Decoding the Genetic Basis of Aromatic Terpene Metabolism in Rosemary: Ancient Whole-Genome Duplications and Ancestral Karyotypes Shed Light on Evolutionary Signatures

Dongfeng Yang[1], Ying Cheng[1], * Xin-Yu[1], Jun-Jie Wu[1], Xuan Zhou[1], Gang-Gui Lou[1], Cathie Martin[2], Yue Chen[1], Zhuo-Ni Hou[1], Bei-Mi Cui[3], Fei-Yan Wang[1], Zhe-Chen Qi[1], and Zong-Suo Liang[1]

[1]Zhejiang Sci-Tech University
[2]John Innes Centre
[3]The University of Edinburgh Institute of Molecular Plant Sciences

October 3, 2023

Abstract

Salvia rosmarinus, a commonly known aromatic plant belonging to the *Salvia* genus, is valued for its medicinal properties, derived primarily from the terpenoids present in its leaves. We have successfully created a chromosome-level genome assembly of *S. rosmarinus*, covering 1.24 Gb, with a scaffold N50 value of 107.45 Mb and 61,717 annotated protein-coding genes. Our analysis highlights a recent whole genome duplication (WGD) event as the primary driver of genomic rearrangement and fusion following speciation. As a result of the WGD, key genes involved in monoterpene biosynthesis, such as *HMGR*, 1,8-cineole synthase, and limonene synthase, underwent tandem duplication and double punctuation. Limonene synthase experienced a nonpolar mutation that favored structural diversity in monoterpene biosynthesis, while 1,8-cineole synthase underwent a polar mutation that favored 1,8-cineole(eucalyptol) accumulation. In addition, our analysis revealed differences in the mechanisms of diterpene biosynthesis between *S. rosmarinus* and *S. miltiorrhiza*, as evidenced by the tandem duplication, covariance, and high-level expression of genes essential for carnosol biosynthesis, specifically *CYP76AK6–8*. These findings no punctuation for understanding the molecular-level diversity of terpenoids in *S. rosmarinus* and will facilitate molecular breeding and quality improvement efforts for this economically important plant.

Introduction

Rosemary (*Salvia rosmarinus* Schleid.) is a well-known Mediterranean perennial shrub that belongs to the mint family (Lamiaceae). It has been cultivated worldwide for its culinary, aromatic, ornamental, and therapeutic properties (Allegra, Tonacci, Pioggia, Musolino, & Gangemi, 2020; Degner, Papoutsis, & Romagnolo, 2009; Freedman, 2019; Neves, Neves, & Oliveira, 2018), with a global market size of rosemary products estimated 2,224 million USD in 2021. The use of rosemary dates back to ancient times, with evidence of its use for embalming in Egyptian tombs 3000 B.C. and as a herbal medicine in ancient Greece and Rome 500 B.C. Currently, rosemary extracts are widely used in cooking, food preservation, cosmetics, and herbal medicine due to their high antimicrobial and antioxidant activities (Degner et al., 2009). Rosemary is considered one of the most effective herbs for treating headaches, poor circulation, inflammatory diseases, and physical and mental fatigue (Nematolahi, Mehrabani, Karami-Mohajeri, & Dabaghzadeh, 2018; Ojeda-Sana, van Baren, Elechosa, Juarez, & Moreno, 2013; Rašković et al., 2014).

Essential oils of rosemary contain more than 30 components, including flavones (genkwanin, isoscutellarein 7- O-glucoside), caffeoyl derivatives (rosmarinic acid), phenolic monoterpenes (1,8-cineole (eucalyptol), α-

pinene, camphene, limonene) (al-Sereiti, Abu-Amer, & Sen, 1999; Angioni et al., 2004; Mena et al., 2016; Sharma, Velamuri, Fagan, & Schaefer, 2020), and diterpenes (carnosic acid, carnosol), which are considered as the major bioactive components. Monoterpenes in rosemary are the main source of the aromatic properties of the fragrance and essential oil (Christopoulou et al., 2021; Micić et al., 2021), which have been shown to possess olfactory properties that influence cognitive performance including memory (Moss, Cook, Wesnes, & Duckett, 2003). The diterpenoids in rosemary leaves are reported to be responsible for their antioxidant, antibacterial, and anticancer properties (Alsamri et al., 2021; Bao et al., 2020; Ngo, Williams, & Head, 2011; Veenstra & Johnson, 2021; M. H. Yu et al., 2013). Despite the commercial interest and increasing demand for rosemary, improvements through breeding have been very limited (Maurizio, Francesconi, Perinu, & Vais, 2002). The lack of high-quality genome information has hindered the understanding of how its terpenoid bioactives are made and any improvements in productivity possible through genetic selection. Therefore, understanding the genes responsible for biosynthesis of the various terpenoids made in rosemary and their regulation will lay a foundation for molecular breeding for improved and sustainable production. Rosemary essential oil is characterized by a high content of monoterpenes, including 1,8-cineole, α-pinene, limonene (Flamini et al., 2022; Rašković et al., 2014). In fact, 1,8-cineole is the major constituent of rosemary, accounting for 23%-49% of the oil (Christopoulou et al., 2021; Flamini et al., 2022; Rašković et al., 2014). Compared to other aromatic plants in the mint family, such as mentha, lavender and ocimum, which contain lower levels of 1,8-cineole, ranging from 0.5% and 8% (Pokajewicz, Białoń, Svydenko, Fedin, & Hudz, 2021; Senthoorraja et al., 2021; Singh & Pandey, 2018; Yang, Jeon, Lee, Shim, & Lee, 2010), rosemary is exceptional in its ability to synthesize large amounts of 1,8-cineole (Raal, Orav, & Arak, 2007).

Monoterpenes, including 1,8-cineole, are derived from the precursor geranyl diphosphate (GPP) via the mevalonate pathway (MVA) in the cytosol (Mendoza-Poudereux et al., 2015; Wu et al., 2020). The phosphate bond of GPP is broken by monoterpene synthase, generating the geranyl cation, which is then isomerized and cyclized to form a terpinyl cation intermediate (N. Srividya, Davis, Croteau, & Lange, 2015; J. Xu et al., 2017). Limonene synthase directly catalyzes the deprotonation of terpinyl cation to synthesize limonene. The product profile of any monoterpene synthase is determined by the conformation of its substrate or intermediate in the active site pocket of the enzyme. The terpinyl cation can be further deprotonated to form a more stable intermediate and generate a variety of monoterpene profiles (Gao, Honzatko, & Peters, 2012). Specifically, 1,8-cineole synthase catalyzes, the cyclization of the terpinyl cation and traps water to generate 1,8-cineole (Piechulla et al., 2016; N. Srividya et al., 2015; Wedler, Pemberton, & Tantillo, 2015). Monoterpene synthases share a common tertiary structure, with similar polar pockets that include conserved active site motifs, such as $RR(X)_8W$, which is responsible for substrate isomerization (Williams, McGarvey, Katahira, & Croteau, 1998), an RXR motif that protects the carbocation intermediate against nucleophilic attack (Starks, Back, Chappell, & Noel, 1997), and a DDXXD motif that provides the main divalent metal binding site (Starks et al., 1997). A NALV motif is necessary to produce 1,8-cineole but not alpha-terpineol (Piechulla et al., 2016). Despite the structural elucidation of 1,8-cineole synthase in *Salvia fruticosa* , the molecular and structural basis of 1,8-cineole synthase activity in rosemary remains unclear.

Carnosic acid and carnosol, which are the primary active diterpenes found in *S. rosmarinus* extracts, exhibit significant antioxidant properties (Veenstra & Johnson, 2021). The biosynthesis of these compounds begins with geranylgeranyl diphosphate (GGPP) supplied by the plastidial methylerythritol phosphate pathway (MEP) (Bergman, Davis, & Phillips, 2019; Forestier, Brown, Harvey, Larson, & Graham, 2021). Diterpene synthases initiated diterpenoid biosynthesis, by cyclizing GGPP to form various hydrocarbon backbone structures. Ent-copalyl diphosphate synthase (CPS) and kaurene synthases (KSL) catalyze the cyclization GGPP to form miltiradiene (Su et al., 2016), which can be spontaneously oxidized to ferruginol. The oxidation network of abietane diterpenes is complex in the genus *Salvia* , with cytochromes P450 enzymes of the subfamily CYP76AK serving as C-20 oxidases, contributing to oxygenations at position C-20 (Bathe, Frolov, Porzel, & Tissier, 2019). In *S. rosmarinus* , the genes *CYP76AK7* and*CYP76AK8* encode enzymes that can catalyze three sequential C-20 oxidations, converting 11-hydroxy ferruginol to carnosic acid (Ignea et al., 2016). However, in *S. miltiorrhiza* , a congeneric medicinal plant in East Asia, the *CYP76AK1* gene was found to catalyze a single hydroxylation at position C-20, resulting in the production of 11,20-

hydroxy ferruginol, which is the precursor of tanshinone biosynthesis (Ignea et al., 2016; Scheler et al., 2016). Notably,*CYP76AK6–8* and *CYP76AK1* accept the same substrate, leading to the diversity of diterpenes present in *S. rosmarinus*and *S. miltiorrhiza* (Bathe et al., 2019; Scheler et al., 2016).

Belonging to genus *Salvia* , *S. rosmarinus* is native to the west coast of the Mediterranean Sea and is a typical European species of*Salvia* , *S. miltiorrhiza* is mainly distributed in Eastern Asia, it has been derived into a separate lineage during the long history of evolution, while *S. splendens* is native to South America and it is a common garden ornamental plant. Rosemary has a long history and a solid position in the spice industry, which stems from the ability of producing essential oils in leaves. Most species of the genus*Salvia* do not possess this ability, including *S. miltiorrhiza* and *S. splendens* . In addition, the antioxidant property of rosemary attributed to the diterpenoids in leaves, including carnosic acid and carnosol. *S. miltiorrhiza* , a traditional medicinal plant, has antioxidant property as well. It was mainly attributed to diterpenoids in the hairy root with different structures, such as tanshinone IIA. The cultivation of *S. splendens* was mainly for ornamental purposes, and few secondary metabolites extracted from the plant. Rosemary is aromatic and its ability to produce essential oils is unique within species of genus *Salvia.* In addition, the high levels of carnosic acid and carnosol in rosemary leaves had not been found in other species of the genus *Salvia*except for *Salvia officinalis* , and the reasons behind the secondary metabolites of rosemary are worth of further exploring.

In this study, we present a reference genome sequence of *S. rosmarinus* that was generated by combining Illumina and PacBio data and assembled using Hi-C technologies. The genome was assembled into twelve pseudochromosomes with a super-N50 of 107.45 Mb, totaling 1.24 Gb. Though the previous article of rosemary genome assembly had revealed the biosynthesis of carnosic acid (Han et al., 2023), essential oil was considered as important metabolite of rosemary. We performed comparative genomic analysis with the published genomes of *S. miltiorrhiza*and *S. splendens* , and identified tandem gene duplications encoding 1,8-cineole synthase and limonene synthase, which are highly expressed in leaves and contribute to the large accumulation of monoterpenes, particularly 1,8-cineole. Additionally, we identified the*CYP76AK6–8* genes responsible for carnosol synthesis and used molecular docking to reveal the differential diterpenoid synthesis mechanism between *S. rosmarinus* and *S. miltiorrhiza.*

Material and methods

Ethics statement :

This study did not involve animal experiments

Plant Material

Salvia rosmarinus (rosemary) plants of the same age and variety were obtained from Shanghai Chenshan Botanical Garden. The growth conditions of the seedlings were maintained in a controlled environment with identical temperature and light conditions. Material from a single plant were used for genome sequencing, Hi-C sequencing, metabolite extraction, and RNA-Seq generation. The plant's young leaves were used for genome sequencing and Hi-C library construction, while tissues from three different organs (leaves, stems and roots) were used for metabolites and RNA extraction.

Genome sequencing

Genomic DNA was extracted by SteadyPure plant genomic DNA extraction kit (https://agbio.com.cn/) and sequenced on the Illumina HiSeq X Ten platform and PacBio platform. After quality control of the generated reads, clean data was obtained. We obtained a total of 107 Gb short reads and 60.5 Gb long reads on PacBio, respectively.

Fresh rosemary plant leaf tissue was used for creating the Hi-C library. The tissue sample was cross-linked with formaldehyde for 30 minutes at room temperature. After purification, the DNA was digested with restriction enzymes. Following digestion, the fragments were biotin-labeled, blunt-ended ligated, and DNA

was extracted. The purified DNA was digested into 300– 700 bp fragments, which were used to construct a DNA library and sequenced on the Illumina HiSeq platform, yielding a total of 111 Gb reads.

Genome assembly and quality assessment

To estimate the genome size of *Salvia rosmarinus* , the Illumina genomic reads were used as the input of the Jellyfish (v1.1.10) tool to obtain the k-mer frequency. Genome size was then predicted to be about 1.2 Gb using GenomeScope (v2.0) (Vurture et al., 2017), with a k-mer length of 31.

Due to the high heterozygosity of 1.7% from GenomeScope results, genome assembly was conducted by combining accurate short-reads with long reads to enhance the assembly performance. The PacBio reads were assembled using Canu (v2.0) (Koren et al., 2017) and FALCON-Unzip (v0.4.0) (Chin et al., 2016), which generated the best primary contigs. We derived a reference genome assembly by selecting the best assembly using Canu, and a phased genome assembly by selecting the best assembly using FALCON-Unzip. Primary contigs were then minced in the form of haplotigs pair, and haplotypes were collapsed. The phased genome assembly and Hi-C libraries were provided to Falcon-Phase (Kronenberg et al., 2021) to obtain a normalized contact matrix, which was used to phase the genome into haplotigs. To extend the genome from contig level to scaffold level, the Canu assembly results were used as the reference genome and Hi-C data were compared to the reference genome by BWA, which was set to strict mode (-n 0) in order to improve the linkage quality, and read pairs were spliced to scaffolds when compared to different contigs. Ultimately, we obtained a phased chromosome-level genome assembly of*S. rosmarinus* .

To assess the quality of the assembly, short reads were mapped to the*S. rosmarinus* genome assembly using BWA software (v0.7.12) (H. Li & Durbin, 2010), with low-quality reads were filtered out (Phreads < 30). Annotation of *S. rosmarinus* and *S. baicalensis* were added to Allele table with BLASTN identity $< 60\%$ and coverage $< 80\%$ in order to filter out noisy signals. All contigs were assigned to 12 pseudochromosomes by partitioning and rescuing. After ordering and format conversion, the rosemary genome was finally assembled. The Benchmarking Universal Single-Copy Orthologs (BUSCOs) (v5.1.2) (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015) pipeline was utilized to conduct an independent assessment of the assembly quality.

Gene prediction and functional annotation

BLASTP (E-value cutoff 1e-05) (*https://blast.ncbi.nlm.nih.gov/Blast.cgi*) comparison searches were first performed between the predicted protein sequences of genes and the entries in the public protein sequence database, including NR and Swiss-Prot, to obtain functional annotations. InterProScan (Jones et al., 2014) was then employed to compared protein domains and functional site databases to further identify protein function. Gene Ontology (GO) terms were derived from corresponding InterPro or Pfam entries. Pathways reconstruction was carried out using KOBAS and Kyoto Encyclopedia of Genes and Genome (KEGG) databases (*http://www.genome.jp/kegg/*).

Repeat sequences were annotated using a combined strategy. We first used LTR_FINDER (http://tlife.fudan.edu.cn/ltr_finder/) to search RepBase database with default settings, and then constructed a *de novo* library using RepeatModeler (http://www.repeatmasker.org/RepeatModeler/html/). Known repeat sequences were identified using RepeatMasker (http://www.repeatmasker.org) based on the Repbase-derived RepeatMasker library and the de novo library. We predicted rRNAs using RNAmmer (http://www.cbs.dtu.dk/services/RNAmmer/) and annotated ncRNAs and sRNAs with tRNAscan-SE (http://lowelab.ucsc.edu/tRNAscan-SE/). In addition, we identified other types of RNAs, including miRNAs and snRNAs, by searching the Rfam database with INFERNAL (http://infernal.janelia.org/).

Phylogenomic analysis

To construct a phylogenetic tree, we used the predicted protein files from the *S. rosmarinus* genome and 23 other species (*Antirrhinum majus* , *Arabidopsis thaliana* , *Daucus carota* , *Boea hygrometrica* , *Coffea canephora* ,*Amborella trichopoda* , *Erythranthe guttata* , *Sesamum indicum* , *Scutellaria baicalensis* , *Glycine*

max ,*Handroanthus impetiginosus* , *Olea europaea* , *Oryza sativa* , *Populus trichocarpa* , *Beta vulgaris* , *Striga asiatica* , *Solanum lycopersicum* , *Vitis vinifera* , *Zea mays* , *Ocimum tenuiflorum* , *Salvia miltiorrhiza* ,*Salvia splendens* , *Tectona grandis*).. Based on 465 single-copy genes, MUSCLE was used for the sequence alignments and matrix construction. A maximum-likelihood phylogenetic tree was then constructed by IQtree2 (v20151210) with the 'MFP+MERGE' model. The divergence time of among 24 plants was predicted using BEAST2 (v2.5.4) (http://www.beast2.org/) with a strict molecular clock model, and time scales were calibrated by the divergence time of the fossil record of species from TimeTree (http://www.timetree.org). The results of OrtherFinder, CAFÉ 5 (https://github.com/hahnlab/CAFE) (Mendes, Vanderpool, Fulton, & Hahn, 2020) were used to analyze the expansion and contraction of gene families. Gene families were regarded as significantly expanded or contracted if the p-value was less than 0.05 in all species. Covariance analysis within genes were performed by jcvi (Tang et al., 2015). A gene pair was considered to have a covariance relationship while cscore was greater than 0.7. Homologous gene pairs were marked with strips and target genes with colorful trips.

Ancient WGD event prediction

To identify the ancient replication events of *S. rosmarinus* , WGD analysis was performed with the predicted proteins file from its genome. We used wgd (v1.0.1) (https://github.com/arzwa/wgd) (Zwaenepoel & Van de Peer, 2019) and MCScanX (Wang et al., 2012) to detect genome-wide replication events in the species by calculating the synonymous substitution rate (*Ks*). On the basis of the most representative transcripts identified using the script of CAFÉ 5 ('cafetutorial_longest_iso.py') (https://github.com/hahnlab/CAFE) (Mendes et al., 2020), sequence alignments and analysis of all gene clustering were performed by wgd. We then calculated the *Ks* distribution and plotted this using R (v4.0.1) (https://www.r-project.org/). And we calculated and compared the Ks distribution on the internal collinearity of gene pairs of *S. rosmarinus* and the Ks distribution based on the orthologous gene pairs between *S. rosmarinus* and *S. miltiorrhiza* and*S. splendens* . According to the normal distribution peaks in the distribution, putative whole-genome duplication events could be identified within species.

Reconstructing the ancestral karyotypes of S. rosmarinus, S. miltiorrhiza and S. splendens

AEK (ancestral eudicot karyotype) was reconstructed from a grape–cacao–peach comparison , and obtained seven protochromosomes with 6,284 ordered protogenes. AEK genome was used for karyotype projections of *S. rosmarinus* , *S. miltiorrhiza* and *S. splendens* . To obtain the karyotype changes of species, we inferred the possible chromosomal evolution process through the comparison of homologous regions between species and AEK genome, and represented seven protochromosomes with different colors.

Karyotype projections indicated the large amount of chromosomal rearrangements occurred in *S. rosmarinus* , *S. miltiorrhiza*and *S. splendens* . Chromosomes rearrangements occurred along with WGD-2 in *S. rosmarinus* . The MRCA (most recent common ancestor) of*S. miltiorrhiza* and *S. splendens* could be considered as the ancestor of S. rosmarinus. Collinear homologous regions were identified by blastp in *S. miltiorrhiza* and *S. splendens* , based on karyotype projections results, protochromosomes were inferred in MRCA of *S. miltiorrhiza* and *S. splendens* . And then chromosomal rearrangements along with WGD-2 were inferred based on karyotype projectory in *S. rosmarinus* and collinear comparison between *S. rosmarinus* and *S. miltiorrhiza* . Finally, we inferred protochromosomes of the MRCA of *S. rosmarinus, S. miltiorrhiza* and *S. splendens* , and represented the process of chromosomes karyotype evolution in *S. rosmarinus, S. miltiorrhiza*and *S. splendens* .

Metabolic analysis

Tissues of rosemary leaves, stems and roots (three biological replicates) were used for RNA extraction and metabolite analyses. High performance liquid chromatography (Waters e2695) and high-performance gas chromatography (Agilent Technologies 7890B-5977B) were used for metabolite analysis of *S. rosmarinus* . The rosemary samples were oven-dried to a constant weight at 60°C, crushed with a grinder, and sieved through a 60-mesh screen. An aliquot of the test sample powder (0.02g) was added to 1 mL of 70% methanol, ultrasonically extracted for 45 minutes, centrifuged at 8,000 rpm/min for 10 minutes, and the supernatant

was collected and passed through a 0.22 μm filter membrane to obtain the test solution, each group of extracts (3 replicates) was placed in a 4 refrigerator for later use.

A Waters e2695 high performance liquid chromatography system was used for analysis, the detector model was a Waters2998 ultraviolet light detector, the chromatographic column model was a Waters sunfire C18 reversed-phase chromatographic column, and the chromatographic acquisition software was Empoeder 2. The chromatographic conditions were as follows: flow rate 1 mL/min, column temperature 30°C, sample loading volume 20 μL. The sample running time was 96 minutes, 0.02 percent phosphoric acid water and acetonitrile, utilizing gradient elution.

Qualitative and quantitative analysis of volatile components in different parts of the rosemary plant was carried out using GC-MS. Precisely weighed 1 g of fresh sample of rosemary was placed in a 20 mL headspace bottle, sealed, and then the sample was injected directly into the headspace sample tray. The chromatographic and headspace conditions were as follows: chromatography initial temperature 35°C, held for 2 min; 3°C/min to 130°C, 25°C/min to 250°C, held for 3 minutes; cooling to 35°C. Headspace conditions: Equilibrium: 80 for 30 min, oven temperature: 80, loop temperature: 90, transfer line temperature: 100, GC cycle time: 70 min, split ratio: 1:5, flow rate: 1mL/min, M/Z: 35-600, EI: 70eV, MS ion source temperature: 230degC, detector temperature: 260degC.

RNA sequencing and analyses

The RNAprep Pure Plant Plus Kit was used to extract total RNA from the roots, stems, and leaves of rosemary samples according to the manufacturer's instructions. The RNA-seq library was constructed on the Illumina HiSeq X Ten platform and sequenced to 150-bp paired reads, and then de novo assembled. The quality control of the assembled sequence included low-quality filtering, removal of N-containing bases, and removal of 3' and 5' end sequencing adapters. Raw data were filtered using SOAPnuke software (www.bgitechsolutions.com) with the following filtering parameters: -n 0.01 -l 20 -q 0.4 -A0.25 –cutAdaptor -Q 2 -G –polyX50 –minLen 150. After the raw data had been filtered, FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to monitor data quality. Three sets of biological replicates were made for each of the three parts of the roots, stems and leaves, and 9 sample RNA sequences were obtained for subsequent transcriptome analyses.

The 'WGCNA' package of R (v4.0) was used to separate all genes into modules of color blocks, calculate the PCCs between gene expression and the content of differential metabolites, and construct a weighted gene co-expression network based on PCCs. In addition, genes were regarded as highly corelated to candidate metabolites if PCCs were > 0.6 and p-values < 0.05.

Homology modeling and molecular docking

Homology modeling of *SmCYP76AK1* and SrCYP76AKs were performed by submitting the complete protein sequence to RoseTTAFold(Baek et al.), using the crystal structure of *SmCYP76AH1* (PDBid:5ylw) as a template, and evaluating the predicted protein model by calculating the root mean square deviation (RMSD) between the generated model and the standard structure. We used AutoDock (v4.2.6)(Bitencourt-Ferreira, Pintro, & de Azevedo, 2019) to dock the predicted protein model with the substrate, and visualized the docking results using PyMOL v2.5 (*https://www.pymol.org/*).

Results

Genome sequencing and assembly

DNA for sequencing the genome of *S. rosmarinus* was extracted from a single plant maintained at Zhejiang Sci-Tech University, and sequencing was performed using both Illumina and PacBio technologies. The initial genome assembly was generated using 107 Gb (84.92x) Illumina reads and 60.5 Gb (48.01x) PacBio reads, resulting in a genome size of 1263.45 Mb estimated through k-mer analysis (Figure S1, Tables S1, S2). After interactive error correction of PacBio reads and assembly of primary contigs using Canu (v2.1.1) and Falcon (v0.0.3) respectively, the genome was phased and polished using Pilon. After assembly of the PacBio

long reads and error correction with Illumina short reads, the final predicted genome size was 1239.46 Mb, with a scaffold N50 of 109.261 Kb (Table S3). The genome was further refined using an 88.10x coverage Hi-C library (111Gb), resulting in 19,878 scaffolds that were placed in 12 pseudochromosomes (ranging from 71.21 Mb to 144.54 Mb), and contained approximately 94.08% of the assembled sequences (Tables S1, S4). The final size of rosemary assembly reached 1,24 Gb, with scaffold N50 of 107.45 Mb. To assess the completeness of the assembly, RNA-seq reads from three different tissues were mapped to the genome, resulting in mapping rate of between 67.95% and 92.40% (Table S14). Additionally, BUSCO analysis (Simao et al., 2015) (Benchmarking Universal Single-copy Orthologs) showed a high level of completeness of both the genome assembly (96.8%) and annotation (88.5%) (Table S5 and S6), supporting the high quality of the *S. rosmarinus* genome assembly (Figure 1).

Genome Annotation

The draft genome was annotated using de novo predictions, homology-based predictions, and transcriptome data from RNA-seq of leaves, stems, and roots. In total, 61,716 genes were annotated, with 88.03% of them (54,326 genes) supported by transcriptome data (Table S9). The average gene length was 3,903 bp, with 5.57 exons per gene (Table S9). We submitted all gene models to five protein databases for annotation: NR (56,281, 92.74%), SwissProt (43,566, 71.79%), GO (24,080, 39.68%), KEGG (54,939, 90.53%), and Pfam (40,976, 67.52%). At least one database funtionally annotated 98.33% (60,687 genes) of the genes (Figure S4, Table S8). We named the genes according to the nomenclature used for *Arabidopsis* (*Arabidopsis* Genome Initiative, 2000) to indicate the relative positions of genes on the pseudochromosomes.

The draft rosemary genome contained 68.46% repetitive sequences, with 67.26% of the genome consisting interspersed repeats. Long terminal repeat (LTR) retroelements comprised 34.16% of the genome, with 24.28% LTR/*Gypsy* and 9.54% *Copia* elements being the predominant elements. DNA transposons accounted for 3.01% of the rosemary genome (Table S7). We detected noncoding RNAs (ncRNAs) using tRNAscan-SE and RNAmmer, which generated 413 microRNAs (miRNAs), 1,629 transfer RNAs (tRNAs), and 362 ribosomal RNAs (rRNAs) (Table S7). Figure 1 provides an overview of the genes, repeats, non-coding RNA densities, and all detected segmental duplications.

Rosemary-specific WGD-2 event lead to massive chromosome rearrangements

The phylogenetic position of S. rosmarinus.

We performed a comparative analysis of our assembly with 23 other genomes from twelve Lamiales (*Olea europaea* , *Boea hygrometrica* , *Antirrhinum majus* , *Sesamum indicum* ,*Handroanthus impetiginosus* , *Striga asiatica* ,*Erythranthe guttata* , *Tectona grandis* , *Scutellaria baicalensis* , *Ocimum tenuiflorum* , *Salvia miltiorrhiza* ,*S. splendens*), eight other eudicots (*Vitis vinifera* ,*Glycine max* , *Populus trichocarpa* , *Arabidopsis thaliana* , *Beta vulgaris* , *Daucus carota* , *Solanum lycopersicum*), two monocots (*Zea mays* , *Oryza sativa*), and*Amborella trichopoda* , which represents a species at the base of the angiosperm as a sister group to all other flowering plants (Table S10). We identified 38,709 gene families (consisting of 813,356 genes) by analyzing gene family clustering. Of these, 1,658 were specific to*S. rosmarinus* , while 5,256 were shared by all species, including 456 single-copy gene families (Figure S5, Table S18).

Compared to congenra *S. miltiorrhiza* and *S. splendens* ,*S. rosmarinus* displayed 5,695 expanded genes and 931 contracted genes, consistent with the previous findings (Bornowski et al., 2020). The majority of the expanded gene families in *S. rosmarinus* were associated with secondary metabolites, with a significant enrichment in "Biosynthesis of other secondary metabolites", involving 123 genes (Figure S11, Table S19). KEGG analysis also revealed erichment in pathways related to terpenoid metabolites, such as "Terpenoid backbone biosynthesis" and "Diterpene biosynthesis" (Table S19). Secondary metabolism-related genes, particularly those related to terpene metabolism, underwent significant expansion in *S. rosmarinus* , which likely contributed to the abundance of terpenoids in *S. rosmarinus* plants.

We retrieved 465 single-copy orthologous genes from 24 species, multi-aligned them, and produced a super-alignment matrix, which was used to construct a dated phylogeny. The topology and time frame in the

tree were consistent with previously reported phylogenomic analysis in angiosperms. The divergence of and within Lamiaceae (*S. rosmarinus* , *S. miltiorrhiza* , *S. splendens* ,*Scutellaria baicalensis* , *Ocimum tenuiflorum* and*Tectona grandis*) were estimated to be around 59.16 Mya and 52.10 Mya, respectively. The origin time of *Salvia rosmarinus* was estimated to be around 21.47 million years ago (Mya), with a separation into a *S. miltiorrhiza* and *S. splendens* clade (Figure 2a).

Whole genome duplication in S. romarinus.

According to both genomic covariance and paralogous homologous gene analysis, evidence supports the occurrence of ancient genome-wide duplication (WGD) events in *S. rosmarinus* . In particular, the identification of 125,489 homologous gene pairs in *S. rosmarinus*(Table S20) and the observation of two peaks in the distribution of substitutions per synonymous site (*Ks*), with *Ks* values of approximately 0.19 (WGD-2) and 0.94 (WGD-1) (Figure 2b, Figure S12) provides strong evidence for WGD events. However, there is a whole-genome triplication occurred with Ks peaks at 1.92 (WGT-γ) (Figure S12), which is not obvious with Ks ditribution for long time ago. Additionally, homology analysis showed that WGD-2 occurred after the divergence of *S. rosmarinus* , *S. miltiorrhiza* and *S. splendens* , as evidenced by the identificaiton of 17,521 orthologous gene pairs between *S. rosmarinus* and *S. miltiorrhiza* , and 16,664 orthologous gene pairs between *S. rosmarinus* and *S. splendens* , with *Ks* values peaking at approximately 0.20 and 0.28, respectively (Figure 2a, Table S20). Furthermore, paralogous gene analysis identified 67,287, and 26,794 paralogous gene paris in *S. miltiorrhiza* and *S. splendens* respectively, with *Ks*values distributions that peaked at approximately 0.98 and 1.24, respectively (Figure 2b, Table S20). Based on the phylogenetic analysis, WGD-1 occurred prior to the divergence of *S. rosmarinus* , *S. miltiorrhiza* and *S. splendens* and was estimated to have occurred between 70.87 and 101.37 Mya, which is consistent with the findings of the *Scutellaria baicalensis* genome study (Z. Xu et al., 2020). Whole-genome triplication (WGT-γ) had been reported shared in core eudicots, WGT-γoccurred at about 144.87-207.32 Mya according to the Ks distribution, close to the previous report(Murat, Armero, Pont, Klopp, & Salse, 2017).

The Ks distribution value of *S. rosmarinus* , which was found to be 0.18, indicated the occurrence of a whole genome duplication event (WGD-2) after the speciation of *S. rosmarinus* , approximately 8.80 Mya (Table S20). The genome syntenic analysis revealed that *S. rosmarinus* had four copies of syntenic blocks corresponding to*Vitis vinifera* blocks (Figure S10 b). This suggests that the entire genome of *S. rosmarinus* was duplicated twice during evolution, corresponding to WGD-1 and WGD-2 respectively. In addition, two copies of syntenic blocks from *S. rosmarinus* corresponding*S. miltiorrhiza* blocks were also found, indicating that the most recent genome-wide duplication of the *S. rosmarinus* genome occurred after the divergence of *S. rosmarinus* and *S. miltiorrhiza* (Figure S8). Therefore, WGD-1 was shared by *S. rosmarinus* , *S. miltiorrhiza* and *S. splendens* , while WGD-2 was unique to *S. rosmarinus* .

Deducing trajectories of S. rosmarinus.

Angiosperms have been proposed to derive from an ancestral eudicot karyotype (AEK) structured with seven protochromosomes. AEK experienced a known whole-genome triplication (WGT γ event) generating a 21-chromosome intermediate for the formation of the modern chromosomes of most eudicots (Bowers, Chapman, Rong, & Paterson, 2003; Jaillon et al., 2007). To infer the chromosome evolution of rosemary, we identified collinearity blocks across *S. rosmarinus* , *S. miltiorrhiza* ,*S. splendens* and *V. vinifera* genomes (Figure S8), with*V. vinifera* used as the reference and to represent the ancestral eudicot karyotype (AEK) genome due to its stable structure among core eudicots (Jaillon et al., 2007; Murat et al., 2017). *S. miltiorrhiza* and *S. splendens* have the closest relationships with rosemary among species with whole genome sequenced, therefore,*S. miltiorrhiza* and *S. splendens* were used for infering the evolutionary trajectory of rosemary chromosomes. The complexity of rosemary chromosome evolution was demonstrated by collinearity analysis, which revealed that the integrities of grape chromosomes were not preserved in rosemary, (Figure S8). To investigate the possible source of rosemary chromosomes, we analyzed the collinearity relationship between the genomes of rosemary and *S. miltiorrhiza.* Our analysis showed that the orthologous regions of *S. miltiorrhiza* chromosome Chr8 with rosemary genome were distributed on rosemary chromosomes Chr2, Chr3, Chr4 and Chr12, indicating that Sm8 was scattered in these rosemary chromosomes after the divergence of

rosemary and *S. miltiorrhiza* (Figure S8). The remaining main part of Chr2, Chr3, Chr4, Chr12 were merged from other *S. miltiorrhiza* chromosomes. Chr12 had orthologous regions with *S. miltiorrhiza* chromosome Chr2, Chr4 and Chr7, and the corresponding orthologous regions in grape genome were obtained by using the homologous relationship between *S. miltiorrhiza* and grape, then evolutionary trajectory of chromosome Chr12 was obtained. The proto-chromosomes of rosemary and *S. miltiorrhiza* were the orthologous regions shared between them (Figure 2d). We inferred the chromosomal evolution trajectories of rosemary and*S. miltiorrhiza* , and showed traces of 26 proto-chromosomes in extant chromosomes.

Monoterpenes and diterpenes are the major metabolites of S. rosmarinus

To investigate the variation in metabolites among different organs of*S. rosmarinus* , we collected and analyzed plant samples from roots, stems and leaves using HPLC-MS and GC-MS. Our analysis identified a total of 85 metabolites, including terpenoids, phenolic acids and flavonoids (Tables S11, S12). As expected, the essential oil extracted from rosemary leaves, which is widely used as a spice, has a distinctive aroma and was found to be rich in monoterpenes (Figure 3a). GC-MS analysis showed that 28 monoterpenes were identified in rosemary leaves, accounting for 99.43% of the volatile components (Table S11). These monoterpnenes included 1,8-cineole (16.66%), camphor (10.20%), limonene (7.57%) and α-pinene (7.55%), which dominated the volatile components in *S. rosmarinus* . Furthermore, LC-MS analysis revealed a high concentration of the diterpenoid carnosol, which accounted for 20.79% of the diterpenes and was found to be the main antioxidant component of *S. rosmarinus* (Table S12). Interestingly, the metabolites extracted from stems were similar to those from leaves in terms of chemical species and relative contents.

In constract, the metabolites identified in roots of *S. rosmarinus* were rather limited, with only 25 terpenoids and 6 phenolic acids (Figure S13). The main components extracted from roots were α-Cubebene (41.16%), camphene (11.13%), γ-muurolene (9.39%), as identified by GC-MS analysis. However, the quantity of monoterpenes and diterpenes extracted from roots was significantly lower compared to that extracted from leaves. For instance, 1,8-cineole, which had a high content in leaves, was present only in a low concentration of 3.04% in roots, and carnosol accounted for only 16.32% of the root components (Figure S13, Tables S11, S12).

Co-expression analysis screened out the set of terpene-related genes

To investigate the transcriptomic differences in leaves, stems, and roots of *S. rosmarinus* , three biological replicates were collected from each organ to ensure accuracy. Raw data underwent filtering, resulting in 7.02-7.10 million 150 bp paired-end reads (Table S13), which were mapped to our assembly at the rate of 67.95%– 92.40% (Table S14). Differentially expressed genes (DEGs) were identified using DESeq2 (Varet, Brillet-Gueguen, Coppee, & Dillies, 2016), with a stringent threshold of Log2|FoldChange| >1 and p-value < 0.05. We identified a total of 16,052 DEGs, with 8,833 up-regulated and 7,219 down-regulated genes in the root vs. Leaf comparison, 11,982 DEGs (6,496 up-regulated, 5,486 down-regulated) in the root vs. Stem comparison, and 15,598 DEGs (8,241 up-regulated, 7,357 down-regulated) in the leaf vs. Stem comparison (Figure S14). In addition, 3,475 genes were differentially expressed in all three groups (Figure 3e).

To gain insight into the metabolic processes involved in different organs of *S. rosmarinus* , KEGG enrichment analysis was performed. The DEGs between roots and leaves were significantly enriched in "Biosynthesis of other secondary metabolites" and "Metabolism of terpenoids and polyketides" (p-value< 0.05) (Figure S15, Table S15). While the DEGs between roots and stems were significantly enriched in "Metabolism of terpenoids and polyketides" and "Terpenoid backbone biosynthesis" (p-value< 0.05) (Figure S15, Table S15). Notably, the DEGs between stems and leaves were significantly enriched in "Biosynthesis of other secondary metabolites," "Flavonoid biosynthesis," and "Phenylpropanoid biosynthesis" (p-value < 0.05) (Figure S15, Table S16). These pathways provided a transcriptomic-level understanding of the metabolic processes in different organs of *S. rosmarinus* .

To construct the WGCNA co-expression network, we used all DEGs and differential metabolites. By calculating Pearson correlation coefficients (PCCs) between contents of components and gene expression modules (Figure S16, Tables S11, S12), we identified a set of genes associated with monoterpenoids and diterpenoids

(PCCs > 0.6). Our results suggested that *SrCYP71D8* , *SrWRKY2* (with p-value of 0.0315 and 0.0237, respectively) were highly correlated with the accumulation of monoterpene in *S. rosmarinus* organs. We also screened key genes, including *SrHMGR* (with p-value located at 0.0019– 0.0401) and limonene synthase, which was found to be involved in the biosynthesis process of monoterpene. Moreover, we identified *SrCYP81Q32* , and *SrMYB1* (with p-values of 0.0004 and 0.0004, respectively) as highly correlated with the accumulation of diterpenoids. Finally, we found that *SrGGPPS* (with p-values of 0.0004– 0.0475) was highly correlated with diterpenoids biosynthesis (Table S21). These findings provide important evidence to support further exploration of terpenoid biosynthesis in *S. rosmarinus* .

Evolution of primary monoterpene biosynthesis genes in rosemary

Expanding, clustering and high expression of HMGR in S. rosmarinus leaf tissue

To investigate the genetic basis of monoterpene accumulation, we analyzed the first key enzyme in the monoterpene synthesis pathway, 3-hydroxy-3-methylglutaryl-CoA reductase (*HMGR*). Our gene family contraction and expansion analysis showed that the family of genes encoding *SrHMGR* was expanded in *S. rosmarinus* compared to sister clades, with 12 copies in *S. rosmarinus* and only 6 in*S. miltiorrhiza* (Table S22). To further understand the expansion of genes encoding *SrHMGR* , we constructed a maximum likelihood (ML) tree of *HMGRs* from 24 species and found that *HMGRs*could be divided into three subgroups (Figure S18). *SrHMGR7* ,*SrHMGR9* , *SrHMGR10* and *SrHMGR12* were grouped in the same sub-clade with high amino acid sequence identity (88.87%–97.96%) (Table S24). These genes clustered within 0.19 Mb on pseudochromosome 7 (Table S25),. indicating *SrHMGR* genes on pseudochromosome 7 had expanded and replicated in clusters in *S. rosmarinus.* Covariance analysis of *HMGRs* in *S. rosmarinus* , *S. miltiorrhiza* , *S. baicalensis* , and *S. splendens* confirmed that the expanded *SrHMGR* genes on pseudochromosome 7 may be an important genetic basis for monoterpene accumulation (Figure S17). Additionally, transcriptome data showed that *SrHMGR7* ,*SrHMGR9* , *SrHMGR10* , and *SrHMGR12* were expressed at 2.40-fold higher levels overall in leaves than in roots of *S. rosmarinus* (Figure 4), suggesting that the upregulation of *HMGR*expression in the MVA pathway may facilitate the biosynthesis of monoterpenes.

Clustering and high expression of 1,8-cineole synthase genes in S. rosmarinus leaf tissue

The co-expression results showed that limonene synthases in *S. rosmarinus* were highly correlated with the synthesis of monoterpenes (p-value = 0.0102, Table S21). A total of three limonene synthases and three 1,8-cineole synthases were identified in *S. rosmarinus* . All genes encoding limonene synthases (*SrLS-1* , *SrLS-2* ,*SrLS-3*) and 1,8-cineole synthases (*SrCinS-3* ,*SrCinS-4*) are clustered on pseudochromosome 3, within a 1.1 Mb region (Figure 5b, Table S25). Transcriptome analysis revealed that the expression levels of the three limonene synthases (*SrLS-1* ,*SrLS-2* , *SrLS-3*) were 8.44, 9.80 and 6.67 times higher in leaves than in roots (Figure 5c), respectively, while the expression levels of the two 1,8-cineole synthases (*SrCinS-3* ,*SrCinS-4*) were 5.57 and 8.81 times higher in leaves than in roots (Figure 5c). Additionally, covariance analysis was performed on *S. rosmarinus* , *S. miltiorrhiza* and *S. splendens* , revealing that the genes encoding 1,8-cineole synthase on pseudochromosome 2 did not have homologues in *S. miltiorrhiza* and *S. splendens.*

We constructed an evolutionary tree of 24 species to analyze limonene synthases and 1,8-cineole synthases, which revealed that *SrCinSs*and *SrLSs* formed two clades with 40.47%–99.16% sequence identity ((Figure S19, Table S29). 3D models of *SrCinSs* and*SrLSs* (pdbid: 2ong) were generated and validated using Ramachandran plots (Figure S21). The models of both enzymes were highly similar in stereospecificity, as indicated by the average root means square displacements (RMSDs) (0.38Å–1.03Å) between the predicted models (Figure S22, S21a). Docking studies showed that. intermediate terpinol cations are located near the active pockets of *SrCinS-3* and*SrCinS-4* (Figure 5d). And eight amino acid residues of the active pocket (Cys-250, Trp-253, Asn-274, Thr-278, Met-458, His-502, Tyr-496 and Ser-454) that all lie within 10 Å distance of the docking site and have a direct effect on biosynthesis of 1,8-cineole were examined (Kampranis et al., 2007; Piechulla et al., 2016; N. Srividya et al., 2015; J. Xu et al., 2017). Cys-379, Trp-382, Tyr-626, and Thr-278 maintained their original polarity in *SrCinS-3* and*SrCinS-4.* Mutations S512G and A278T in silico analysis indicated that these residues retained their original polarity in *SrCinS-3*and *SrCinS-4* . However,

A278T became more hydrophilic and M622I lost its original polarity (Figure S24, Table S27). Based on these findings, we hypothesize that SrCinS-3 and SrCinS-4 are responsible for 1,8-cineole biosynthesis. Furthermore, we found that S512G and A278T, which promote the accumulation of 1,8-cineole in tobacco terpene synthases (Piechulla et al., 2016), were retained in SrCinS-3 and SrCinS-4 in rosemary, their association with high 1,8-cineole accumulation in rosemary.

The limonene synthases in *S. rosmarinus* were observed to dock with terpinyl cations near the active pocket (Figure S23), indicating that the crystallographic structures of limonene synthase could accept terpinyl cation intermediate. Mutations in the key sites of the active pocket can affect the product diversity of terpene synthases as the product profile is determined by the conformation of the substrate or intermediate. Analysis of the active pockets of *SrLSs* showed that they deviate from the ancestral limonene synthase pattern (N. Srividya et al., 2015; Narayanan Srividya, Lange, & Lange, 2020), with changes in polarity observed for several important sites (Thr-278, His-502, Tyr-496 and Ser-454, see Table S27). M519I and T279V mutations directly led the loss of polarity of original residues, but compensatory mutations were observed for Asn-274 (to Phe-274) and Cys-250 (to Asn-250), which maintained the polarity of the active site (J. Xu et al., 2017) (Figure S23, Table S27). The polarity changes in the active sites resulted in a larger active pocket, potentially enhancing the production of more abundant terpenoids by attenuating the stability of the carbon positive ion.

The transcriptional expression of genes started from the specific binding of the promoter region upstream of the gene to RNA polymerase; therefore, we extracted the promoter sequences of *SrCinSs* and*SrLSs* and analyzed the promoter elements using Promoter 2.0 (Knudsen, 1999). TATA-boxes are essential for binding to RNA polymerase and activating transcription among the various elements (Orphanides, Lagrange, & Reinberg, 1996). The results showed the number of TATA-boxes nearly doubled in the promoter sequence of *SrCinS-3*(Table S28), greatly enhancing its expression and promoting the accumulation of 1,8-cineole. Additionally, we examined G-box elements in the promoter region of *SrCinSs,* and five G-boxes were targeted on*SrCinS-3* , significantly increasing its binding probability with the *bHLH* gene family. Furthermore, we discovered two*SrbHLH143* near the gene cluster (Figure 5b), and their expression was 2.24-fold and 1.63-fold higher in leaves than in roots (Figure 5c), leading to a speculation that *SrbHLH143* may play a regulatory role in the biosynthesis of 1,8-cineole. The expression of the bHLH family was reported to be significantly correlated with changes in 1,8-cineole content in *Artemisia absinthium* (Yi et al., 2021), and G-box was cis-acting DNA regulatory element which could bind with bHLH transcription factor(Qian et al., 2007). Conversely, the promoter region of other 1,8-cineole synthases had incomplete G-box elements, resulting in low expression.

Evolution of carnosic acid biosynthesis in S. rosemarinus.

Diterpene gene cluster involved in abietane-type diterpenoids in S. rosmarinus

Comparisons of microsynteny blocks detected in *S. rosmarinus* ,*S. miltiorrhiza* and *S. splendens* , revealed the presence of two diterpene gene clusters (DGC) in *S. rosmarinus* distributed in Chr1 and Chr2, spanning 220 Kb and 200 Kb, respectively (Figure S27). Previous studies have reported the loss of shoot *KSL* and three *CYP76AH* genes (*CYP76AH59* , *CYP76AH58* , and*CYP76AH56*) in *S. militiorrhiza* DGC, and silencing or inactivity of *SmCPS2* in tanshinone biosynthesis, leading to the abrogation of abietane-type diterpenoid biosynthesis in shoots (C.-Y. Li et al., 2022). Despite the loss of *CYP76AH59* on rosemary chromosomes, *SrCYP76AH58* , *SrCYP76AH56* , *SrCPS* , and*SrKSL* were identified in the rosemary DGC, providing precursors for abietane-type diterpenoid biosynthesis. *CPS2* had been silenced in DGCS of *S. miltiorrhiza* and rosemary, however,*SrCPS1* on Chr1 and Chr2 highly expressed in rosemary leaves (4.03-fold and 6.42-fold), and remained active in diterpene biosynthesis. HPLC-MS results also revealed that the accumulation of carnosic acid and carnosolin rosemary leaves. These findings demonstrate that the, two DGCs in rosemary retain the ability of produce abietane-type diterpenoids in its leaves.

WGD-2 causes duplication of SrCYP76AK6–8 genes and site-specific mutations in molecular docking sites

Carnosic acid and carnosol are the primary diterpenes in *S. rosmarinus* leaves, the biosynthesis of them have

been elaborated. These compounds are derived from precursors (IPP and DMAPP) through MEP pathway in the plastids, and are catalyzed by downstream genes including diterpene synthases and cytochrome P450. In *S. rosmarinus* genome, we identified three genes encoding *SrCYP76AK6* , two encoding*SrCYP76AK7* , and two encoding *SrCYP76AK5* on pseudochromosome 11. All of these genes were clustered within a 0.33 Mb region (Figure 6d), and one, four, and one homologous gene were identified in the syntenic positions in *S. miltiorrhiza* , *S. splendens* , and *S. baicalensis* , respectively (Figure 5e). These findings suggest that substantial duplication of *SrCYP76AK5, SrCYP76AK6* and *SrCYP76AK7* occurred on pseudochromosome 11 after speciation of *S. rosmarinus.* Moreover, *SrCYP76AK5-2* ,*SrCYP76AK6-1* , and *SrCYP76AK6-2* are highly expressed in rosemary leaves, with expression levels 6.59-fold, 5.64-fold, and 6.25-fold higher than in roots, respectively (Figure 7d). Therefore, the clustering, expansion, and high expression of the genes encoding*SrCYP76AK5* , *SrCYP76AK6* and *SrCYP76AK7* might have contributed to the accumulation of carnosol in *S. rosemarinus* .

In addition to the *SrCYP76AK* genes on pseudochromosome 11, we have also identified one *SrCYP76AK5* gene and one*SrCYP76AK8* gene on pseudochromosome 3. Our analysis of the evolutionary trajectory for the chromosomes of *S. rosemarinus*suggested that the duplication of *CYP76AK8* occurred as result of the WGD-2 and subsequently underwent chromosomal rearrangements and fusions on pseudochromosomes 3 and 11, respectively (Figure S29 b). The Ks values between homologous gene pairs (*SrCYP76AK5-1* vs*SrCYP76AK5-2* and *SrCYP76AK5-1* vs *SrCYP76AK6-2)* were all close to *Ks* value of WGD-2, indicating their duplication occurred during this event, and then *SrCYP76AK6-2* replicated to*SrCYP76AK7-2* . The duplication of *SrCYP76AK7-1* to*SrCYP76AK7-2* and *SrCYP76AK6-1* to *SrCYP76AK6-2*occurred close to the present (Table S31). We hypothesize that*SrCYP76AK5-1* and *SrCYP76AK8-1* on Chr3 were copied to Chr11 during the event of WGD-2, following a tandem duplication occurred recently on Chr11. It was followed by replications of*SrCYP76AK6-2* and *SrCYP76AK6-3* to form the cluster of six*SrCYP76AK* copies. Moreover, further duplications of chromosome fragments led to the clustering of six *SrCYP76AK6–8* genes on pseudochromosome 11 within a 0.33 Mb region (Figure 7e). This is supported by our phylogenetic analysis of the proteins encoded by homologous gene pairs (Figure S30), and *Ks* calculations (Table S29).

To gain a comprehensive understanding of the evolution of *CYP76AK*subfamily, we examined the proteins encoded by *CYP76AK1* and*CYP76AK6-8* in 24 different species and extracted a total of 18 protein sequences, mainly from *Salvia* species. Using *Ocimum basilicum CYP76* gene as an outgroup, a maximum likelihood (ML) tree of CYP76 genes were reconstructed (Figure S29 a). The phylogenetic relationships revealed that the proteins encoded by *CYP76AK1s* ,*CYP76AK2s* , *CYP76AK3s* , *CYP76AK5s* , *CYP76AK6s* ,*CYP76AK7s* and *CYP76AK8s* align into four distinct groups, respectively. The evolutionary tree of the *CYP76AK* subfamily showed two clades, the clade of the gene encoding the *CYP76AK3*and *CYP76AK7* was sister to the clade of *CYP76AK5* ,*CYP76AK6* , *CYP76AK8* , *CYP76AK1* and *CYP76AK2. CYP76AK6, CYP76AK1* and *CYP76AK2* did not form the independent clade, which indicated that CYP76AK6, *CYP76AK1* and*CYP76AK2* were evolved from *CYP76AK8, CYP76AK1* and*CYP76AK2* were evolved from *CYP76AK6*

We observed that *SrCYP76AK6–8* catalyzed the conversion of 11-hydroxy ferruginol into capraldehyde in *S. rosmarinus* , while*SmCYP76AK1* catalyzed the production of 11,20-dihydroxy ferruginol in *S. miltiorrhiza* (Figure 7a). To further investigate the catalytic mechanism of *CYP76AK* subfamily, we performed homology modeling and molecular docking to infer the key amino acid sites on*SmCYP76AK1* and *SrCYP76AKs* . The latter were highly expressed in leaves of rosemary. Using SmCYP76AH1 (PDBid: 5ym3) structure as a PDB template, we generated 3D models of SmCYP76AK1 and SrCYP76AKs, and docked them to the substrate 11-hydroxy-ferroginol. Our results showed that position C-20 in 11-hydroxy-ferruginol, which docked with *SrCYP76AK5-2* , *SrCYP76AK6-1* , and *SrCYP76AK6-2* , was closer to heme iron than that with *SmCYP76AK1* (Figure 6b). This closer proximity may have led to a sequential oxidation reaction at C-20 that resulted in the accumulation of carnosol precursors. We hypothesized that mutations in essential amino acids could result in functional differentiation of *CYP76AKs* , leading to the accumulation of carnosol in the leaves of *S. rosmarinus* and tanshinone in the roots of *S. miltiorrhiza* , respectively.

Furthermore, we investigated amino acid mutations within 8 Å of the active pocket in order to understand their potential influence on the proximity of the ligands to the heme iron (Figure 7c). To identify key residues involved in docking sites, we compared differential amino acid residues within this range between *SrCYP76AKs* and*SmCYP76AK1* , and found. nine candidate residues. We then conducted remodeling and docking experiments by replacing the corresponding residues of *SmCYP76AKs* with those of *SmCYP76AK1* , and vice versa. Specifically, we mutated S445 and I449 of *SrCYP76AK5-2* ,*SrCYP76AK6-1* , and *SrCYP76AK6-2* to I445 and M449, respectively, to mimic *SmCYP76AK1* . Conversely, we mutated I445 and M449 of SmCYP76AK1 to S445 and I449, respectively, to mimic SrCYP76AK5-2, SrCYP76AK6-1, and SrCYP76AK6-2. We then docked these remodeled proteins with 11-hydroxy-ferruginol. whereas I445S, M449I with*SmCYP76AK1* . The results showed that the co-mutation of S445I and I449M in SrCYP76AK5-2, SrCYP76AK6-1, SrCYP76AK6-2 led to ligands docking away from heme iron at the docking sites, while co-mutation of I445S and M449 in SmCYP76AK1 resulted in docking close to heme iron (Figure S31). Therefore, we hypothesized that S445I and I449M played a significant role in determining the distance of ligand from heme iron, and may have contributed to the functional divergence of *SmCYP76AK1* from*SrCYP76AK6–8* . Our findings suggest that these residues are critical for ligand binding and may have important implications for understanding the functional differences between these two enzymes.

Discussion

The expansion of a large number of duplicated genes and chromosomal rearrangements, which may contribute to the increase of metabolite content, is frequently accompanied by polyploidy, which is regarded as a key factor in species divergence (Ren et al., 2018; Van de Peer, Mizrachi, & Marchal, 2017), Polyploidy events were prevalent in dicotyledons and contributed to the accumulation of metabolites in plants. The polyploidization event in *Hippophae rhamnoides* may have contributed to the increased accumulation of fatty acid synthesis, AsA, and aldonic acid in its fruits (L. Yu et al., 2022). The positive effect of gene amplification on metabolite accumulation was discovered in the report of *Artemisia argyi* genome, which demonstrated its adaptability in the face of environmental stress (Miao et al., 2022). Our study uncovered that *S. rosmarinus* underwent an independent polyploidization event approximately 8.8 million years ago. Through analyzing the replication events of genes in the biosynthetic pathways of monoterpenes and diterpenes, we observed traces of a second WGD, suggesting that WGD-2 contributed to the expansion of genes involved in terpene biosynthesis, resulting in the mass accumulation of terpenoids in rosemary leaves.

Carnosol, the major active phenolic diterpene in *S. rosmarinus* , is found in various Mediterranean plants such as *S. officinalis* ,*Thymus mongolicus* and *Origanum vulgare* . Rosemary extracts was a source of high antioxidant compounds (Petiwala & Johnson, 2015). The biosynthetic pathways of diterpenoids in *S. rosmarinus* and*S. miltiorrhiza* diverged at the step of catalyzing the formation of 11-hydroxy-ferruginol. In *S. rosmarinus* , SrCYP76AK6–8 introduces a carbonyl group at C-20 to produce carnosaldehyde, which is further converted to carnosic acid and its derivatives (Ignea et al., 2016; Scheler et al., 2016). In *S. miltiorrhiza,SmCYP76AK1* encodes an enzyme with hydroxylation activity at C-20 that primarily produces 11,20-dihydroxy-ferruginol for further biosynthesis of tanshinone using 11-hydroxy-ferruginol as a substrate (Guo et al., 2016). CYP76AK1 and CYP76AK6–8 were key enzymes in the same protein subfamily controlling the biosynthesis of diterpenoids in*S. rosmarinus* and *S. miltiorrhiza* , repectively, as a result of the evolutionary divergence of the ancestral *CYP76AK* gene (Bathe et al., 2019). Intrerestingly, *CYP76AK1* was not identified in *S. rosmarinus* , and *CYP76AK6–8* were not identified in *S. miltiorrhiza* . The results of the genealogical evolution of the genus *Salvia* indicated that the *Salvia* in Europe differentiated earlier than *Salvia* in Eastern Asia (Hu et al., 2018). *CYP76AK1* and CYP76AK2 evolved from *CYP76AK6* and *CYP76AK8*. The evolutionary direction of *CYP76AK6* and*CYP76AK8* to *CYP76AK1* and *CYP76AK2* was consistent with the genealogical evolution of Europe *Salvia* to Eastern Asia*Salvia* , which laid the genetic basis for structural differences in diterpenes.

Conclusion

The study assembled a chromosome-level genome of rosemary, demonstrating a high level of genomic integrity in comparison to previously reported genomes (Bornowski et al., 2020). Notably, *S. rosmarinus* differs

significantly from other *Salvia* species in both of plant morphology and secondary metabolites. To investigate the genetic basis for these differences, our study provides a genomic resource for exploring the structural and genetic basis of monoterpene and diterpenoid accumulation in *S. rosmarinus* , which will facilitate the development of molecular breeding and quality improvement of this species.

Acknowledgements

The authors declare no competing interests.

Funding

This work was financially supported by Zhejiang Provincial Natural Science Foundation of China (LR21H280002); Key Scientific and Technological Grant of Zhejiang for Breeding New Agricultural Varieties (No.2021C02074); National Natural Science Foundation of China for State Key Laboratory (81973415) and Key project of the Central Government: Capacity Building of sustainable Utilization of Traditional Chinese Medicine Resources (2060302).

Data availability

All data sets (genome sequencing, genome assembly, RNA-seq and metabolic profiles) are depositing into the required databases. All data with the article can be requested by contacting Dong-feng Yang (ydf807@sina.com).

References

al-Sereiti, M. R., Abu-Amer, K. M., & Sen, P. (1999). Pharmacology of rosemary (Rosmarinus officinalis Linn.) and its therapeutic potentials.*Indian J Exp Biol, 37* (2), 124-130.

Allegra, A., Tonacci, A., Pioggia, G., Musolino, C., & Gangemi, S. (2020). Anticancer Activity of Rosmarinus officinalis L.: Mechanisms of Action and Therapeutic Potentials. *Nutrients, 12* (6). doi:10.3390/nu12061739

Alsamri, H., Hasasna, H. E., Baby, B., Alneyadi, A., Dhaheri, Y. A., Ayoub, M. A., . . . Iratni, R. (2021). Carnosol Is a Novel Inhibitor of p300 Acetyltransferase in Breast Cancer. *Front Oncol, 11* , 664403. doi:10.3389/fonc.2021.664403

Angioni, A., Barra, A., Cereti, E., Barile, D., Coïsson, J. D., Arlorio, M., . . . Cabras, P. (2004). chemical composition, plant genetic differences, antimicrobial and antifungal activity investigation of the essential oil of Rosmarinus officinalis L. *J Agric Food Chem, 52* (11), 3530-3535. doi:10.1021/jf049913t

Baek, M. A.-O., DiMaio, F. A.-O., Anishchenko, I. A.-O., Dauparas, J. A.-O. X., Ovchinnikov, S. A.-O., Lee, G. A.-O., . . . Baker, D. A.-O. Accurate prediction of protein structures and interactions using a three-track neural network. (1095-9203 (Electronic)).

Bao, T. Q., Li, Y., Qu, C., Zheng, Z. G., Yang, H., & Li, P. (2020). Antidiabetic Effects and Mechanisms of Rosemary (Rosmarinus officinalis L.) and its Phenolic Components. *Am J Chin Med, 48* (6), 1353-1368. doi:10.1142/s0192415x20500664

Bathe, U., Frolov, A., Porzel, A., & Tissier, A. (2019). CYP76 Oxidation Network of Abietane Diterpenes in Lamiaceae Reconstituted in Yeast. *J Agric Food Chem, 67* (49), 13437-13450. doi:10.1021/acs.jafc.9b00714

Bergman, M. E., Davis, B., & Phillips, M. A. (2019). Medically Useful Plant Terpenoids: Biosynthesis, Occurrence, and Mechanism of Action.*Molecules, 24* (21). doi:10.3390/molecules24213961

Bitencourt-Ferreira, G., Pintro, V. O., & de Azevedo, W. F., Jr. (2019). Docking with AutoDock4. *Methods Mol Biol, 2053* , 125-148. doi:10.1007/978-1-4939-9752-7_9

Bornowski, N., Hamilton, J. P., Liao, P., Wood, J. C., Dudareva, N., & Buell, C. R. (2020). Genome sequencing of four culinary herbs reveals terpenoid genes underlying chemodiversity in the Nepetoideae. *DNA Res, 27* (3). doi:10.1093/dnares/dsaa016

Bowers, J. E., Chapman, B. A., Rong, J., & Paterson, A. H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature, 422* (6930), 433-438. doi:10.1038/nature01521

Chin, C. S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., . . . Schatz, M. C. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods, 13* (12), 1050-1054. doi:10.1038/nmeth.4035

Christopoulou, S. D., Androutsopoulou, C., Hahalis, P., Kotsalou, C., Vantarakis, A., & Lamari, F. N. (2021). Rosemary Extract and Essential Oil as Drink Ingredients: An Evaluation of Their Chemical Composition, Genotoxicity, Antimicrobial, Antiviral, and Antioxidant Properties.*Foods, 10* (12). doi:10.3390/foods10123143

Degner, S. C., Papoutsis, A. J., & Romagnolo, D. F. (2009). Chapter 26 - Health Benefits of Traditional Culinary and Medicinal Mediterranean Plants. In R. R. Watson (Ed.), *Complementary and Alternative Therapies and the Aging Population* (pp. 541-562). San Diego: Academic Press.

Flamini, G., Najar, B., Leonardi, M., Ambryszewska, K. E., Cioni, P. L., Parri, F., . . . Pistelli, L. (2022). Essential oil composition of Salvia rosmarinus Spenn. wild samples collected from six sites and different seasonal periods in Elba Island (Tuscan Archipelago, Italy).*Nat Prod Res, 36* (7), 1919-1925. doi:10.1080/14786419.2020.1824229

Forestier, E. C. F., Brown, G. D., Harvey, D., Larson, T. R., & Graham, I. A. (2021). Engineering Production of a Novel Diterpene Synthase Precursor in Nicotiana benthamiana. *Front Plant Sci, 12* , 757186. doi:10.3389/fpls.2021.757186

Freedman, P. (2019). History of Spices. In H. L. Meiselman (Ed.),*Handbook of Eating and Drinking: Interdisciplinary Perspectives*(pp. 1-15). Cham: Springer International Publishing.

Gao, Y., Honzatko, R. B., & Peters, R. J. (2012). Terpenoid synthase structures: a so far incomplete view of complex catalysis. *Nat Prod Rep, 29* (10), 1153-1175. doi:10.1039/c2np20059g

Guo, J., Ma, X., Cai, Y., Ma, Y., Zhan, Z., Zhou, Y. J., . . . Huang, L. (2016). Cytochrome P450 promiscuity leads to a bifurcating biosynthetic pathway for tanshinones. *New Phytol, 210* (2), 525-534. doi:10.1111/nph.13790

Han, D., Li, W., Hou, Z., Lin, C., Xie, Y., Zhou, X., . . . Yang, C. (2023). The chromosome-scale assembly of the Salvia rosmarinus genome provides insight into carnosic acid biosynthesis. *Plant J, 113* (4), 819-832. doi:10.1111/tpj.16087

Hu, G. X., Takano, A., Drew, B. T., Liu, E. D., Soltis, D. E., Soltis, P. S., . . . Xiang, C. L. (2018). Phylogeny and staminal evolution of Salvia (Lamiaceae, Nepetoideae) in East Asia. *Ann Bot, 122* (4), 649-668. doi:10.1093/aob/mcy104

Ignea, C., Athanasakoglou, A., Ioannou, E., Georgantea, P., Trikka, F. A., Loupassaki, S., . . . Kampranis, S. C. (2016). Carnosic acid biosynthesis elucidated by a synthetic biology platform. *Proc Natl Acad Sci U S A, 113* (13), 3681-3686. doi:10.1073/pnas.1523787113

Jaillon, O., Aury, J. M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., . . . Wincker, P. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.*Nature, 449* (7161), 463-467. doi:10.1038/nature06148

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., . . . Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics, 30* (9), 1236-1240. doi:10.1093/bioinformatics/btu031

Kampranis, S. C., Ioannidis, D., Purvis, A., Mahrez, W., Ninga, E., Katerelos, N. A., . . . Johnson, C. B. (2007). Rational conversion of substrate and product specificity in a Salvia monoterpene synt-

hase: structural insights into the evolution of terpene synthase function. *Plant Cell, 19* (6), 1994-2005. doi:10.1105/tpc.106.047779

Knudsen, S. (1999). Promoter2.0: for the recognition of PolII promoter sequences. *Bioinformatics, 15* (5), 356-361. doi:10.1093/bioinformatics/15.5.356

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res, 27* (5), 722-736. doi:10.1101/gr.215087.116

Kronenberg, Z. N., Rhie, A., Koren, S., Concepcion, G. T., Peluso, P., Munson, K. M., . . . Kingan, S. B. (2021). Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. *Nat Commun, 12* (1), 1935. doi:10.1038/s41467-020-20536-y

Li, C.-Y., Yang, L., Liu, Y., Xu, Z.-G., Gao, J., Huang, Y.-B., . . . Chen, X.-Y. (2022). The sage genome provides insight into the evolutionary dynamics of diterpene biosynthesis gene cluster in plants. *Cell Reports, 40* (7), 111236. doi:*https://doi.org/10.1016/j.celrep.2022.111236*

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics, 26* (5), 589-595. doi:10.1093/bioinformatics/btp698

Maurizio, M., Francesconi, A., Perinu, B., & Vais, E. (2002). Selection of Rosemary (Rosmarinus officinalis L.) Cultivars to Optimize Biomass Yield. *Journal of Herbs, Spices & Medicinal Plants* , 133-138. doi:10.1300/J044v09n02_19

Mena, P., Cirlini, M., Tassotti, M., Herrlinger, K. A., Dall'Asta, C., & Del Rio, D. (2016). Phytochemical Profiling of Flavonoids, Phenolic Acids, Terpenoids, and Volatile Fraction of a Rosemary (Rosmarinus officinalis L.) Extract. *Molecules, 21* (11). doi:10.3390/molecules21111576

Mendes, F. K., Vanderpool, D., Fulton, B., & Hahn, M. W. (2020). CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* . doi:10.1093/bioinformatics/btaa1022

Mendoza-Poudereux, I., Kutzner, E., Huber, C., Segura, J., Eisenreich, W., & Arrillaga, I. (2015). Metabolic cross-talk between pathways of terpenoid backbone biosynthesis in spike lavender. *Plant Physiol Biochem, 95* , 113-120. doi:10.1016/j.plaphy.2015.07.029

Miao, Y., Luo, D., Zhao, T., Du, H., Liu, Z., Xu, Z., . . . Huang, L. (2022). Genome sequencing reveals chromosome fusion and extensive expansion of genes related to secondary metabolism in Artemisia argyi. *Plant Biotechnol J* . doi:10.1111/pbi.13870

Micić, D., Đurović, S., Riabov, P., Tomić, A., Šovljanski, O., Filip, S., . . . Blagojević, S. (2021). Rosemary Essential Oils as a Promising Source of Bioactive Compounds: Chemical Composition, Thermal Properties, Biological Activity, and Gastronomical Perspectives. *Foods, 10* (11). doi:10.3390/foods10112734

Moss, M., Cook, J., Wesnes, K., & Duckett, P. (2003). Aromas of rosemary and lavender essential oils differentially affect cognition and mood in healthy adults. *Int J Neurosci, 113* (1), 15-38. doi:10.1080/00207450390161903

Murat, F., Armero, A., Pont, C., Klopp, C., & Salse, J. (2017). Reconstructing the genome of the most recent common ancestor of flowering plants. *Nat Genet, 49* (4), 490-496. doi:10.1038/ng.3813

Nematolahi, P., Mehrabani, M., Karami-Mohajeri, S., & Dabaghzadeh, F. (2018). Effects of Rosmarinus officinalis L. on memory performance, anxiety, depression, and sleep quality in university students: A randomized clinical trial. *Complement Ther Clin Pract, 30* , 24-28. doi:10.1016/j.ctcp.2017.11.004

Neves, J. A., Neves, J. A., & Oliveira, R. C. M. (2018). Pharmacological and biotechnological advances with Rosmarinus officinalis L. *Expert Opin Ther Pat, 28* (5), 399-413. doi:10.1080/13543776.2018.1459570

Ngo, S. N., Williams, D. B., & Head, R. J. (2011). Rosemary and cancer prevention: preclinical perspectives. *Crit Rev Food Sci Nutr, 51* (10), 946-954. doi:10.1080/10408398.2010.490883

Ojeda-Sana, A. M., van Baren, C. M., Elechosa, M. A., Juarez, M. A., & Moreno, S. (2013). New insights into antibacterial and antioxidant activities of rosemary essential oils and their main components.*FOOD CONTROL, 31* (1), 189-195. doi:10.1016/j.foodcont.2012.09.022

Orphanides, G., Lagrange, T., & Reinberg, D. (1996). The general transcription factors of RNA polymerase II. *Genes Dev, 10* (21), 2657-2683. doi:10.1101/gad.10.21.2657

Petiwala, S. M., & Johnson, J. J. (2015). Diterpenes from rosemary (Rosmarinus officinalis): Defining their potential for anti-cancer activity. *Cancer Letters, 367* (2), 93-102. doi:*https://doi.org/10.1016/j.canlet.2015.07.005*

Piechulla, B., Bartelt, R., Brosemann, A., Effmert, U., Bouwmeester, H., Hippauf, F., & Brandt, W. (2016). The α-Terpineol to 1,8-Cineole Cyclization Reaction of Tobacco Terpene Synthases. *Plant Physiol, 172* (4), 2120-2131. doi:10.1104/pp.16.01378

Pokajewicz, K., Białoń, M., Svydenko, L., Fedin, R., & Hudz, N. (2021). Chemical Composition of the Essential Oil of the New Cultivars of Lavandula angustifolia Mill. Bred in Ukraine. *Molecules, 26* (18). doi:10.3390/molecules26185681

Qian, W., Tan, G., Liu, H., He, S., Gao, Y., & An, C. (2007). Identification of a bHLH-type G-box binding factor and its regulation activity with G-box and Box I elements of the PsCHS1 promoter.*Plant Cell Rep, 26* (1), 85-93. doi:10.1007/s00299-006-0202-x

Raal, A., Orav, A., & Arak, E. (2007). Composition of the essential oil of Salvia officinalis L. from various European countries. *Nat Prod Res, 21* (5), 406-411. doi:10.1080/14786410500528478

Rašković, A., Milanović, I., Pavlović, N., Ćebović, T., Vukmirović, S., & Mikov, M. (2014). Antioxidant activity of rosemary (Rosmarinus officinalis L.) essential oil and its hepatoprotective potential.*BMC Complement Altern Med, 14* , 225. doi:10.1186/1472-6882-14-225

Ren, R., Wang, H., Guo, C., Zhang, N., Zeng, L., Chen, Y., . . . Qi, J. (2018). Widespread Whole Genome Duplications Contribute to Genome Complexity and Species Diversity in Angiosperms. *Mol Plant, 11* (3), 414-428. doi:10.1016/j.molp.2018.01.002

Scheler, U., Brandt, W., Porzel, A., Rothe, K., Manzano, D., Božić, D., . . . Tissier, A. (2016). Elucidation of the biosynthesis of carnosic acid and its reconstitution in yeast. *Nat Commun, 7* , 12942. doi:10.1038/ncomms12942

Senthoorraja, R., Subaharan, K., Manjunath, S., Pragadheesh, V. S., Bakthavatsalam, N., Mohan, M. G., . . . Basavarajappa, S. (2021). Electrophysiological, behavioural and biochemical effect of Ocimum basilicum oil and its constituents methyl chavicol and linalool on Musca domestica L. *Environ Sci Pollut Res Int, 28* (36), 50565-50578. doi:10.1007/s11356-021-14282-x

Sharma, Y., Velamuri, R., Fagan, J., & Schaefer, J. (2020). Full-Spectrum Analysis of Bioactive Compounds in Rosemary (Rosmarinus officinalis L.) as Influenced by Different Extraction Methods.*Molecules, 25* (20). doi:10.3390/molecules25204599

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics, 31* (19), 3210-3212. doi:10.1093/bioinformatics/btv351

Singh, P., & Pandey, A. K. (2018). Prospective of Essential Oils of the Genus Mentha as Biopesticides: A Review. *Front Plant Sci, 9* , 1295. doi:10.3389/fpls.2018.01295

Srividya, N., Davis, E. M., Croteau, R. B., & Lange, B. M. (2015). Functional analysis of (4S)-limonene synthase mutants reveals determinants of catalytic outcome in a model monoterpene synthase.*Proc Natl*

Acad Sci U S A, 112 (11), 3332-3337. doi:10.1073/pnas.1501203112

Srividya, N., Lange, I., & Lange, B. M. (2020). Determinants of Enantiospecificity in Limonene Synthases. *Biochemistry, 59* (17), 1661-1664. doi:10.1021/acs.biochem.0c00206

Starks, C. M., Back, K., Chappell, J., & Noel, J. P. (1997). Structural basis for cyclic terpene biosynthesis by tobacco 5-epi-aristolochene synthase. *Science, 277* (5333), 1815-1820. doi:10.1126/science.277.5333.1815

Su, P., Tong, Y., Cheng, Q., Hu, Y., Zhang, M., Yang, J., . . . Huang, L. (2016). Functional characterization of ent-copalyl diphosphate synthase, kaurene synthase and kaurene oxidase in the Salvia miltiorrhiza gibberellin biosynthetic pathway. *Sci Rep, 6* , 23057. doi:10.1038/srep23057

Tang, H., Zhang, X., Miao, C., Zhang, J., Ming, R., Schnable, J. C., . . . Lu, J. (2015). ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol, 16* . doi:10.1186/s13059-014-0573-1

Van de Peer, Y., Mizrachi, E., & Marchal, K. (2017). The evolutionary significance of polyploidy. *Nat Rev Genet, 18* (7), 411-424. doi:10.1038/nrg.2017.26

Varet, H., Brillet-Gueguen, L., Coppee, J. Y., & Dillies, M. A. (2016). SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data. *Plos One, 11* (6), e0157022. doi:10.1371/journal.pone.0157022

Veenstra, J. P., & Johnson, J. J. (2021). Rosemary (Salvia rosmarinus): Health-promoting benefits and food preservative properties. *Int J Nutr, 6* (4), 1-10.

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics, 33* (14), 2202-2204. doi:10.1093/bioinformatics/btx153

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., . . . Paterson, A. H. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res, 40* (7), e49. doi:10.1093/nar/gkr1293

Wedler, H. B., Pemberton, R. P., & Tantillo, D. J. (2015). Carbocations and the Complex Flavor and Bouquet of Wine: Mechanistic Aspects of Terpene Biosynthesis in Wine Grapes. *Molecules, 20* (6), 10781-10792. doi:10.3390/molecules200610781

Williams, D. C., McGarvey, D. J., Katahira, E. J., & Croteau, R. (1998). Truncation of limonene synthase preprotein provides a fully active 'pseudomature' form of this monoterpene cyclase and reveals the function of the amino-terminal arginine pair. *Biochemistry, 37* (35), 12213-12220. doi:10.1021/bi980854k

Wu, L., Zhao, Y., Zhang, Q., Chen, Y., Gao, M., & Wang, Y. (2020). Overexpression of the 3-hydroxy-3-methylglutaryl-CoA synthase gene LcHMGS effectively increases the yield of monoterpenes and sesquiterpenes. *Tree Physiol, 40* (8), 1095-1107. doi:10.1093/treephys/tpaa045

Xu, J., Ai, Y., Wang, J., Xu, J., Zhang, Y., & Yang, D. (2017). Converting S-limonene synthase to pinene or phellandrene synthases reveals the plasticity of the active site. *Phytochemistry, 137* , 34-41. doi:*https://doi.org/10.1016/j.phytochem.2017.02.017*

Xu, Z., Gao, R., Pu, X., Xu, R., Wang, J., Zheng, S., . . . Song, J. (2020). Comparative Genome Analysis of Scutellaria baicalensis and Scutellaria barbata Reveals the Evolution of Active Flavonoid Biosynthesis. *Genomics Proteomics Bioinformatics, 18* (3), 230-240. doi:10.1016/j.gpb.2020.06.002

Yang, S. A., Jeon, S. K., Lee, E. J., Shim, C. H., & Lee, I. S. (2010). Comparative study of the chemical composition and antioxidant activity of six essential oils and their components. *Nat Prod Res, 24* (2), 140-151. doi:10.1080/14786410802496598

Yi, X., Wang, X., Wu, L., Wang, M., Yang, L., Liu, X., . . . Shi, Y. (2021). Integrated Analysis of Basic Helix

Loop Helix Transcription Factor Family and Targeted Terpenoids Reveals Candidate AarbHLH Genes Involved in Terpenoid Biosynthesis in Artemisia argyi. *Front Plant Sci, 12* , 811166. doi:10.3389/fpls.2021.811166

Yu, L., Diao, S., Zhang, G., Yu, J., Zhang, T., Luo, H., . . . Zhang, J. (2022). Genome sequence and population genomics provide insights into chromosomal evolution and phytochemical innovation of Hippophae rhamnoides. *Plant Biotechnol J, 20* (7), 1257-1273. doi:*https://doi.org/10.1111/pbi.13802*

Yu, M. H., Choi, J. H., Chae, I. G., Im, H. G., Yang, S. A., More, K., . . . Lee, J. (2013). Suppression of LPS-induced inflammatory activities by Rosmarinus officinalis L. *Food Chem, 136* (2), 1047-1054. doi:10.1016/j.foodchem.2012.08.085

Zwaenepoel, A., & Van de Peer, Y. (2019). wgd—simple command line tools for the analysis of ancient whole-genome duplications.*Bioinformatics, 35* (12), 2153-2155. doi:10.1093/bioinformatics/bty915

Figures

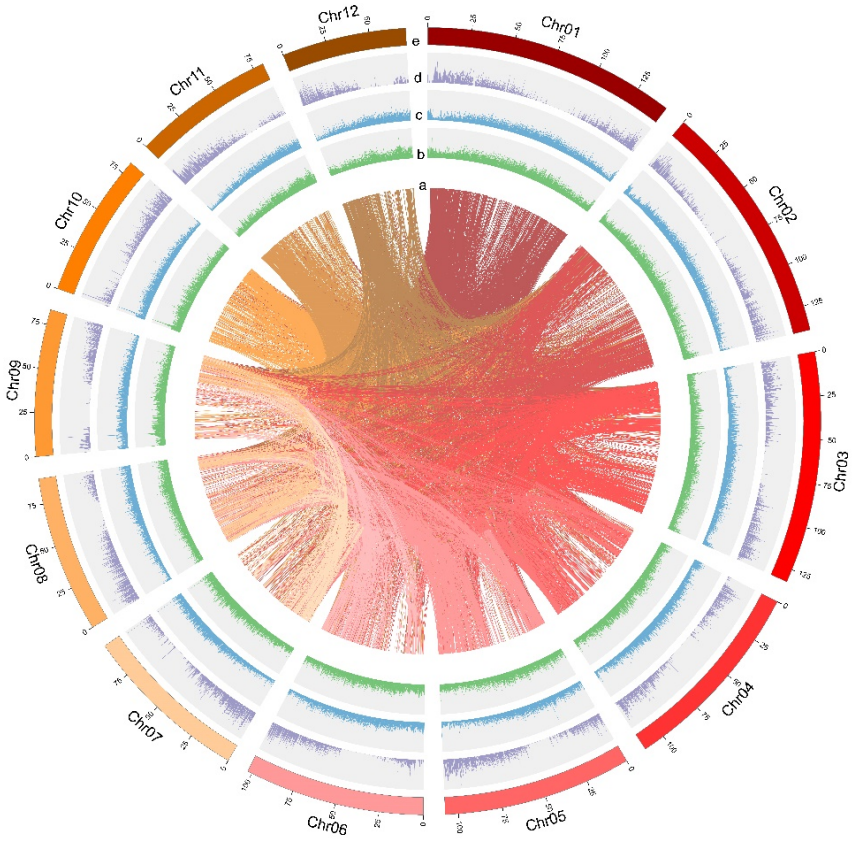

Figure 1 Overview of *S. rosmarinus* genome Assembly

The outer lines represent the 12 pseudochromosomes of *S. rosmarinus* . Within the circle plot, colorful strips represent the density, and all detected gene duplications are indicated with links inside the circle. **a** Syntenic blocks, **b** GC (guanine-cytosine) density, **c** Repeats density, **d** Genes density, **e** Length (Mb) of 12 chromosomes. Scales showed chromosomes in a 10-kb window, genes density in a 100-kb window (0-27), repeats density in a 100-kb window (0-5004), and GC density in a 100-kb window (0-7053).

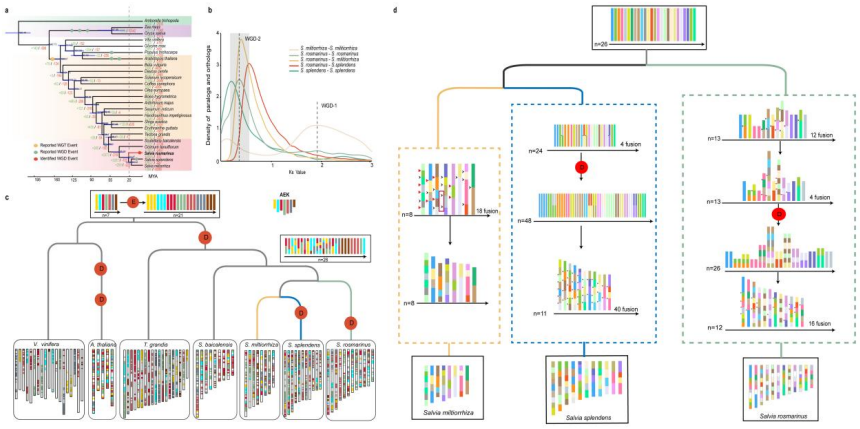

Figure 2 Comparative Genomic Analysis.

a The phylogenetic tree based on the Bayesian inference method was constructed using 465 single-copy homologous genes from 24 species. The basal angiosperm *Amborella trichopoda* was selected as outgroup. A red branche represent *S. rosmarinus* , which diverged from *S. miltiorrhiza* and *S. splendens* at ca. 21.47 Mya. Orange ellipse represent the reported WGT event, and green and red ellipses represent reported WGD events and the identified WGD event in this study, respectively. **b** Synonymous substitution rate (*Ks*) distributions of syntenic blocks for the paralogs and orthologs of *S. rosmarinus, S. miltiorrhiza* and *S. splendens.* The gray box indicates the unique WGD event in *S. rosmarinus* . **c** The karyotype evolution from seven chromosomes of a eudicot ancestor to seven species (*Vitis vinfera* ,*Arabidopsis thaliana* , *Scutellaria baicalensis* ,*Tectona grandis* , *S. miltiorrhiza* , *S. splendens* and*S. rosmarinus*). Genome polyploidy events are indicated by red circles, and the lower columns show the retention of ancestral genes in the chromosomes of seven species. The yellow branch represents *S. miltiorrhiza* , and the green branch represents *S. rosmarinus* .**d** The karyotype evolution of *S. rosmarinus, S. miltiorrhiza* and *S. splendens* from their recent common ancestor to the present. The karyotype evolution speculation map of *S. rosmarinus* in the green dashed box, *S. miltiorrhiza* in the yellow dashed box and *S. splendens* in the blue dashed box. D represents the occurrence of whole genome duplicated event.

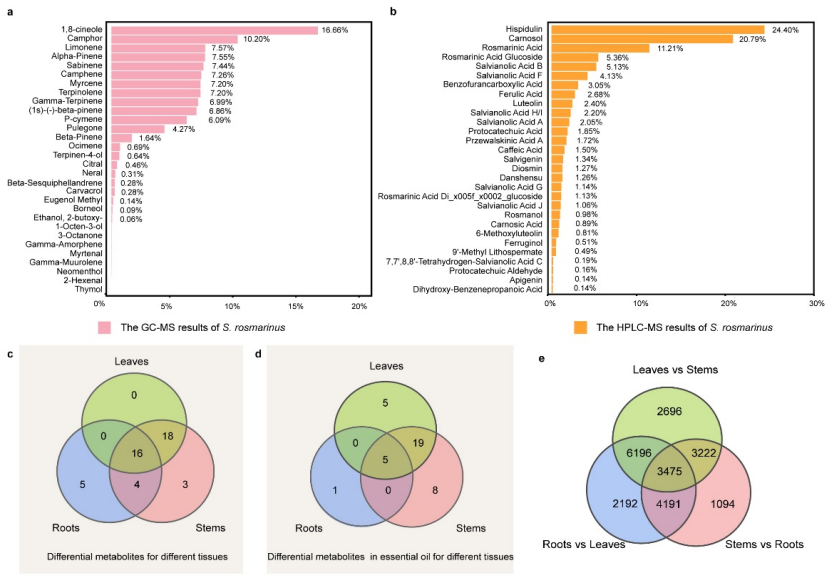

Figure 3 Metabolic analysis and transcriptome analysis

a-b The GC-MS results and HPLC-MS results in the leaves of*S. rosmarinus* . Components were ordered by proportions and labeled. Pink bars and orange bars represent the result of GC-MS and HPLC-MS, respectively. **c-d** Venn diagram of differential metabolites for roots, stems and leaves. Nonoverlapping regions represent the metabolites that are specific to the different tissues, while overlapping regions represent the metabolites that are common to several different tissues. **e** Venn diagram of DEGs for roots vs. leaves, stems vs. roots, and leaves vs. stems. Nonoverlapping regions represent the DEGs that are specific to roots, stems and leaves, while overlapping regions represent the genes differential expressed in three tissues.

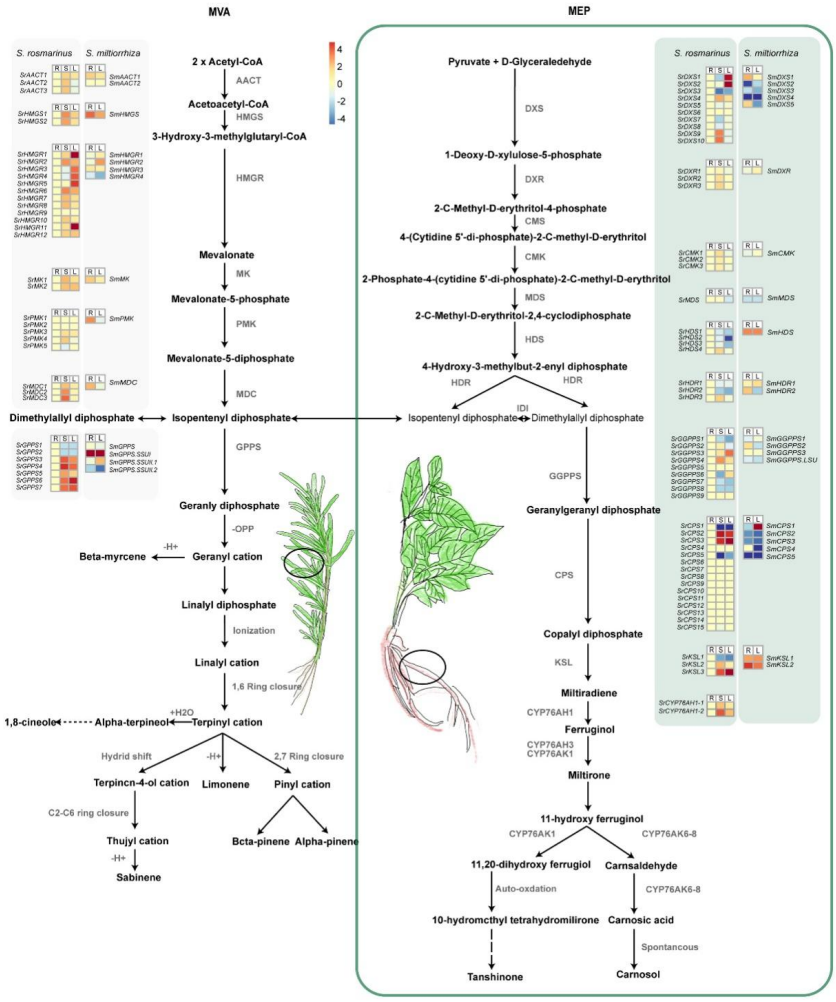

Figure 4 The synthesis of terpenoids in *S. rosmarinus*

Tissue-specific relative expression profiles (red–blue scale) of genes implicated in terpenoid biosynthesis (heat map). Intermediates are shown in black, and the enzymes involved at each step are shown in gray. The genes involved in the pathways exhibit high level of expression, which may contribute to the biosynthesis of large amounts of terpenes. Monoterpene was biosynthesized by MVA pathway in cytoplasm, and diterpene was biosynthesized by MEP pathway in plastid, framed with green box. Grey boxes represent the genes on MVA pathway of *S. rosmarinus* (left) and *S. miltiorrhiza* (right), and light green boxes represent the genes on MEP pathway of *S. rosmarinus* (left) and *S. miltiorrhiza* (right). All genes involved in the pathways of MEP and MVA almost expanded in *S. rosmarinus* . MVA pathway mevalonate pathway, MEP pathway mevalonate-independent (deoxyxylulose phosphate) pathway, R root, S stem, L leaves.

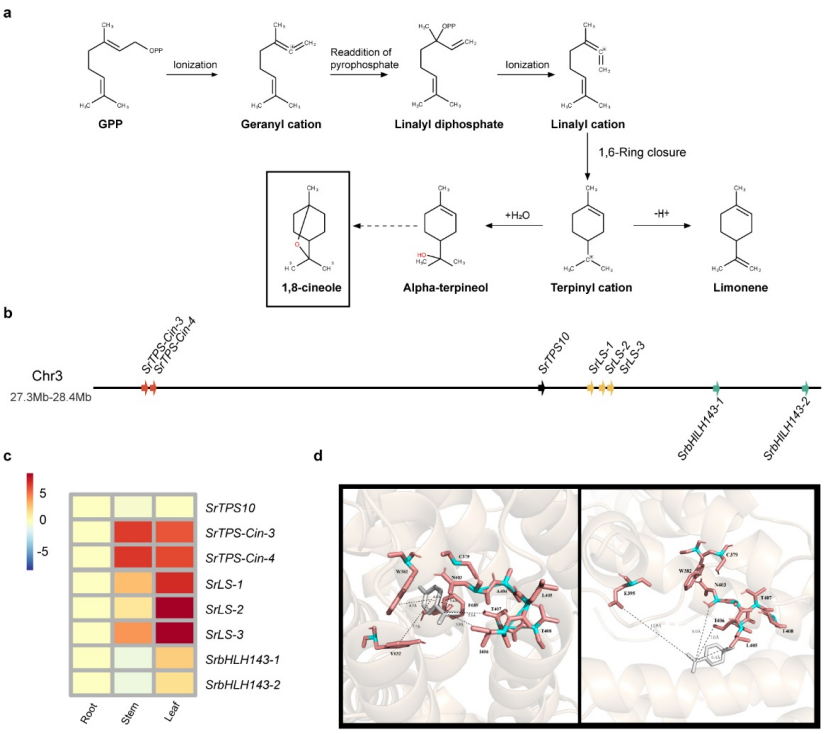

Figure 5 Biosynthetic pathway of 1,8-cineole and gene structure analysis

a Downstream pathway of 1,8-cineole biosynthesis and limonene synthase. Intermediates are shown in structural formula with black and bold, and reaction process involved in each step are shown in grey color. Dotted arrows indicate predicted or unknown reaction. The solid boxes mean 1,8-cineole and limonene. **b** The cluster of 1,8-cineole synthase and limonene synthase in pseudochromosome 3 of*S. rosmarinus* . Genes are labeled with italic, arrows present the position and strand of genes. Red arrows represent *SrCinSs* in cluster, yellow arrows represent SrLSs in cluster, black arrow means the other TPS of the cluster, and blue arrows present the transcription factors of the cluster. **c** The expression of *SrCinSs* and*SrLSs* in three tissues of *S. rosmarinus* plants. Genes are shown in italic. Red, high expression; blue, low expression. **d**Docking sites demonstrating active-site amino acid substitutions of*SrCinS-4* (left) and *SrCin-5* (right), which highly express in the leaves of *S. rosmarinus* . Ligand alpha-terpinyl is indicated with a white stick, and residues in active pocket are shown in sticks, helixes are shown in blue, loops are shown in pink. Residues and position are labeled in black and bold. The dotted lines with numbers above in gray and bold indicate the distance of residues and alpha-terpinyl.

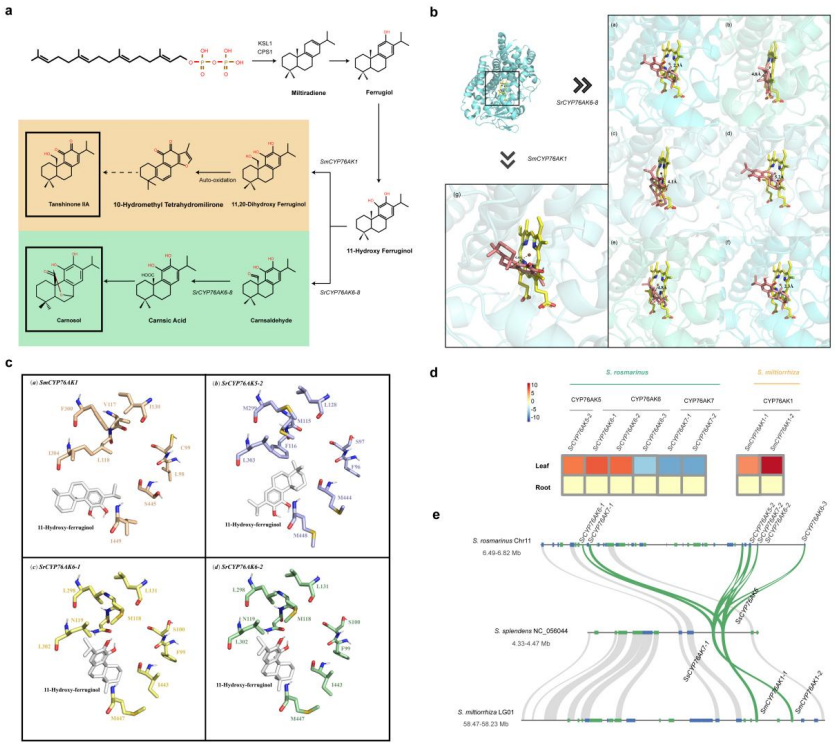

Figure 6 Structure comparison of *SrCYP76AK6-8* and *SmCYP76AK1* reveals the divergent evolution.

a Downstream pathway of carnosol and tanshinone biosynthesis, starting from geranylgeranyl diphosphate. Immediate are shown in structure and labeled by bold. Enzymes involved in each step are labeled by italic. The divergence of carnosol and tanshinone biosynthesis start from 11-hydroxy-ferruginol. *CYP76AK6-8* catalyze 11-hydroxy-ferruginol and synthesize carnosol precursors in *S. rosmarinus* , which is colored by pink block. *CYP76AK1* synthesize tanshinone precursors in *S. miltiorrhiza* . Dotted arrows indicate omitted reaction. **b** Homology modeling and docking analysis of*SrCYP76AK6-8* and *SmCYP76AK1* from *S. rosmarinus* and*S. miltiorrhiza* . **a** Homology modeling of*SrCYP76AK6-1* . Docking poses of *SrCYP76AK7-1* **b** ,*SrCYP76AK5-2* **c** , *SrCYP76AK7-2* (d),*SrCYP76AK6-2* (e), *SrCYP76AK6-3* (f), *SmCYP76AK1* (g). Compound structure is depicted as stick with carbons colored pink and oxygens red. Heme is depicted as stick with carbons colored yellow and iron blue. Distance between C-20 and heme iron is indicated by dashed line with the length indicated in Å. **c** Residues within 8 Å around docking sites of *SrCYP76AKs* , which are different from*SmCYP76AK1* . **a** Residues pose of *SmCYP76AK1* in pink color and ligand 11-hydroxy-ferruginol in white. Residues pose of*SrCYP76AK5-2* in blue color **b** , *SrCYP76AK6-1* in yellow **c** , *SrCYP76AK6-2* in green (d). Residues in active pocket are shown in sticks, with oxygen atom in red, hydrogen in white and nitrogen atom in blue. Ligand 11-hydroxy-ferruginol is shown in white sticks. **d** The expression of *SrCYP76AKs* and*SmCYP76AK1* in roots and leaves. Red, high expression; blue, low expression. Genes are labeled in italic which are shown based on functionally annotation in bold. Genes of *S. rosmarinus* are marked in green and genes of *S. miltiorrhiza* are marked in orange. **e** Syntenic blocks of *CYP76AK* regions within *S. rosmarinus, S. miltorrhiza* and *S. splendens* . Green curves represent the co-linearity of *CYP76AKs.SrCYP76AK5-8* in *S. rosmarinus* can be linked to*SmCYP76AK1* in *S. miltiorrhiza* through *SsCYP76AK7* in*S. splendens* , which provide evidence for the evolutionary of*CYP76AK* subfamily.