reverse transcriptase (Invitrogen, www.invitrogen.com) following the manufacturer's protocol. *Actin* and *GAPDH* served as internal controls for normalization in the RT–qPCR analyses.

DATA AND CODE AVAILABILITY

We have uploaded the original sequence data for RNA sequencing and whole-genome assembly of *H. cordata* to NCBI and the Genome Sequence Archive at the China National Genomics Data Center (https://ngdc.cncb.ac.cn). The BioProject numbers are PRJNA940964 and PRJCA015615.

FUNDING

This project was funded by the National Natural Science Foundation of China, China (grant number 32360074), the Guizhou Provincial Natural Science Foundation of Department of Education,China ([2022]077), The Karst Mountain Ecological Security Engineering Research Center, China (KY [2021]007), and the Joint Fund of the National Natural Science Foundation of China and the Karst Science Research Center of Guizhou Province, China (U1812401).

ACKNOWLEDGMENTS

We would like to thank Prof. Hong Ma from Penn State University for helpful discussions. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

AUTHOR CONTRIBUTIONS

Z.Y., R.X., and J.-X.L. designed the research. R.X., Y.M., and F.W. performed the data analysis. Y.A., H.Y., and M.T. collected the samples. F.H. and S.F. uploaded the raw data and performed the experiments. R.X., K.L., and J.-X.L. analyzed the results. J.-X.L., Z.Y., K.L., and R.X. wrote the manuscript. All authors contributed to the article and approved the submitted version.

SUPPLEMENTAL INFORMATION

Supplemental information is available at *Plant Communications Online*.

Received: February 6, 2024
Revised: June 7, 2024
Accepted: August 29, 2024
Published: September 2, 2024

REFERENCES

Altman, T., Travers, M., Kothari, A., Caspi, R., and Karp, P.D. (2013). A systematic comparison of the MetaCyc and KEGG pathway databases. BMC Bioinf. **14**:112–115. https://doi.org/10.1186/1471-2105-14-112.

Amborella Genome Project. (2013). The Amborella genome and the evolution of flowering plants. Science (New York, N.Y.) **342**:1241089.

Baeza, J.A., Barata, R., Rajapakse, D., Penaloza, J., Harrison, P., Haberski, A., Harrison, P., and Haberski, A. (2023). Mitochondrial Genomes Assembled from Non-Invasive eDNA Metagenomic Scat Samples in Critically Endangered Mammals. Genes **14**:657. https://doi.org/10.3390/genes14030657.

Bahadur Gurung, A., Ajmal Ali, M., Lee, J., Abul Farah, M., Mashay Al-Anazi, K., and Al-Hemaid, F. (2021). Identification of SARS-CoV-2 inhibitors from extracts of *Houttuynia cordata* Thunb. Saudi J. Biol. Sci. **28**:7517–7527. https://doi.org/10.1016/j.sjbs.2021.08.100.

Bai, L., Li, X., He, L., Zheng, Y., Lu, H., Li, J., Zhong, L., Tong, R., Jiang, Z., Shi, J., et al. (2019). Antidiabetic Potential of Flavonoids from Traditional Chinese Medicine: A Review. Am. J. Chin. Med. **47**:933–957. https://doi.org/10.1142/s0192415x19500496.

Banno, H., Ikeda, Y., Niu, Q.W., and Chua, N.H. (2001). Overexpression of *Arabidopsis* ESR1 induces initiation of shoot regeneration. Plant Cell **13**:2609–2618. https://doi.org/10.1105/tpc.010234.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods **12**:59–60. https://doi.org/10.1038/nmeth.3176.

Cao, Y., Li, K., Li, Y., Zhao, X., and Wang, L. (2020). MYB Transcription Factors as Regulators of Secondary Metabolism in Plants. Biology **9**:61. https://doi.org/10.3390/biology9030061.

Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics **25**:1972–1973. https://doi.org/10.1093/bioinformatics/btp348.

Chang, J., Stahlke, A.R., Chudalayandi, S., Rosen, B.D., Childers, A.K., and Severin, A.J. (2023). polishCLR: A Nextflow Workflow for Polishing PacBio CLR Genome Assemblies. Genome Biol. Evol. **15**:evad020. https://doi.org/10.1093/gbe/evad020.

Chen, C., Wu, Y., Li, J., Wang, X., Zeng, Z., Xu, J., Liu, Y., Feng, J., Chen, H., He, Y., et al. (2023a). TBtools-II: A "one for all, all for one" bioinformatics platform for biological big-data mining. Mol. Plant **16**:1733–1742. https://doi.org/10.1016/j.molp.2023.09.010.

Chen, J., Wang, W., Shi, C., and Fang, J. (2014). A comparative study of sodium houttuyfonate and 2-undecanone for their in vitro and in vivo anti-inflammatory activities and stabilities. Int. J. Mol. Sci. **15**:22978–22994. https://doi.org/10.3390/ijms151222978.

Chen, S. (2023b). Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta **2**:e107. https://doi.org/10.1002/imt2.107.

Chen, X., and Ma, J. (2021). Weighted gene co-expression network analysis (WGCNA) to explore genes responsive to Streptococcus oralis biofilm and immune infiltration analysis in human gingival fibroblasts cells. Bioengineered **12**:1054–1065. https://doi.org/10.1080/21655979.2021.1902697.

Chen, Y.C., Smith, M., Ying, Y.L., Makridakis, M., Noonan, J., Kanellakis, P., Rai, A., Salim, A., Murphy, A., Bobik, A., et al. (2023c). Quantitative proteomic landscape of unstable atherosclerosis identifies molecular signatures and therapeutic targets for plaque stabilization. Commun. Biol. **6**:265. https://doi.org/10.1038/s42003-023-04641-4.

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., and Li, H. (2021a). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat. Methods **18**:170–175. https://doi.org/10.1038/s41592-020-01056-5.

Cheng, Q.Q., Ouyang, Y., Tang, Z.Y., Lao, C.C., Zhang, Y.Y., Cheng, C.S., and Zhou, H. (2021b). Review on the Development and Applications of Medicinal Plant Genomes. Front. Plant Sci. **12**:791219. https://doi.org/10.3389/fpls.2021.791219.

Cheng, T., Xu, C., Wu, D., Yan, G., Wang, C., Wang, T., and Shao, J. (2023). Sodium houttuyfonate derived from *Houttuynia cordata* Thunb improves intestinal malfunction via maintaining gut microflora stability in Candida albicans overgrowth aggravated ulcerative colitis. Food Funct. **14**:1072–1086. https://doi.org/10.1039/d2fo02369e.

Cui, L., Ma, Z., Wang, D., and Niu, Y. (2022). Ultrasound-assisted extraction, optimization, isolation, and antioxidant activity analysis of

flavonoids from Astragalus membranaceus stems and leaves. Ultrason. Sonochem. **90**:106190. https://doi.org/10.1016/j.ultsonch.2022.106190.

Datta, P.C., and Dasgupta, A. (1977). Comparison of vegetative anatomy of piperales. I. Juvenile xylem of twigs. Acta Biol. Acad. Sci. Hung. **28**:81–96.

Di Petrillo, A., Orrù, G., Fais, A., and Fantini, M.C. (2022). Quercetin and its derivates as antiviral potentials: A comprehensive review. Phytother Res. **36**:266–278. https://doi.org/10.1002/ptr.7309.

Dias, M.C., Pinto, D.C.G.A., and Silva, A.M.S. (2021). Plant Flavonoids: Chemical Characteristics and Biological Activity. Molecules **26**:5377. https://doi.org/10.3390/molecules26175377.

Dong, N.Q., and Lin, H.X. (2021). Contribution of phenylpropanoid metabolism to plant development and plant-environment interactions. J. Integr. Plant Biol. **63**:180–209. https://doi.org/10.1111/jipb.13054.

Dong, S., Zhao, C., Chen, F., Liu, Y., Zhang, S., Wu, H., Zhang, L., and Liu, Y. (2018). The complete mitochondrial genome of the early flowering plant Nymphaea colorata is highly repetitive with low recombination. BMC Genom. **19**:614. https://doi.org/10.1186/s12864-018-4991-4.

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 Syst. **3**:95–98. https://doi.org/10.1016/j.cels.2016.07.002.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**:1792–1797. https://doi.org/10.1093/nar/gkh340.

Ekalu, A., and Habila, J.D. (2020). Flavonoids: isolation, characterization, and health benefits. Beni. Suef. Univ. J. Basic Appl. Sci. **9**:45. https://doi.org/10.1186/s43088-020-00065-9.

Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. **20**:238. https://doi.org/10.1186/s13059-019-1832-y.

Goel, M., and Schneeberger, K. (2022). Plotsr: visualizing structural similarities and rearrangements between multiple genomes. Bioinformatics **38**:2922–2926. https://doi.org/10.1093/bioinformatics/btac661.

Goel, M., Sun, H., Jiao, W.B., and Schneeberger, K. (2019). SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol. **20**:277–313. https://doi.org/10.1186/s13059-019-1911-0.

Grissom, S.F., Lobenhofer, E.K., and Tucker, C.J. (2005). A qualitative assessment of direct-labeled cDNA products prior to microarray analysis. BMC Genom. **6**:36. https://doi.org/10.1186/1471-2164-6-36.

Guo, A.Y., Chen, X., Gao, G., Zhang, H., Zhu, Q.H., Liu, X.C., Zhong, Y.F., Gu, X., He, K., and Luo, J. (2008). PlantTFDB: a comprehensive plant transcription factor database. Nucleic Acids Res. **36**:D966–D969. https://doi.org/10.1093/nar/gkm841.

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 Biol. **9**:R7. https://doi.org/10.1186/gb-2008-9-1-r7.

Han, H., Wang, C., Yang, X., Wang, L., Ye, J., Xu, F., Liao, Y., and Zhang, W. (2023). Role of bZIP transcription factors in the regulation of plant secondary metabolism. Planta **258**:13. https://doi.org/10.1007/s00425-023-04174-4.

Han, L. (1995). On the evolution and distribution in saururaceae. Acta Bot. Yunnanica **17**:1–3.

Han, X., Li, C., Sun, S., Ji, J., Nie, B., Maker, G., Ren, Y., and Wang, L. (2022). The chromosome-level genome of female ginseng (Angelica sinensis) provides insights into molecular mechanisms and evolution of coumarin biosynthesis. Plant J. **112**:1224–1237. https://doi.org/10.1111/tpj.16007.

He, X., Hu, M., Song, C., Ni, M., Liu, L., Chen, C., and Wu, D. (2023). Sodium New Houttuyfonate Effectively Improves Phagocytosis and Inhibits the Excessive Release of Inflammatory Factors by Repressing TLR4/NF-Kb Pathway in Macrophages. Curr. Pharm. Biotechnol. **24**:1964–1971. https://doi.org/10.2174/1389201024666230418163100.

Huang, F., Chen, P., Tang, X., Zhong, T., Yang, T., Nwafor, C.C., Yang, C., Ge, X., An, H., Li, Z., et al. (2023). Genome assembly of the Brassicaceae diploid Orychophragmus violaceus reveals complex whole-genome duplication and evolution of dihydroxy fatty acid metabolism. Plant Commun. **4**:100432. https://doi.org/10.1016/j.xplc.2022.100432.

Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., von Mering, C., and Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Mol. Biol. Evol. **34**:2115–2122. https://doi.org/10.1093/molbev/msx148.

Hung, P.Y., Ho, B.C., Lee, S.Y., Chang, S.Y., Kao, C.L., Lee, S.S., and Lee, C.N. (2015). *Houttuynia cordata* targets the beginning stage of herpes simplex virus infection. PLoS One **10**:e0115475. https://doi.org/10.1371/journal.pone.0115475.

Jiu, S., Chen, B., Dong, X., Lv, Z., Wang, Y., Yin, C., Xu, Y., Zhang, S., Zhu, J., Wang, J., et al. (2023). Chromosome-scale genome assembly of Prunus pusilliflora provides novel insights into genome evolution, disease resistance, and dormancy release in Cerasus L. Hortic. Res. **10**:uhad062. https://doi.org/10.1093/hr/uhad062.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics **30**:1236–1240. https://doi.org/10.1093/bioinformatics/btu031.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. **30**:772–780. https://doi.org/10.1093/molbev/mst010.

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. https://doi.org/10.1038/s41587-019-0201-4.

Lagesen, K., Hallin, P., Rødland, E.A., Staerfeldt, H.H., Rognes, T., and Ussery, D.W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. **35**:3100–3108. https://doi.org/10.1093/nar/gkm160.

Laldinsangi, C. (2022). The therapeutic potential of *Houttuynia cordata*: A current review. Heliyon **8**:e10386. https://doi.org/10.1016/j.heliyon.2022.e10386.

Lee, J.H., Ahn, J., Kim, J.W., Lee, S.G., and Kim, H.P. (2015). Flavonoids from the aerial parts of *Houttuynia cordata* attenuate lung inflammation in mice. Arch Pharm. Res. (Seoul). **38**:1304–1311. https://doi.org/10.1007/s12272-015-0585-8.

Letunic, I., and Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. **44**:W242–W245. https://doi.org/10.1093/nar/gkw290.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics **30**:923–930. https://doi.org/10.1093/bioinformatics/btt656.

Li, H. (2021). New strategies to improve minimap2 alignment accuracy. Bioinformatics **37**:4572–4574. https://doi.org/10.1093/bioinformatics/btab705.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup, and Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics **25**:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Li, W., Liu, C., Liu, J., Bai, Z., and Liang, Z. (2020). Transcriptomic analysis reveals the GRAS family genes respond to gibberellin in Salvia miltiorrhiza hairy roots. BMC Genom. **21**:727. https://doi.org/10.1186/s12864-020-07119-3.

Lin, C.H., Chao, L.K., Lin, L.Y., Wu, C.S., Chu, L.P., Huang, C.H., and Chen, H.C. (2022). Analysis of Volatile Compounds from Different Parts of *Houttuynia cordata* Thunb. Molecules **27**:8893. https://doi.org/10.3390/molecules27248893.

Liu, J.X., Srivastava, R., Che, P., and Howell, S.H. (2007). An endoplasmic reticulum stress response in *Arabidopsis* is mediated by proteolytic processing and nuclear relocation of a membrane-associated transcription factor, bZIP28. Plant Cell **19**:4111–4119. https://doi.org/10.1105/tpc.106.050021.

Liu, R., Ruan, Y., Liu, Z., and Gong, L. (2019). The role of fluoroalcohols as counter anions for ion-pairing reversed-phase liquid chromatography/high-resolution electrospray ionization mass spectrometry analysis of oligonucleotides. Rapid Commun. Mass Spectrom. **33**:697–709. https://doi.org/10.1002/rcm.8386.

Liu, W., Feng, Y., Yu, S., Fan, Z., Li, X., Li, J., and Yin, H. (2021a). The Flavonoid Biosynthesis Network in Plants. Int. J. Mol. Sci. **22**:12824. https://doi.org/10.3390/ijms222312824.

Liu, X., Zhong, L., Xie, J., Sui, Y., Li, G., Ma, Z., and Yang, L. (2021b). Sodium houttuyfonate: A review of its antimicrobial, anti-inflammatory and cardiovascular protective effects. Eur. J. Pharmacol. **902**:174110. https://doi.org/10.1016/j.ejphar.2021.174110.

Lou, Y., Guo, Z., Zhu, Y., Kong, M., Zhang, R., Lu, L., Wu, F., Liu, Z., and Wu, J. (2019). *Houttuynia cordata* Thunb. and its bioactive compound 2-undecanone significantly suppress benzo(a)pyrene-induced lung tumorigenesis by activating the Nrf2-HO-1/NQO-1 signaling pathway. J. Exp. Clin. Cancer Res. **38**:242. https://doi.org/10.1186/s13046-019-1255-3.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25**:955–964. https://doi.org/10.1093/nar/25.5.955.

Luo, Q., Meng, P.H., Jiang, D.W., Han, Z.M., Wang, Z.H., Tan, G.F., and Zhang, J. (2022). Comprehensive Assessment of *Houttuynia cordata* Thunb., an Important Medicinal Plant and Vegetable. Agronomy-Basel **12**:2582. https://doi.org/10.3390/agronomy12102582.

Mapoung, S., Umsumarng, S., Semmarath, W., Arjsri, P., Srisawad, K., Thippraphan, P., Yodkeeree, S., and Dejkriengkraikul, P. (2021). Photoprotective Effects of a Hyperoside- Enriched Fraction Prepared from *Houttuynia cordata* Thunb. on Ultraviolet B-Induced Skin Aging in Human Fibroblasts through the MAPK Signaling Pathway. Plants **10**:2628. https://doi.org/10.3390/plants10122628.

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics **27**:764–770. https://doi.org/10.1093/bioinformatics/btr011.

Mendes, F.K., Vanderpool, D., Fulton, B., and Hahn, M.W. (2021). CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics **36**:5516–5518. https://doi.org/10.1093/bioinformatics/btaa1022.

Meng, L., Song, W., Chen, S., Hu, F., Pang, B., Cheng, J., He, B., and Sun, F. (2022). Widely targeted metabolomics analysis reveals the mechanism of quality improvement of flue-cured tobacco. Front. Plant Sci. **13**:1074029. https://doi.org/10.3389/fpls.2022.1074029.

Michala, A.S., and Pritsa, A. (2022). Quercetin: A Molecule of Great Biochemical and Clinical Value and Its Beneficial Effect on Diabetes and Cancer. Diseases **10**. https://doi.org/10.3390/diseases10030037.

Mildrum, S., Hendricks, A., Stortchevoi, A., Kamelamela, N., Butty, V.L., and Levine, S.S. (2020). High-throughput Minitaturized RNA-Seq Library Preparation. J. Biomol. Tech. **31**:151–156. https://doi.org/10.7171/jbt.20-3104-004.

Murat, F., Armero, A., Pont, C., Klopp, C., and Salse, J. (2017). Reconstructing the genome of the most recent common ancestor of flowering plants. Nat. Genet. **49**:490–496. https://doi.org/10.1038/ng.3813.

Nabavi, S.M., Samec, D., Tomczyk, M., Milella, L., Russo, D., Habtemariam, S., Suntar, I., Rastrelli, L., Daglia, M., Xiao, J.B., et al. (2020). Flavonoid biosynthetic pathways in plants: Versatile targets for metabolic engineering. Biotechnol. Adv. **38**:107316. https://doi.org/10.1016/j.biotechadv.2018.11.005.

Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics **29**:2933–2935. https://doi.org/10.1093/bioinformatics/btt509.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. **32**:268–274. https://doi.org/10.1093/molbev/msu300.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., Lugo, C.S.B., Elliott, T.A., Ware, D., Peterson, T., et al. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. **20**:275. https://doi.org/10.1186/s13059-019-1905-y.

Ou, S., Chen, J., and Jiang, N. (2018). Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Res. **46**:e126. https://doi.org/10.1093/nar/gky730.

Persson, X.M.T., Blachnio-Zabielska, A.U., and Jensen, M.D. (2010). Rapid measurement of plasma free fatty acid concentration and isotopic enrichment using LC/MS. J. Lipid Res. **51**:2761–2765. https://doi.org/10.1194/jlr.M008011.

Pradhan, S., Rituparna, S., Dehury, H., Dhall, M., and Singh, Y.D. (2023). Nutritional profile and pharmacological aspect of *Houttuynia cordata* Thunb. and their therapeutic applications. Pharmacological Research - Modern Chinese Medicine **9**:100311. https://doi.org/10.1016/j.prmcm.2023.100311.

Qin, L., Hu, Y., Wang, J., Wang, X., Zhao, R., Shan, H., Li, K., Xu, P., Wu, H., Yan, X., et al. (2021). Insights into angiosperm evolution, floral development and chemical biosynthesis from the *Aristolochia fimbriata* genome. Nat. Plants **7**:1239–1253. https://doi.org/10.1038/s41477-021-00990-2.

Rafiq, S., Hao, H., Ijaz, M., and Raza, A. (2022). Pharmacological Effects of *Houttuynia cordata* Thunb (*H. cordata*): A Comprehensive Review. Pharmaceuticals **15**:1079. https://doi.org/10.3390/ph15091079.

Ranallo-Benavidez, T.R., Jaron, K.S., and Schatz, M.C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat. Commun. **11**:1432. https://doi.org/10.1038/s41467-020-14998-3.

Ross, J., Li, Y., Lim, E., and Bowles, D.J. (2001). Higher plant glycosyltransferases. Genome Biol. **2**:Reviews3004. https://doi.org/10.1186/gb-2001-2-2-reviews3004.

Schellinger, A.P., and Carr, P.W. (2006). Isocratic and gradient elution chromatography: A comparison in terms of speed, retention reproducibility and quantitation. J. Chromatogr. A **1109**:253–266. https://doi.org/10.1016/j.chroma.2006.01.047.

Seppey, M., Manni, M., and Zdobnov, E.M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. Methods Mol. Biol. **1962**:227–245. https://doi.org/10.1007/978-1-4939-9173-0_14.

Shen, Y.H., Cheng, M.H., Liu, X.Y., Zhu, D.W., and Gao, J. (2021). Sodium Houttuyfonate Inhibits Bleomycin Induced Pulmonary Fibrosis in Mice. Front. Pharmacol. **12**:596492. https://doi.org/10.3389/fphar.2021.596492.

Shi, P., and Gu, M. (2020). Transcriptome analysis and differential gene expression profiling of two contrasting quinoa genotypes in response to salt stress. BMC Plant Biol. **20**:568. https://doi.org/10.1186/s12870-020-02753-1.

Slater, G.S.C., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinf. **6**:31. https://doi.org/10.1186/1471-2105-6-31.

Smith, S.Y., and Stockey, R.A. (2007). Establishing a fossil record for the perianthless Piperales: Saururus tuckerae sp. nov. (*Saururaceae*) from the Middle Eocene Princeton Chert. Am. J. Bot. **94**:1642–1657. https://doi.org/10.3732/ajb.94.10.1642.

Soltis, D.E., Albert, V.A., Leebens-Mack, J., Bell, C.D., Paterson, A.H., Zheng, C., Sankoff, D., Pamphilis, C.W.D., Wall, P.K., and Soltis, P.S. (2009). Polyploidy and angiosperm diversification. Am. J. Bot. **96**:336–348.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**:1312–1313. https://doi.org/10.1093/bioinformatics/btu033.

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics **24**:637–644. https://doi.org/10.1093/bioinformatics/btn013.

Stanke, M., Tzvetkova, A., and Morgenstern, B. (2006). AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. Genome Biol. **7**:S11. https://doi.org/10.1186/gb-2006-7-s1-s11.

Sun, P., Jiao, B., Yang, Y., Shan, L., Li, T., Li, X., Xi, Z., Wang, X., and Liu, J. (2022). WGDI: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. Mol. Plant **15**:1841–1851. https://doi.org/10.1016/j.molp.2022.10.018.

Su, X., Yang, L., Wang, D., Shu, Z., Yang, Y., Chen, S., and Song, C. (2022). 1 K Medicinal Plant Genome Database: an integrated database combining genomes and metabolites of medicinal plants. Hortic. Res. **9**:uhac075. https://doi.org/10.1093/hr/uhac075.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. **34**:W609–W612. https://doi.org/10.1093/nar/gkl315.

Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., Kanazawa, M., VanderGheynst, J., Fiehn, O., and Arita, M. (2015). MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. Nat. Methods **12**:523–526. https://doi.org/10.1038/nmeth.3393.

Vanholme, R., Sundin, L., Seetso, K.C., Kim, H., Liu, X., Li, J., De Meester, B., Hoengenaert, L., Goeminne, G., Morreel, K., et al. (2019). COSY catalyses trans-cis isomerization and lactonization in the biosynthesis of coumarins. Nat. Plants **5**:1066–1075. https://doi.org/10.1038/s41477-019-0510-0.

Wang, G., Wang, Y., Yao, L., Gu, W., Zhao, S., Shen, Z., Lin, Z., Liu, W., and Yan, T. (2022a). Pharmacological Activity of Quercetin: An Updated Review. Evid. Based. Complement. Alternat. Med. **2022**:3997190. https://doi.org/10.1155/2022/3997190.

Wang, J., Wang, D., Yin, Y., Deng, Y., Ye, M., Wei, P., Zhang, Z., Chen, C., Qin, S., and Wang, X. (2022b). Assessment of Combined Karyotype Analysis and Chromosome Microarray Analysis in Prenatal Diagnosis: A Cohort Study of 3710 Pregnancies. Genet. Res. **2022**:6791439. https://doi.org/10.1155/2022/6791439.

Wang, W., Sun, C., Mao, L., Ma, P., Liu, F., Yang, J., and Gao, Y. (2016). The biological activities, chemical stability, metabolism and delivery systems of quercetin: A review. Trends Food Sci. Technol. **56**:21–38. https://doi.org/10.1016/j.tifs.2016.07.004.

Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H., Jin, H., Marler, B., Guo, H., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. **40**:e49. https://doi.org/10.1093/nar/gkr1293.

Wang, Z., Li, Y., Sun, P., Zhu, M., Wang, D., Lu, Z., Hu, H., Xu, R., Zhang, J., Ma, J., et al. (2022). A high-quality *Buxus austro-yunnanensis* (*Buxales*) genome provides new insights into karyotype evolution in early eudicots. BMC Biol. **20**:216. https://doi.org/10.1186/s12915-022-01420-1.

Wei, P., Luo, Q., Hou, Y., Zhao, F., Li, F., and Meng, Q. (2024). *Houttuynia Cordata* Thunb.: A comprehensive review of traditional applications, phytochemistry. Phytomedicine **123**:155195. https://doi.org/10.1016/j.phymed.2023.155195.

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation **2**:100141. https://doi.org/10.1016/j.xinn.2021.100141.

Xu, G.C., Xu, T.J., Zhu, R., Zhang, Y., Li, S.Q., Wang, H.W., and Li, J.T. (2019). LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. GigaScience **8**:giy157. https://doi.org/10.1093/gigascience/giy157.

Xu, X., Xu, H., Shang, Y., Zhu, R., Hong, X., Song, Z., and Yang, Z. (2021). Development of the general chapters of the Chinese Pharmacopoeia 2020 edition: A review. J. Pharm. Anal. **11**:398–404. https://doi.org/10.1016/j.jpha.2021.05.001.

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., et al. (2023). The *Saururus chinensis* genome provides insights into the evolution of pollination strategies and herbaceousness in magnoliids. Plant J. **113**:1021–1034. https://doi.org/10.1111/tpj.16097.

Yang, D., Wang, T., Long, M., and Li, P. (2020). Quercetin: Its Main Pharmacological Activity and Potential Application in Clinical Medicine. Oxid. Med. Cell. Longev. **2020**:8825387. https://doi.org/10.1155/2020/8825387.

Yang, L., Ji, W., Zhong, H., Wang, L., Zhu, X., and Zhu, J. (2019). Anti-tumor effect of volatile oil from *Houttuynia cordata* Thunb. on HepG2 cells and HepG2 tumor-bearing mice. RSC Adv. **9**:31517–31526. https://doi.org/10.1039/c9ra06024c.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24**:1586–1591. https://doi.org/10.1093/molbev/msm088.

Yang, Z., and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. **17**:32–43. https://doi.org/10.1093/oxfordjournals.molbev.a026236.

Yang, Z.T., Fan, S.X., Wang, J.J., et al. (2023). The plasma membrane-associated transcription factor NAC091 regulates unfolded protein response in *Arabidopsis thaliana*. Plant Science **334**:111777.

Yang, Z.T., Lu, S.J., Wang, M.J., Bi, D.L., Sun, L., Zhou, S.F., Song, Z.T., and Liu, J.X. (2014a). A plasma membrane-tethered transcription factor, NAC062/ANAC062/NTL6, mediates the unfolded protein response in *Arabidopsis*. Plant J. **79**:1033–1043. https://doi.org/10.1111/tpj.12604.

Yang, Z.T., Wang, M.J., Sun, L., Lu, S.J., Bi, D.L., Sun, L., Song, Z.T., Zhang, S.S., Zhou, S.F., and Liu, J.X. (2014b). The membrane-associated transcription factor NAC089 controls ER-stress-induced

programmed cell death in plants. PLoS Genet. **10**:e1004243. https://doi.org/10.1371/journal.pgen.1004243.

Yao, N., Jing, L., Zheng, H., Chen, M., and Shen, Y. (2018). [Research progress of jasmonate-responsive transcription factors in regulating plant secondary metabolism]. Zhongguo Zhongyao Zazhi **43**:897–903. https://doi.org/10.19540/j.cnki.cjcmm.20180109.001.

Yin, Y., Wang, D., Wu, D., He, W., Zuo, M., Zhu, W., Xu, Y., and Wang, L. (2023). Two New 4-Hydroxy-2-pyridone Alkaloids with Antimicrobial and Cytotoxic Activities from Arthrinium sp. GZWMJZ-606 Endophytic with *Houttuynia cordata* Thunb. Molecules **28**:2192. https://doi.org/10.3390/molecules28052192.

Yuan, H., Liu, L., Zhou, J., Zhang, T., Daily, J.W., and Park, S. (2022). Bioactive Components of *Houttuynia cordata* Thunb and Their Potential Mechanisms Against COVID-19 Using Network Pharmacology and Molecular Docking Approaches. J. Med. Food **25**:355–366. https://doi.org/10.1089/jmf.2021.K.0144.

Zarour, M., Alenezi, M., Ansari, M.T.J., Pandey, A.K., Ahmad, M., Agrawal, A., Kumar, R., and Khan, R.A. (2021). Ensuring data integrity of healthcare information in the era of digital health. Healthc. Technol. Lett. **8**:66–77. https://doi.org/10.1049/htl2.12008.

Zhang, L., Lv, H., Li, Y., Dong, N., Bi, C., Shan, A., Wu, Z., and Shi, B. (2020). Sodium houttuyfonate enhances the intestinal barrier and attenuates inflammation induced by Salmonella typhimurium through the NF-κB pathway in mice. Int. Immunopharmacol. **89**:107058. https://doi.org/10.1016/j.intimp.2020.107058.

Zhou, L., Jiao, Y., Tang, J., Zhao, Z., Zhu, H., Lu, Y., and Chen, D. (2022). Ultrafiltration isolation, structure and effects on H1N1-induced acute lung injury of a heteropolysaccharide from *Houttuynia cordata*. Int. J. Biol. Macromol. **222**:2414–2425. https://doi.org/10.1016/j.ijbiomac.2022.10.027.

Zhu, D.W., Yu, Q., Sun, J.J., and Shen, Y.H. (2021). Evaluating the Therapeutic Mechanisms of Selected Active Compounds in *Houttuynia cordata* Thunb. in Pulmonary Fibrosis via Network Pharmacology Analysis. Front. Pharmacol. **12**:733618. https://doi.org/10.3389/fphar.2021.733618.