the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2023) **113,** 819–832 doi: 10.1111/tpj.16087

The chromosome-scale assembly of the *Salvia rosmarinus* genome provides insight into carnosic acid biosynthesis

Danlu Han[1], Wenliang Li[1], Zhuangwei Hou[2], Chufang Lin[1], Yun Xie[1], Xiaofang Zhou[3], Yuan Gao[4], Junwen Huang[1], Jianbin Lai[1], Li Wang[2], Liangsheng Zhang[5,*] and Chengwei Yang[1,6,*]

[1]*Guangdong Provincial Key Laboratory of Biotechnology for Plant Development, School of Life Science, South China Normal University, 510631, Guangzhou, China,*

[2]*Shenzhen Branch Guangdong Laboratory for Lingnan Modern Agriculture/Genome Analysis Laboratory of the Ministry of Agriculture/Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China,*

[3]*Guangdong Laboratory for Lingnan Modern Agriculture, Guangdong Province Key Laboratory of Microbial Signals and Disease Control, Integrative Microbiology Research Center, South China Agricultural University, 510642, Guangzhou, China,*

[4]*Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China,*

[5]*Genomics and Genetic Engineering Laboratory of Ornamental Plants, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310058, China, and*

[6]*SCNU Qingyuan Institute of Science and Technology Innovation Co., Ltd., Qingyuan 511517, China*

Received 9 August 2022; revised 16 December 2022; accepted 26 December 2022; published online 29 December 2022.

*For correspondence (e-mail yangchw@scnu.edu.cn and zls83@zju.edu.cn).

SUMMARY

Rosemary (*Salvia rosmarinus*) is considered a sacred plant because of its special fragrance and is commonly used in cooking and traditional medicine. Here, we report a high-quality chromosome-level assembly of the *S. rosmarinus* genome of 1.11 Gb in size; the genome has a scaffold N50 value of 95.5 Mb and contains 40 701 protein-coding genes. In contrast to other diploid Labiataceae, an independent whole-genome duplication event occurred in *S. rosmarinus* at approximately 15 million years ago. Transcriptomic comparison of two *S. rosmarinus* cultivars with contrasting carnosic acid (CA) content revealed 842 genes significantly positively associated with CA biosynthesis in *S. rosmarinus*. Many of these genes have been reported to be involved in CA biosynthesis previously, such as genes involved in the mevalonate/methylerythritol phosphate pathways and CYP71-coding genes. Based on the genomes and these genes, we propose a model of CA biosynthesis in *S. rosmarinus*. Further, comparative genome analysis of the congeneric species revealed the species-specific evolution of CA biosynthesis genes. The genes encoding diterpene synthase and the cytochrome P450 (CYP450) family of CA synthesis-associated genes form a biosynthetic gene cluster (*CPSs–KSLs–CYP76AHs*) responsible for the synthesis of leaf and root diterpenoids, which are located on *S. rosmarinus* chromosomes 1 and 2, respectively. Such clustering is also observed in other sage (*Salvia*) plants, thus suggesting that genes involved in diterpenoid synthesis are conserved in the Labiataceae family. These findings provide new insights into the synthesis of aromatic terpenoids and their regulation.

Keywords: Salvia rosmarinus, genome assembly, carnosic acid, Salvia.

INTRODUCTION

Rosemary (*Salvia rosmarinus*), also known as ‘sea dew’ or ‘antos’, is a small aromatic shrub belonging to the genus *Salvia* in the family Lamiaceae, which originates in Europe and along the Mediterranean coast (Sasikumar, 2012). As one of the oldest Mediterranean herbs, *S. rosmarinus* has been used as a food flavoring agent and natural medicine for over a thousand years. Greek healers found *S. rosmarinus* was beneficial for treating conditions related to the brain, heart, and eyes. They also believe that placing it under pillows can dispel negative emotions, prevent nightmares, and enhance memory. Moreover, in ancient Rome, the dead were embalmed by placing *S. rosmarinus* in tombs, and in the Middle Ages, leaves and twigs of *S. rosmarinus* were burned to kill bacteria in hospitals (Kumar et al., 2016). *Salvia rosmarinus* was introduced to China in 220 B.C. and has been passed on as a folk herb ever since.

Salvia rosmarinus belongs to the genus *Salvia* in the family Labiataceae and contains various terpenoids and volatile oils: carnosic acid (CA), carnosol (CO), rosmarinic acid, ursolic acid, alpha-pinene, camphor, and camphene (Drew et al., 2017). Notably, CA is a natural antioxidant and has antibacterial properties, and therefore it has been widely used for seasonings and food preservation (González-Minero et al., 2020). It also exhibits anti-aging, fat-reducing, and anticancer effects, and it can be used to treat cardiovascular diseases (Kamli et al., 2022). *Salvia rosmarinus* contains higher amounts of CA than other plant species (3–50 mg g^{-1} dry weight [DW] vs. approximately 0.1–21.8 mg g^{-1} DW) (Richheimer et al., 1996; Schwarz & Ternes, 1992b; Wenkert et al., 1965), and therefore it is suitable for exploring CA biosynthesis.

CA consists of a tricyclic C_{20} abietane skeleton with a typical catechol function in the C ring (Birtić et al., 2015). This molecule is thought to exhibit antioxidant activity through its oxidative conversion to the quinone form. Although both CA and CO exhibit antioxidant properties, CA can be spontaneously oxidated to CO, whereas CO is less antioxidative than CA. CA biosynthesis and accumulation occur only in the photosynthetic green tissues of *S. rosmarinus* (leaves, sepals, and petals) (Munné-Bosch & Alegre, 2001). CA also accumulates in glandular trichomes (Brückner et al., 2014). Owing to the oxidative properties of CA, it is partially depleted during leaf development and senescence and during seasonal changes, under conditions such as high temperatures and droughts.

Terpene metabolites are formed mainly via the mevalonate (MVA) and methylerythritol phosphate (MEP) pathways to form the C_5-isoprenoid precursors isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP). Then, isoprenoid transferase or isoprenoid diphosphate synthase catalyzes the fusion of the C_5-units of IPP and DMAPP to produce all isoprene-like linear precursors. Finally, acyclic C_5- to C_{20}-prenyl diphosphate intermediates are produced by diphosphate synthase and through cytochrome catalysis (Tholl & Lee, 2011). The process of CA biosynthesis has been characterized by reconstitution yeast and *Escherichia coli* systems (Božić et al., 2015; Ignea et al., 2016; Wang et al., 2020). CA is a diterpene compound, and recent studies mainly suggest that it originates from the liposomal 1-deoxyxylulose 5-phosphate (DXP) pathway or the cytoplasmic MVA pathway (Gao et al., 2009). Several genes involved in CA synthesis have been identified in *S. rosmarinus* and *Salvia fruticosa*, such as copalyl diphosphate synthase (CPS) and miltiradiene synthase (MDS/KSL), which are responsible for the formation of the CA precursors copalyl diphosphate and miltiradiene, respectively. Ferruginol synthases (FSs, CYP76AH1) oxidize abietatriene to ferruginol, and then 11-hydroxyferruginol synthase (HFS, CYP76AH22–24) oxidizes ferruginol to form 11-hydroxyferruginol, which is converted to CA by CYP76AK6–8 with C_{20}-oxidase (C20Ox) activity (Božić et al., 2015; Ignea et al., 2016; Scheler et al., 2016). However, the whole-genome sequence of *S. rosmarinus* remains undetermined, which limits the study of CA synthetase and the analysis of genome evolution.

Here, we utilized both short- and long-read genome sequencing and high-throughput chromosome conformation (Hi-C) strategies to obtain a high-quality chromosome-level genome assembly of *S. rosmarinus*, carried out genome comparison with other *Salvia* species to unravel the basic characteristics of its genome evolution including the divergence times and whole-genome duplication (WGD) events, and finally explored the key genes in CA biosynthesis. This study provides important insights into the molecular breeding of rosemary and the evolution of diterpenoid synthesis in Labiataceae.

RESULTS

Genome sequencing, assembly, and annotation

The diploid *S. rosmarinus* was estimated to have a genome size of 1.08 Gb and high levels of heterozygosity (1.2%) and repetitive sequences (69.7%) based on flow cytometry (Table S1) and *K*-mer distribution analyses (Figure S1; Table S2). We thus sequenced the *S. rosmarinus* genome using PacBio HiFi long-read sequencing technology, which is advantageous for genomes of high complexity. In total, 39.1 Gb of circular consensus sequencing long reads (average read length: 49.24 kb) were generated (Table S3). The long reads were assembled using Canu and further processed by HaploMerger2 to remove allelic sequences, resulting in a 1.03-Gb draft genome assembly with 160 contigs and an N50 length of 26.56 Mb (Table S4). To obtain a chromosome-scale assembly of the *S. rosmarinus* genome, we mapped a total of 103 Gb high-throughput chromosome conformation capture (Hi-C) data to the draft *S. rosmarinus* genome and anchored the genome sequences to 12 pseudochromosomes with an N50 value of 95.5 Mb (Figure S2; Table S4). The genome assembly was evaluated by Benchmarking Universal Single-Copy Orthologs (BUSCO) using true dicotyledons, green plants, and eukaryotes datasets which all revealed high levels of completeness (97.6%, 99.0%, and 99.6%, respectively), indicating that the *S. rosmarinus* genome assembly is nearly complete (Figure S3; Table S5).

Based on *ab initio*, homology-based, and transcriptome-guided gene prediction methods, the *S. rosmarinus* genome was predicted to contain 49 858 protein-coding genes, among which 81.6% genes were functionally annotated based on sequence similarity and protein domain analyses (Figure S4a,b). The average length of these protein-coding genes was 2990 bp and the average coding sequence (CDS) length was 249 bp, which were similar to those of other annotated species belonging to Labiataceae (Table S6). Furthermore, 3815 non-coding RNAs (ncRNAs) were predicted in the *S. rosmarinus* genome, including 378 rRNAs, 961 micro-RNAs (miRNAs), 878 small nuclear RNAs (snRNAs), and 826 tRNAs (Figure S4c).

A comparative analysis revealed that the genome of *S. rosmarinus* was larger than those of other Labiataceae diploids. We found that 72.72% of the *S. rosmarinus* genome was comprised of repetitive sequences, with long terminal repeat (LTR) retrotransposons (RTs) accounting for 68.6% of the genome and two families, Ty1/Copia (17.19%) and gypsy/DIRS1 (36.57%), dominating the genome (Figure 1a, Table S7). A total of 1959 intact Copia family LTR-TRs and 4605 intact Gypsy family LTR-TRs were identified in the *S. rosmarinus* genome; both appeared to explode approximately 2 million years ago (MYA) (Figure S5). Thus, the recent LTR insertion and activation is likely a critical factor underlying the size expansion of the *S. rosmarinus* genome.

Phylogenetic analysis and whole-genome duplication events of the *S. rosmarinus* genome

To infer the evolutionary history of *S. rosmarinus*, phylogenetic reconstruction and divergence time estimation analyses were conducted based on 107 single-copy genes shared by *S. rosmarinus* and 16 plant species, including 11 previously sequenced Labiataceae species, namely *Salvia officinalis*, *Salvia bowleyan*, *Salvia miltiorrhiza*, *Salvia splendens*, *Scutellaria baicalensis*, *Scutellaria barbata*, *Origanum vulgare*, *Origanum majorana*, *Tectona grandis*, *Callicarpa americana*, and *Hyssopus officinalis*, as well as five

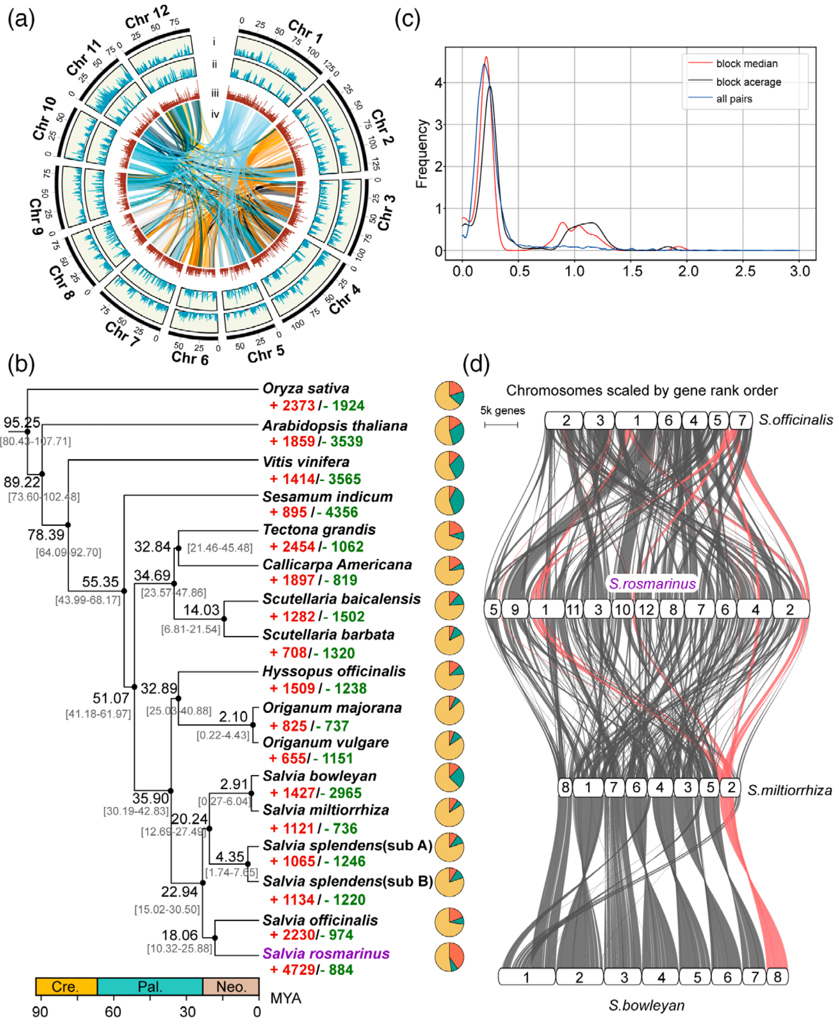

Figure 1. The *S. rosmarinus* genome annotation. (a) Characterization of different elements on the *S. rosmarinus* chromosome. The 12 chromosomes of the assembled *S. rosmarinus*. (i) distribution of Ty1/Copia in the *S. rosmarinus* genome, (ii) distribution of gypsy/DIRS1, (iii) GC content, (iv) collinear blocks on chromosomes in the *S. rosmarinus* genome. Block size = 25 kb. (b) The phylogenetic tree indicates the number of gene families that are expanded (red) or contracted (green) in each of the 17 species. The pie charts show the percentage of expanded (red), contracted (green), and conserved (yellow) gene families across all gene families. Estimated divergence times (in millions of years) are shown below the phylogenetic tree in black. The tree is rooted with *Oryza sativa* as the outgroup. (c) *Ks* frequency distribution chart of *S. rosmarinus*. (d) Syntenic plots present a syntenic relationship between the *S. rosmarinus* genome with *S. bowleyana*, *S. officinalis*, and *S. miltiorrhiza* genomes.

model plants. Consistent with previous reports, the estimated divergence time for Labiataceae is approximately 51.07 MYA, with a 95% highest posterior density of 41.18–61.97 MYA. The four Salvia species (i.e., *S. Rosmarinus*, *S. officinalis*, *S. bowleyan*, and *S. miltiorrhiza*) form a monophyletic clade, with an estimated divergence time of approximately 22.94 MYA with a 95% highest posterior density of 15.02–30.50 MYA. *Salvia rosmarinus* and *S. officinalis* are sister to each other, with an estimated divergence time of approximately 18.06 MYA with a 95% highest posterior density of 10.32–25.88 MYA (Figure 1b).

At the same time, 33 947 homologous gene families were identified from these 17 plant species, and gene family evolutionary analysis was conducted using CAFE 5. Families in the *S. rosmarinus* genome showed more expansion than those in other species, with 4729 expanded orthogroups and 884 contracted orthogroups (Figure 1b). Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis demonstrated that the expanded gene families in *S. rosmarinus* were enriched in terpene skeleton synthesis and cytochrome P450 (CYP450) (Figure S6a). In the Gene Ontology (GO) analysis, the expanded orthogroups were significantly enriched in flavonoid metabolic processes and diterpene phytoalexin metabolic process (Figure S6b,c). Thus, the specific expansion of orthogroups for these biological processes may be related to the crucial role of *S. rosmarinus* in metabolic processes.

The expansion of gene families during evolution can be mainly attributed to proximal duplication (PD), tandem duplication (TD), and WGD. Particularly, WGD is a major force contributing to the rapid evolution of flowering plants. On analyzing the distribution of synonymous substitution rates (*Ks*) between paralogs within the *S. rosmarinus* genome (Figure S7a), we found significant peaks at 0.25, 0.8, and 1.8 (Figure 1c), indicating that *S. rosmarinus* underwent three WGD events during evolution (Figure 1c). The oldest event was estimated to have occurred around 115–132 MYA, corresponding to the whole-genome triplication (WGT) in the core gymnosperm ancestor, and the second event occurring around 48 MYA is likely shared by the Labiatae family, while the most recent one occurred approximately 15 MYA and is unique to *S. rosmarinus*. Relative to the other eight species of the Labiataceae family with a 4-fold degenerate regression rate (4Dtv), *S. rosmarinus* exhibited characteristic peaks at approximately 0.066 and 0.27. In addition, its direct homologs with *S. bowleyana* and *S. barbata* had a peak at 0.069 and 0.237, respectively (Figure S7b). These peaks represent a WGD event for *S. rosmarinus* after divergence from *S. bowleyana* and *S. officinalis* at approximately 15.1 MYA, with a 95% highest posterior density of 9.8–24.0 MYA. An earlier WGD event was shared with *S. barbata* before divergence from Labiataceae at around 32.7–53.7 MYA. This is further corroborated by the results of 4Dtv analysis and syntenic analysis (Figure 1d; Figures S7b and S8).

Based on the classification of duplicated genes, 22 224 WGDs, 1086 TDs, 1094 PDs, 8204 transposed duplications (TRDs), and 8514 dispersed duplications (DSDs) were identified in the *S. rosmarinus* genome (Figure S9a,b). The GO and KEGG analyses of these duplication types revealed that the results were consistent with the previous analysis of expansion/contraction genes and TD- and PD-type duplications mainly involved in the diterpenoid biosynthesis pathway and metabolic biological processes (Figures S9c and S10).

Transcriptome analysis of terpenoid biosynthesis pathways in *S. rosmarinus*

To investigate the CA synthesis pathway in *S. rosmarinus*, we selected one of 12 different cultivars, U2 (II-03-US from the USA), which had significantly lower CA content than wild type (WT; Rosemary Morocco) and other varieties, for the present study (Figure 2a; Table S8). Old and young leaves of the two cultivars were subjected to CA content analysis and transcriptome sequencing (three biological replicates per sample). Principal component analysis (PCA) revealed that the transcriptome sequencing data were of high quality, and the three biological samples were reproducible (Figure S11a) and therefore suitable for subsequent analysis.

Based on the *S. rosmarinus* genome, 67 terpene skeleton biosynthesis genes were identified and their expression patterns were further analyzed (Figure 2b). Most genes were predicted to encode rate-limiting enzymes such as 3-hydroxy-3-methylglutaryl-CoA reductase (HMGR) and geranylgeranyl diphosphate synthase (GGPPS) in the MVA pathway of terpene skeleton biosynthesis, indicating the vital and diverse roles of these enzymes in *S. rosmarinus* terpene biosynthesis. Pearson correlation analyses showed that the gene expression patterns of *hydroxymethylglutaryl-CoA synthase* (*HMGS*), *HMGR*, *phosphomevalonate kinase* (*PMK*), *diphosphomevalonate decarboxylase* (*MVD*), and *isopentenyl diphosphate isomerase* (*IDI*) were positively correlated with the pattern of CA accumulation, whereas the gene expression patterns of *4-hydroxy-3-methylbut-2-enyl diphosphate reductase* (*HDR*) and *1-deoxy-D-xylulose-5-phosphate synthase* (*DXS*) genes were negatively correlated with CA accumulation (Figure S12). In addition, genomic collinearity analysis was conducted to identify tandem duplicates of *HDR* and *HMGR*, particularly *HMGR*, whose gene numbers were increased compared to *S. officinalis* (Figure 2c). The phylogenetic trees of *HMGR* genes of *S. officinalis* and *S. rosmarinus* indicated that *Ro085620* and *Ro085630* expansion occurred before the speciation of *S. rosmarinus* and *S. officinalis*, but for *Ro495030* and *Ro495020* tandem duplication events occurred only in *S. rosmarinus*. Diterpenoids such as CA

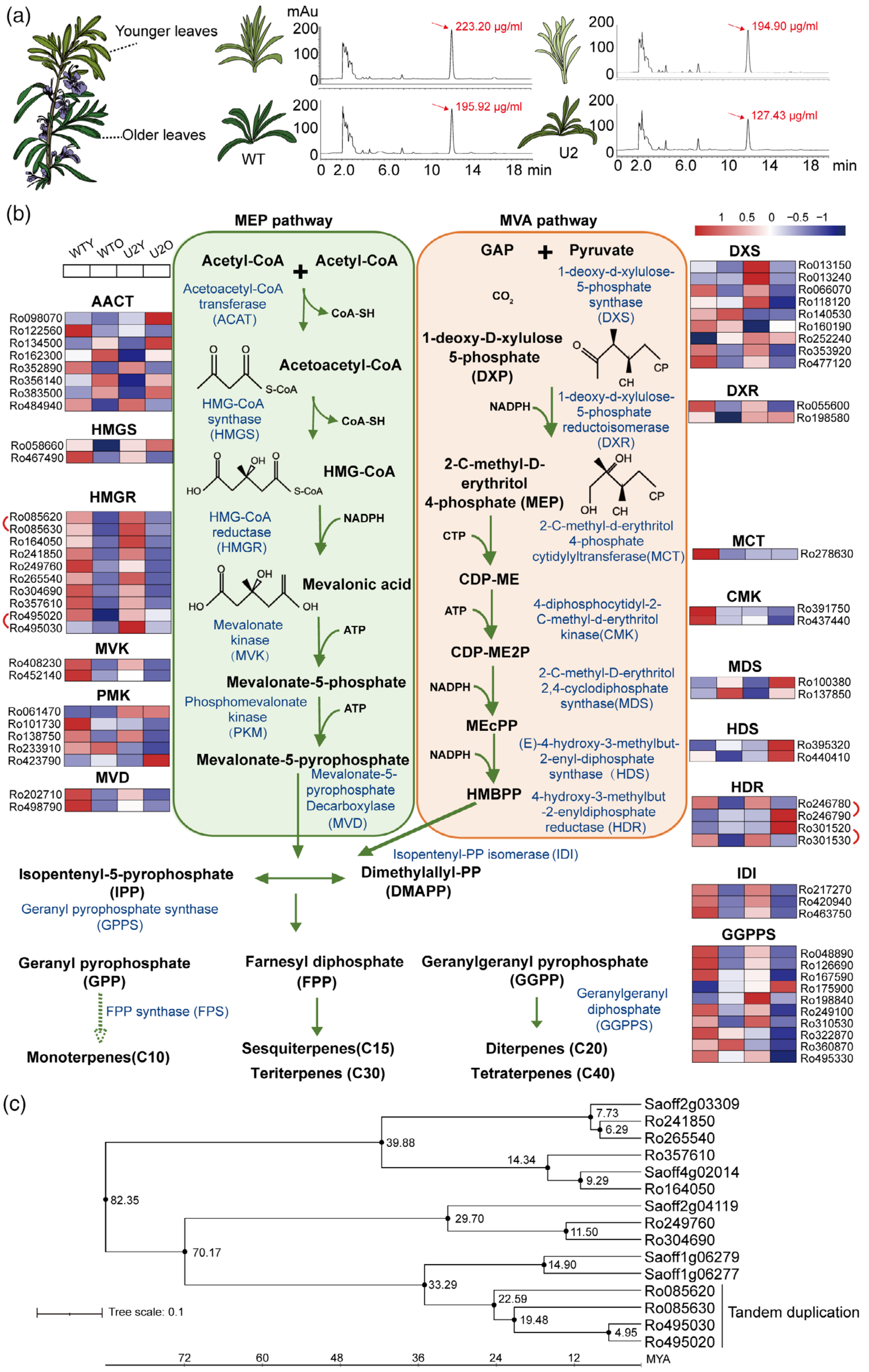

Figure 2. Determination of CA content in *S. rosmarinus* and gene expression analysis of the CA biosynthetic pathway. (a) Determination of CA contents in different tissue samples through HPLC. (b) Heatmap representation of transcriptional regulation of the CA synthesis pathway. Gene representation of the enzymes predicted in *S. rosmarinus*. Relative expression levels are shown as $\log_2$(fold change in transcript levels) values. Red, high expression; blue, low expression. (c) Evolutionary tree analysis of the 3-hydroxy-3-methylglutaryl-CoA reductase (HMGR) genes of *S. officinalis* and *S. rosmarinus*, constructed by the neighbor-joining (NJ) method by comparing protein sequences. Nodes indicate the time of divergence. MYA, million years ago.

and CO are the main antioxidant components in *S. rosmarinus*. To identify CA synthesis-related genes, we further carried out differential gene expression analysis between old and young leaves of WT and U2 (Figure S11b; Table S9). By centralizing and normalizing the fragments per kilobase of transcript per million mapped reads (FPKM) values of genes, *K*-means clustering analysis was performed. All differentially expressed genes were classified into 10 clusters based on the variation in expression trends (Figure 3a). Genes in Cluster 8 exhibited expression patterns that are highly positively correlated with the CA content of the four samples (i.e., WTY > WTO > U2Y > U2O), whereas genes in Cluster 7 showed an opposite trend. KEGG analysis (Figure S13; Table S10) revealed that the differentially expressed genes in these two clusters were mainly enriched in metabolism-related pathways, with significant changes in the terpene skeleton biosynthesis pathway (KO00900) and the diterpene biosynthesis pathway (KO00904). Therefore, it is possible to further explore similar expression patterns to screen for relevant genes involved in CA synthesis.

A total of 9225 and 10 279 genes were differentially expressed in older and younger leaves of WT and U2, respectively. Overall, 6290 (WT) and 6765 (U2) genes were expressed at lower levels in younger leaves, and only 2965 (WT) and 3514 (U2) genes were expressed at higher levels in younger leaves than in older leaves. In total, 511 genes with higher expression in young leaves encode transcription factors (TFs), which include members of the AP2/ERF, MYB, bHLH, and WRKY families. The expression patterns of these TFs were strongly correlated with the trend of CA accumulation (Figure 3b), suggesting that they might play key roles in CA biosynthesis in *S. rosmarinus*. CA is mainly found in photosynthetic tissues, namely the leaves, sepals, and petals, with the highest levels found in 3-month-old leaves (Luis & Johnson, 2005). To screen for genes more closely related to CA biosynthesis, genes in Cluster 8 were subjected to additional heatmap clustering analysis for their expression levels in roots, stems, leaves, and flowers (Figure 3c). The high-expression genes in young leaves were concentrated in class E (842 genes) and class A (188 genes), which contain several key genes reported to be involved in CA biosynthesis. Besides these previously reported CA-related genes, we have identified several other genes of the CYP71 family involved in diterpenoid synthesis, including *RoCYP77A27* (*Ro422590*), *RoCYP71AT90* (*Ro413350*), *RoCYP716A89* (*Ro053650*), *RoCYP77A28* (*Ro003870*), and *RoCYP77A27* (*Ro422590*), which are all potentially involved in the CA biosynthetic pathway in rosemary (Figure 3c). Notably, GO analysis of genes from these classes revealed that genes of class A were associated with response to jasmonic acid (JA) (Figure S14a). Recent studies have also shown that exogenous methyl jasmonate (MeJA) treatment of *S. rosmarinus* cells in suspension promotes CA accumulation (Yao et al., 2022). This suggests that the phytohormone JA could be involved in CA synthesis in *S. rosmarinus*.

Based on the results of the above analysis and previous reports related to CA biosynthesis, we propose a model for the CA biosynthesis pathway in *S. rosmarinus*. In *S. rosmarinus*, prerequisite substances for CA biosynthesis are obtained via the MVA pathway in the cytoplasm and the MEP pathway in the plastid, with *RoDX3* (*Ro0066070*), *RoDX9* (*Ro477120*), *GGPS2* (*Ro126690*), and *RoIDI2* (*Ro42940*) as key genes. Then abietariene is synthesized via the diterpene synthases RoCPS1 and RoKSL2, and the final steps of CA biosynthesis are mainly catalyzed by the cytochrome CYP76 family (*RoCYP76AK8* [*Ro291170*], *RoCYP76AK7* [*Ro087390*], and *RoCYP76AH22/FS1* [*Ro012750*]) and the CYP71 family (*RoCYP71BE52* [*Ro375970*], *RoCYP77A27* [*Ro422590*], *RoCYP71AT90* [*Ro413350*], *RoCYP716A89* [*Ro053650*], *RoCYP77A28* [*Ro003870*], and *RoCYP77A27* [*Ro422590*]). This process may be regulated by JA, which induces the expression of downstream CA biosynthetic genes via MYB, bHLH, and WRKY family-related TFs (Figure 3d and Figure S14b).

Conserved evolution and functional divergence of CA synthesis genes in *S. rosmarinus*

Diterpenoids such as CA and CO are the main antioxidant components in *S. rosmarinus*. According to current reports, the key genes for CA synthesis mainly belong to the terpene synthase (TPS) family and the CYP450 family (Božić et al., 2015), but these two gene families have not been systematically examined in *S. rosmarinus* due to the lack of genome sequence. Here, we identified 56 members of the TPS family and classified them into five subfamilies, a, b, c, e/f, and g (Figure S15b), according to the classification of TPS families reported in Arabidopsis. The TPS family analysis among the congeneric species revealed they had a similar total number of TPS families as diploid sage (*Salvia*) (Table S11). The TPS-a and TPS-b subfamilies have the highest numbers of genes. The TPS-a subfamily (22 genes) mainly includes genes encoding sesquiterpene synthases in plants, while members of the TPS-b subfamily (20 genes) have an R(R)X_8W motif that is specific to monoterpene synthases. The TPS-g subfamily is closely related to TPS-b; its members also encode monoterpene synthases, but lack the R(R)X_8W motif (Chen et al., 2011). The TPS-c and TPS-e/f subfamilies include genes that encode major components of diterpene synthases. The CPS and kaurene synthase (KS) could convert GGPP to copalyl diphosphate (CPP), which is the first step of CA synthesis (Božić et al., 2015). Through gene alignment and evolutionary tree analysis, the *RoCPS2* (*Ro209030*), which was previously reported to be involved in CA synthesis, belongs to the TPS-c subfamily (Figure S15b). *RoKSL1* (*Ro013350*) and *RoKLS2* (*Ro208980*) are classified in the TPS-e/f subfamily, which encodes KSs (Figure 4a). Based

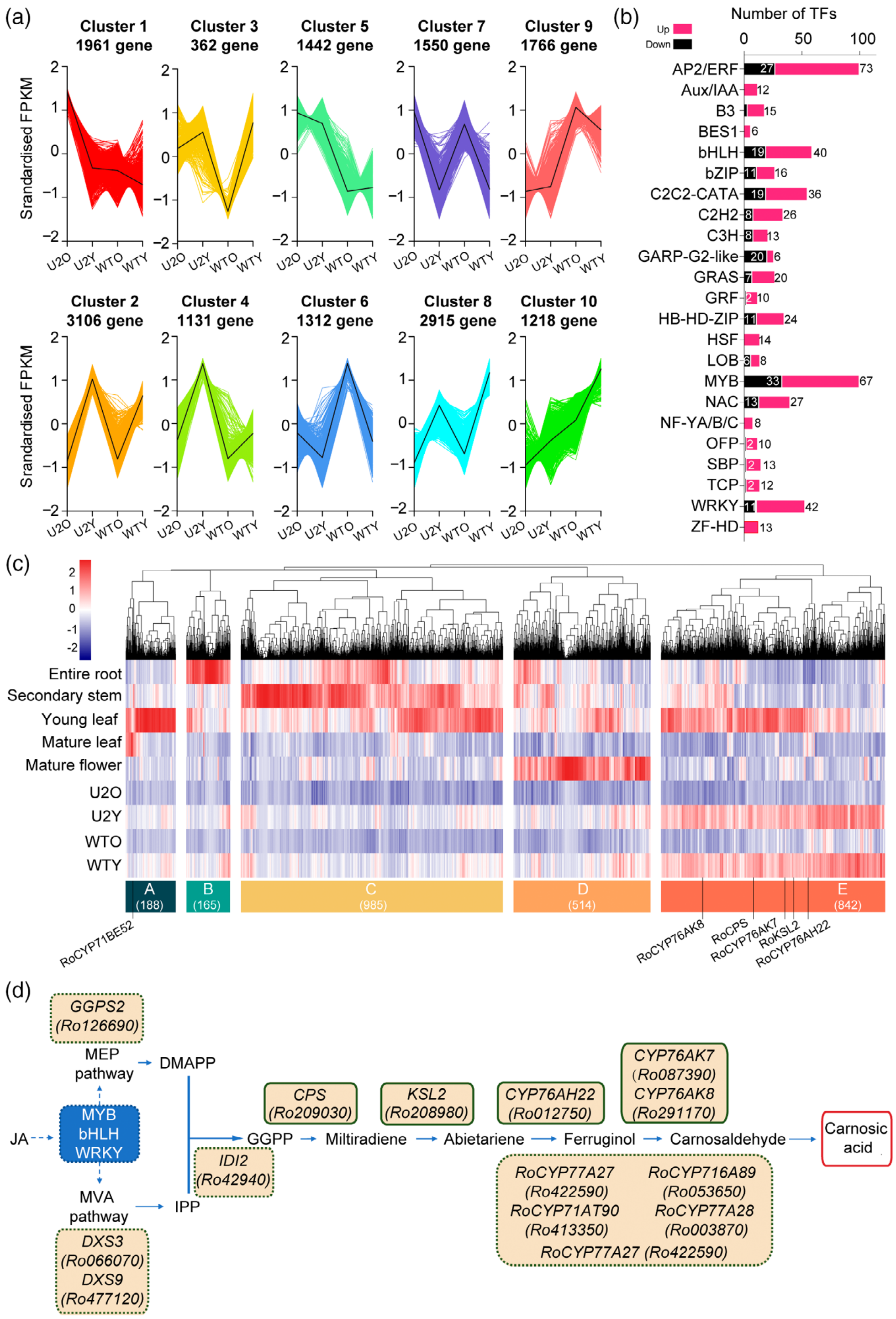

Figure 3. Analysis of differential gene expression in different tissues of *S. rosmarinus*. (a) Clustering analysis of four different transcriptomes, divided into 10 clusters based on trends, with different clusters shown in different colors. WTY, young leaves of wild type; WTO, old leaves of wild type; U2Y, young leaves of U2; U2O, old leaves of U2. (b) Bar graph representing the statistics of differentially expressed genes and TFs in young and old leaves in WT. Red, higher expression of genes in young leaves than in old leaves; blue, lower expression of genes in young leaves than in old leaves; rose, TFs whose expression is higher in young leaves than in old leaves; purple, TFs whose expression is lower in young leaves than in old leaves. (c) Heatmap clustering analysis of genes in Cluster 8 for screening their expression levels in different tissues of *S. rosmarinus*. The positions of CA synthesis-related genes are marked. (d) Model diagram of CA biosynthesis in hypothetical rosemary. The solid boxes indicate genes that have been identified, and the dashed boxes indicate genes predicted to be obtained.

on evolutionary and positional analysis of the genes, we found *RoKLS2* and *Ro208970* are derived from tandem duplication (Figure S15a). The CYP450 superfamily is also essential for diterpenoid synthesis in plants. CYP76AH and AK subfamily P450s reconstituted into yeast produce the terpene skeletal structure sub-tanshinone diene to synthesize CA (Ignea et al., 2016). Based on the conserved structural domains of P450 and related genes in Arabidopsis, we identified 341 *CYP450* gene members in *S. rosmarinus*, including 17 physical gene clusters of CYP450 (identified regions with more than five gene clusters per 500 kb), indicating a relatively high frequency of tandem repeats of *CYP* genes (Figures S16, S17).

To further explore the evolutionary relationship between CPS and KSL in the genus *Salvia*, a collinearity analysis was performed between *S. officinalis* and *S. bowleyan*. Interestingly, we found conserved CA-associated biosynthetic gene clusters (identified regions with genes synthesizing one metabolite-related enzyme per 500 kb). This biosynthetic gene cluster encodes three enzymes that synthesize CA (*CYP*, *KSL*, and *CYP76AH*) and is conserved in the genus *Salvia*; it is localized on chromosome 8 of *S. bowleyan* and on chromosome 1 of *S. officinalis*. But for *S. rosmarinus*, this biosynthetic gene cluster is rearranged and distributed on two different chromosomes (Figure 4a). By microsynteny analysis of biosynthetic gene clusters in *S. officinalis*, *S. bowleyan*, *S. miltiorrhiza*, and *S. rosmarinus*, two types of gene clusters were identified: *KSL1–CPS1–CYP76AHs* and *KSL2–CPS2–CYP76AHs*. Among them, one set of *KSL2–CPS2–CYP76AHs* was conserved in all four sage plants, while *KSL1–CPS1–CYP76AHs* were found only in one set in *S. officinalis*, *S. bowleyan*, and *S. miltiorrhiza* and in two sets in *S. rosmarinus* on chromosome 1 (*RoKSL1–RoCPS1–CYPAH76AH22* and *RoKSL1.1/Ro013350–RoCPS1.1/Ro013310–RoCYP76AH58/Ro013340*, Figure 4b). The evolutionary tree shows that gene divergence between *CPS1/CPS2* occurred prior to the divergence of the Labiatae (91.13 MYA) (Figure 4c), whereas gene divergence between *KSL1* and *KSL2* occurred after the divergence of the Labiatae, roughly at 36.84 MYA (Figure 4d). The CYP76AH family and CYP76AK family also diverged at approximately 85.27 MYA (Figure 4e). For *RoKSL1.1/Ro013350* and *RoCPS1.1*, their protein sequence comparison revealed a high sequence similarity with RoKSL1 and RoCPS1 (Figure S18). Phylogenetic analysis showed that these two genes were generated by tandem duplication after the divergence of *S. rosmarinus* and *S. officinalis* (Figure 4c). *RoCYP76AH22–RoCYP76AH24* and *RoCYP76AK7–RoCYP76AK8* are collinear gene pairs in *S. rosmarinus* (Figure S17). According to the phylogenetic tree analysis (Figure 4e), a divergence could have occurred in *RoCYP76AH22–RoCYP76AH24* during the recent WGD, and the *RoCYP76AK7–RoCYP76AK8* divergence occurred 44.07 MYA (second WGD). The function of RoCYP76AH58/Ro013304 has not been reported yet, and it may originate from the same ancestor as SoCYP76AH58/Saoff1g02124 in *S. officinalis* based on the evolutionary tree (Figure 4e).

To further explore the differences, we focused on the gene expression in different tissues and found that *RoCPS1–RoKSL1* and *RoCPS1.1–RoKSL1.1*, located on chromosome 1 in *S. rosmarinus*, were mainly expressed in the roots. In contrast, *RoKSL2–RoCPS2*, located on chromosome 2, were mainly expressed in young leaves (Figure 4f), which is coincident with the trend of CA accumulation (young leaves > old leaves > stem) (Figure S19). This suggests that the different types of *CPS* and *KSL* are specific for the downstream diterpenoid synthesis. CA is abundant in young leaves; therefore, it may be synthesized mainly through *RoKSL2–RoCPS2*. In contrast, rarely detectable accumulation of CA in the roots, *RoKSL1–RoCPS1*, and *RoKSL1.1–RoCPS1.1* may be responsible for the synthesis of other diterpenoids in the roots. Similarly, the carbonic acid synthase genes *RoCYP76AK7* and *RoCYP76AK8* were also highly expressed in young leaves. However, the biased differences in tissue expression were not significant for the CYP76AH family. *RoCYP76AH22* and *RoCYP65AH24* were highly expressed in both roots and young leaves, except for *RoCYPAH58* and *RoCYPAH59*, which were highly expressed in roots (Figure 4f). This difference may be caused by the production of different diterpenoids.

DISCUSSION

Salvia rosmarinus is a traditional spice and medicinal plant. It is rich in the antioxidant component CA. Moreover, CA can delay ageing, has powerful weight loss and fat reduction properties, and can be used for treatment of cardiovascular disease and cancer. *Salvia rosmarinus* is the richest source of CA and is therefore among the best plants for exploring CA biosynthesis (Richheimer et al., 1996; Schwarz & Ternes, 1992a). This is the first report of a chromosome-level assembly of the *S. rosmarinus* genome by using a combination of third-generation sequencing (PacBio), second-generation sequencing (Illumina), and Hi-C technologies. With a total of 12 chromosomes and a size of 1.11 G, *S. rosmarinus* has a larger genome than other species of the Labiataceae family, partially due to recent expansion of genes and LTRs in the *S. rosmarinus* genome. Notably, *S. rosmarinus* has experienced a total of three WGD events: besides the WGT event common to all dicotyledons (115–132 MYA) and the WGD event shared by all members of Labiataceae (32.7–53.7 MYA), *S. rosmarinus* underwent a lineage-specific WGD that occurred approximately 9.8–24.0 MYA (Figure 1).

Diterpenoid synthesis genes are conserved in the Labiataceae family

The biosynthesis of CA, a diterpenoid metabolite, has not been fully understood *in vivo* in *S. rosmarinus*. Previous

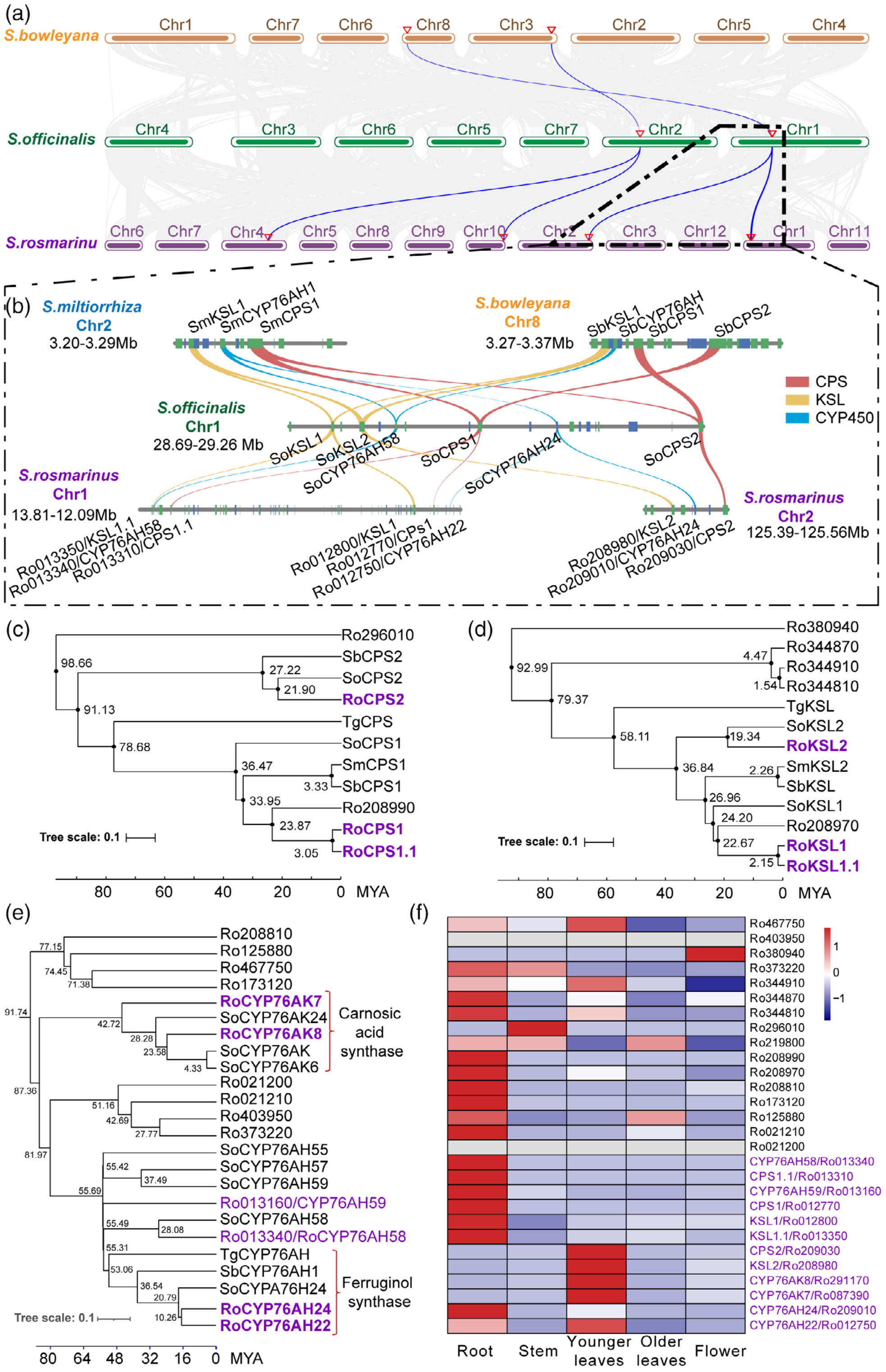

Figure 4. Analysis of *TPS* and *CYP450* gene families related to CA synthesis in sage. (a) The synteny of the CA-related genes in *S. officinalis*, *S. bowleyana*, and *S. rosmarinus*. The triangle indicates the location of the gene. The blue line indicates the synteny area. (b) Microsynteny of the *CPS*– *KSL*–*CYP76AH* in *S. rosmarinus*. Red line indicates CPS genes, yellow indicates KSL genes, and blue indicates *CYP450* genes. (c–e) Time evolution tree analysis. (c) *CPS* genes, (d) *KSL* genes, and (e) *CYP76A* genes in sage (Tg, *T. grandis*; So, *S. officinalis*; Sm, *S. miltiorrhiza*; Sb, *S. bowleyana*). (f) Heatmap of gene expression of *TPS* and *CYP450* gene families related to CA synthesis in *S. rosmarinus*.

studies described *CPS2*, *KSL1*, and *KSL2* belonging to the TPS family and the CYP71/76 subfamily belonging to the P450 family as the key regulatory genes of CA biosynthesis. We found that *CPS2*, *KSL2*, and *CYP76AH22*, which are involved in CA synthesis, were all localized on chromosome 2 in the *S. rosmarinus* genome (Figure 4b) and form a biosynthetic gene cluster. Through comparison with *S. bowleyana*, *S. miltiorrhiza*, and *S. officinalis*, we found that similar biosynthetic gene clusters are conserved in Labiataceae (Figure 4b). *RoCPS1–RoKSL1–RoCYP76AH22* and *RoCPS1.1–RoKSL1.1–RoCYP76AH58* are located on chromosome 1 and are highly expressed in roots. In contrast, *RoCPS2–RoKSL2–RoCYP76AH24* are located on chromosome 2 and are highly expressed in young leaves (Figure 4b,e; Figure S20). This suggests that these *CPS–KSL–CYP76AH* biosynthetic gene clusters might have different roles in the synthesis of downstream diterpene compounds. Previous studies have demonstrated the substrate specificities of RoKSL1 and RoKSL2 (Božić et al., 2015). It is possible that in *S. rosmarinus*, different CPS–KSL biosynthetic gene clusters might synthesize specific products at different sites. Current metabolite synthesis studies have mostly been validated *in vitro* with yeast or *Agrobacterium*, and therefore there may be limitations in studies exploring the functions of diterpenoid synthesis genes in plants. This issue can be further investigated by carrying out tissue culture, using transformation systems, and further verifying if the CYP76AH family division is also somewhat specific in *S. rosmarinus*. Gene clusters and their molecular regulatory mechanisms can be further explored in the future, and this information would be of great significance for the study of CA synthesis and metabolic engineering of *S. rosmarinus*.

A model pathway map for CA biosynthesis in *S. rosmarinus*

CA significantly accumulates in young *S. rosmarinus* leaves. Transcriptomic data from different tissues showed that the expression of key genes *RoCPS*, *RoKSL2*, and *RoCYP76* was significantly higher in young leaves than in old leaves (Figure 4f). By clustering the transcriptomic data, we identified 842 genes most associated with CA biosynthesis in *S. rosmarinus* (Figure 3c). Besides the already reported CA synthesis-related genes, we have identified other genes of the CYP71 family (*RoCYP77A27* [*Ro422590*], *RoCYP71AT90* [*Ro413350*], *RoCYP716A89* [*Ro053650*], *RoCYP77A28* [*Ro003870*], and *RoCYP77A27* [*Ro422590*]) involved in diterpenoid synthesis that may be involved in CA biosynthesis in *S. rosmarinus*. Interestingly, some of these are related to biosynthesis of JA, a phytohormone with essential roles in plant signal transmission (Yao et al., 2022) (Figure S14a). Recent studies have also indicated that exogenous MeJA can promote CA accumulation in cells of *S. rosmarinus* in suspension. In addition, MeJA can enhance the production of metabolites (Zhang et al., 2014); for example, JA-induced set-association reactions in tobacco (*Nicotiana benthamiana*) can increase nicotine biosynthesis (Shoji et al., 2010). In *S. miltiorrhiza*, treatment with MeJA can increase metabolite synthesis (Xiao et al., 2009). Thus, JA is critical for secondary metabolite synthesis in plants. Here we also propose that TFs in *S. rosmarinus* may regulate downstream genes, leading to increased CA synthesis by inducing a signaling cascade in response to JA. Furthermore, the expression of TFs screened in different rosemary tissues was mostly high in young leaves (Figure S14b). Meanwhile, further analysis based on sequencing of the NCBI database (PRJNA80069) revealed that JA could activate the expression of most of these TFs. Therefore, a model pathway map for CA biosynthesis was proposed based on the available analyses (Figure 3d). In summary, the present study reports a high-quality *S. rosmarinus* genome assembly that provides crucial insights for studies of breeding, evolution, and aromatic compound diversity in Lamiaceae.

EXPERIMENTAL PROCEDURES

Plant materials and genome sequencing

Two *S. rosmarinus* plants were collected from the Institute of Highland Forestry Science, Chinese Academy of Forestry: WT (Rosemary Morocco) and U2 (II-03-US from the USA). High-quality DNA extracted from young leaves was used for preparing PacBio, Illumina, and Hi-C sequencing libraries; samples for second-generation sequencing were collected from a mixture of *S. rosmarinus* roots, stems, and leaves. Three replicates of transcriptome sequencing were taken from young and old leaves of the same species of *S. rosmarinus*.

Plant tissues were removed and immediately frozen in liquid nitrogen. Samples collected were extracted for sequencing with high-molecular weight DNA using the cetyltrimethylammonium bromide method. Biomarker (Beijing, China) performed sequencing by constructing a SMRTbell library followed by analysis on a PacBio Sequel II sequencing platform. For Hi-C sequencing, genomic DNA was cross-linked with 1% formaldehyde and digested using restriction enzyme *Dpn*II to construct 300–500-bp Hi-C libraries. These libraries were subsequently sequenced on an Illumina NovaSeq 6000 platform at the Beijing Genomics Institute (Qingdao, China).

Genome survey

The *S. rosmarinus* genome size was estimated through flow cytometry and *k*-mer analysis. Referring to previous methods (Doležel et al., 2007; Doležel & Bartoš, 2005), the plant tissue was minced with a blade and the nuclear suspension was obtained through filtration. Using maize (*Zea mays*) as an internal reference, the suspension of the test sample and the internal reference suspension were proportionately mixed. The stained nuclear suspension samples were assayed on a BD FACSCalibur flow cytometer by excitation at 488 nm and detecting the light emitted by propidium iodide. In total, 10 000 particles were collected for each assay. The coefficient of variation (CV%) was controlled to within 5%. Graphical analysis was performed using Modifit 3.0 analysis software. For *k*-mer analysis, DNA sequencing data based on Illumina

NovaSeq 6000 were used to calculate the read 21-mer frequency using Jellyfish, and the results were visualized using GenomeScope 2.0 (Ranallo-Benavidez et al., 2020).

Genome assembly and chromosome construction

The raw PacBio SMRT data were rescued and assembled using Canu (v.1.7) (Koren et al., 2017) with the following parameters: canu -p Roff -d. genomeSize = 1 g -pacbio-hifi ccs.fasta.gz. Two haploid sets were merged into one using HaploMerger2 (Huang et al., 2017). Then, data from short reads were mapped to the assembled genome using BWA (Li, 2013), and the draft data were anchored to the chromosomes using 3D-DNA (Dudchenko et al., 2017). Hi-C data were processed and paired by Juicer and clustered into chromosomes by 3D-DNA. The obtained data were manually corrected and severed using Juicebox (Robinson et al., 2018) to obtain the final genome sequence.

Genome annotation

The LTRs were initially identified using ltrharvest and LTR_FINDER_-parallel (Ellinghaus et al., 2008; Ou & Jiang, 2019). Then, the identified LTRs were filtered and sorted using LTR_retriever (Ou & Jiang, 2019), and the time of LRT differentiation was further calculated. Gene structure annotation: Repeat sequences of the *S. rosmarinus* genome were masked using RepeatMasker (v.2.0.3). Mixed samples of *S. rosmarinus* root, stem, and leaf tissues were obtained for RNA-seq (Illumina NovaSeq platform). The obtained transcriptome clean reads by HISAT2 (v.2.2.0) were built to index and compare against the masked *S. rosmarinus* genome, and *de novo* prediction was performed using BRAKER2 software (Brůna et al., 2021). To give more credibility to the annotation, we obtained the gene structure based on the PASA process for homoprediction of the genome (Haas et al., 2003). Finally, the aforementioned two results were integrated using EVidenceModeler (EVM) (Haas et al., 2008). Structural annotation of the genome was performed through BUSCO evaluation (with default parameters) (Simão et al., 2015). Gene function was annotated using the online tool eggNOG-Mapper (Cantalapiedra et al., 2021). tRNAs were annotated using tRNAscan-SE (Chan & Lowe, 2019). Other rRNAs, miRNAs, and snRNAs were identified through comparison with the Rfam database.

Genome evolution analysis and divergence time estimates

Gene orthologs of 11 species, including published homologous and species-specific genes of Labiataceae and outgroups, were identified using OrthoFinder (Emms & Kelly, 2019). The single-copy sequences obtained were passed through mcmctree software (Dos Reis & Yang, 2013) to estimate divergence times under a relaxed clock model (JC69 model). Calibration times were obtained from the TimeTree database. Based on the obtained species tree and time tree, CAFE 5 (Mendes et al., 2021) was used to determine the dispersal times of gene families and the genes for expansion and contraction. Pairwise synteny and WGD events were analyzed using MCScan (Tang et al., 2008). WGD events were determined by calculating *Ka*/*Ks* using KaKs_calculator (Wang et al., 2010). The 4-fold synonymous third codon conversion (4DTv) rate was determined using a perl script (https://github.com/JinfengChen/Scripts/calculate_4DTV_correction.pl).

Identification and analysis of CA synthesis-related gene families

The gene families were identified using a combination of the HMMER protein structural domain and the BLAST protein sequence. Structural domains specific to the TPS family, according to the Pfam website, were PF03936.hmm and PF01397.hmm, and the structural domain specific to the P450 family was PF00067.hmm. Using an E-value of 0.01 for filtering, searches were performed with HMMER software based on the hmm files of the gene families. The protein sequences encoded by genes related to the MVA and MEP pathways in Arabidopsis were referenced from previous studies (Tholl & Lee, 2011). The protein sequences of the AtTPS family genes were also referenced from previous studies (Parker et al., 2014), and those of the CYP450 family genes in Arabidopsis were downloaded from the TAIR database (https://www.arabidopsis.org/browse/genefamily/p450.jsp). Protein sequences were searched using BLASTP, and sequences with E-values of <1e−5 were selected. Finally, the HMMER and BLASTP results were pooled to take the intersection and obtain the final family genes. Gene family phylogenetic trees were calculated using IQtree software (v. 2.2.0). Gene motifs were predicted using MEME software (v. 5.4.0). Gene distribution on chromosomes was marked using TBtools (Chen et al., 2020).

HPLC determination of CA and CO

Two steps were followed to determine CA levels through HPLC. The first step involved sample extraction. *Salvia rosmarinus* leaves were smashed using liquid nitrogen. Then, a proper amount of methanol (5–10 mL) was added to the smashed leaves. An ultrasonic extractor was used for extraction for 30 min. The supernatant obtained was filtered using a 0.22-μm nylon filter membrane. In the second step, HPLC detection was performed using an Accucore XL C18 (250 × 4.6 mm, 4 μm, Thermofisher) chromatographic column. The mobile phase was acetonitrile:0.1% trifluoroacetic acid (7:3), the flow rate was 1 mL/min, the injection volume was 10 μL, the column temperature was controlled at 30°C, the UV detection wavelength was 230 nm, and the detection time was 20 min. A standard curve of CA content was plotted using a CA standard sample (Sigma, Germany). First, 2 mg of the CA standard sample was dissolved in a constant volume of 1 mL and diluted to 10, 25, 50, 75, and 100 μg/mL. Then, 1.5 mL of each diluted sample was filtered using a 0.22-μm nylon filter membrane, and the filtrate was subjected to HPLC detection. The aforementioned method is a slightly modified version of that discussed in a previous study (Liu et al., 2011).

Transcriptome sequencing (RNA-seq) and RT-qPCR

Young and old leaves (young stage: leaf length 1–2 cm; old stage: leaf length 3–4 cm) were obtained from different *S. rosmarinus* plants of the same type (WT and U2), and RNA was extracted using the TRIzol method. The samples obtained were used to construct RNA-seq libraries using Illumina's NEBNext® Ultra™ RNA Library Prep Kit. After passing the library check, sequencing was performed on an Illumina HiSeq platform (Metware). Sequencing lengths of 125 bp/150 bp yielded double-end reads. The reads obtained were subjected to splice processing and low-quality base filtering with fastq. The filtered clean reads were compared to the assembled genome of *S. rosmarinus* by HISAT2. Genes were quantified and FPKM values were calculated using featureCounts (v1.6.2). Differential expression analysis ($|\log2(\text{fold change})| \geq 1$ with $P < 0.05$) between the two groups was then performed using DESeq2 (v1.22.1). GO and KEGG analyses were performed using the clusterProfiler package in R (Wu et al., 2021).

The obtained RNA was used to obtain cDNA using the reverse transcription kit (HiScript III 1st Strand cDNA Synthesis Kit, Vazyme #R312-01). qRT-PCRs were performed using 2× TSINGKE Master qPCR Mix (Tsingke, #T-TSE201).

Gene cloning, vector construction, and gene expression assays

RNA from *S. rosmarinus* tissue was extracted using the TRIzol method. Then, 1 μg of RNA was collected and reverse transcribed to cDNA using the HiScript III 1st Strand cDNA Synthesis Kit (Vazyme #R312) for gene cloning and qPCR. Sequences of the clones of TPS-related and CYP450-related genes are referenced to the CDSs after the genome assembly with primers listed in Table S12. The clones are ligated into the vector pZSC-YFP (driven by the 35S promoter) by the endonucleases *Nco*I and *Sal*I. Based on the information from the genome and sequencing data from the transcriptome, primers for related genes were designed (Table S12) to determine the expression levels of CA synthesis-related genes in stems and young and old leaves.

Protoplast preparation and transformation in *Arabidopsis thaliana*

Protoplasts were extracted from 4-week-old Arabidopsis leaves after referring to the published literature (Yoo et al., 2007). The protoplasts were then transformed using PEG buffer, incubated for 6 h at room temperature in the dark, and observed under a Zeiss LSM 800 confocal microscope.

ACKNOWLEDGMENTS

This work was supported by the Major Program of Guangdong Basic and Applied Research (2019B030302006), the National Natural Science Foundation of China (31871222), the Natural Science Foundation of Guangdong Province (2018B030308002, 2019A1515110330), the Guangdong Modern Agro-industry Technology Research System (2021KJ114), the Program for Changjiang Scholars, South China Normal University Young Teachers' Research Incubation Fund Project (21KJ18), and the China Postdoctoral Science Foundation (2022 M711217).

AUTHOR CONTRIBUTIONS

DLH, ZWH, GY, and XFZ performed the data analysis; DLH, WLL, and CFL performed the experiments; DLH, JBL, XFZ, LW, LSZ, and CWY prepared the manuscript. All authors read and approved the final manuscript.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

DATA AVAILABILITY STATEMENT

The raw genome sequence and raw transcriptome sequencing data are available in the NCBI Sequence Read Archive under BioProject ID PRJCA010408. The *S. rosmarinus* genome assemblies and annotations are available at figshare (https://doi.org/10.6084/m9.figshare.21443223.v1). Transcriptome sequencing of other tissues of *S. rosmarinus* and JA treatments from NCBI: PRJNA80069. The genome sequence numbers or reference links used in the article are as follows: *Tectona grandis* (Zhao et al., 2019), *Scutellaria barbata* (Xu et al., 2020), *Scutellaria baicalensis* (Xu et al., 2020), *Salvia bowleyana* (Zheng et al., 2021), *Salvia officinalis* (Li et al., 2022), *Salvia miltiorrhiza* (Song et al., 2020), *Salvia splendens* (Jia et al., 2021), *Origanum vulgare* (Bornowski et al., 2020), and *Callicarpa americana* (Hamilton et al., 2020).

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Estimation of genome size and heterozygosity using GenomeScope. The estimated heterozygosity of *S. rosmarinus* was 1.99% and the estimated genome size was 1.09 Gb. The K-mer size was 21.

Figure S2. Genome scaffolding of *S. rosmarinus* with Juicer. Hi-C data of *S. rosmarinus* were processed and aligned using Juicer and then clustered into 12 hypochromosomes using 3D-DNA.

Figure S3. Genome assembly assessment with BUSCO. The assembly data of *S. rosmarinus* were used to assess the integrity of the transcriptome and to compare it with the publicly available data from BUSCO (eudicots, eukaryotes, and Viridiplantae).

Figure S4. *Salvia rosmarinus* genome annotation statistics. (a) Pie chart indicating the proportion of different TF families in *S. rosmarinus*. (b) Bar chart showing the number of functions of *S. rosmarinus* annotated genes. (c) The numbers of non-coding RNA species annotated with *S. rosmarinus*.

Figure S5. Insertion time of LTRs in *S. rosmarinus* genomes. Copia family LTRs are shown in red and Gypsy family LTRs are shown in blue. The horizontal coordinate indicates the insertion time in millions of years (MYA), and the vertical coordinate indicates the count of LTR types.

Figure S6. KEGG and GO analyses of *S. rosmarinus* expanded and contracted orthogroups. (a) KEGG analysis of *S. rosmarinus* expanded and contracted orthogroups. (b, c) GO analysis of *S. rosmarinus* rapidly expanded (b) and contracted orthogroups (c).

Figure S7. Synteny and collinearity in *S. rosmarinus* genomes. (a) *Salvia rosmarinus*'s own paired synthetic visualization by WGDI. (b) Relative to the other eight species of the Labiataceae family *T. grandis*, *S. barbata*, *S. bowleyana*, *S. baicalensis*, *O. vulgare*, *S. officinalis*, and *C. americana*, with a 4-fold degenerate regression rate (4Dtv)

Figure S8. Collinearity in *S. rosmarinus* genomes in comparison with *V. vinifera*, *S. bowleyana*, *S. miltiorrhiza*, and *S. officinalis* genomes. (a) Collinearity of *S. rosmarinus* and *V. vinifera*. (b) Collinearity of *S. rosmarinus* and *S. bowleyana*. (c) Collinearity of *S. rosmarinus* and *S. miltiorrhiza*. (d) Collinearity of *S. rosmarinus* and *S. officinalis*. (4) Collinearity of *S. rosmarinus* and *S. splendens*.

Figure S9. Evolutionary and KEGG analyses of different types of duplexed genes in the *S. rosmarinus* genome. (a) Box plots of *Ka*/*Ks* density distributions for different replicate types of *S. rosmarinus*. WGD, whole-genome duplication; TD, tandem duplication; PD, proximal duplication; TRD, transposed duplication; DSD, scattered duplication. (b) *Ks*, *Ka*, and 4Dtv frequency distribution of different duplicated genes in the *S. rosmarinus* genome. (c) KEGG analysis of different duplicated genes in the *S. rosmarinus* genome. WGD, whole-genome duplication; TD, tandem duplication; PD, proximal duplication; TRD, transposed repeats; DSD, scattered repeats; 4Dtv, 4-fold degenerate regression rate.

Figure S10. GO analysis of different duplexed genes in the *S. rosmarinus* genome.

Figure S11. Quality control and expression profiling of the *S. rosmarinus* transcriptome. (a) Principal component analysis (PCA) of RNA-seq (b) Heatmap showing the expression of genes in different samples of *S. rosmarinus*.

Figure S12. Correlation analysis of genes involved in MVA and MVP pathways with CA accumulation. Spearman analysis of the

correlation between CA content of *S. rosmarinus* and genes related to the MVA and MEP pathways.

Figure S13. KEGG analysis of Cluster 7 and Cluster 8.

Figure S14. GO analysis of class clustering and expression levels of related TFs and genes in different tissues of *S. rosmarinus*. (a) GO analysis of different classes. (b) Expression levels of related TFs in different tissues of *S. rosmarinus*. (c) Expression levels of CA-associated TFs in different tissues of *S. rosmarinus*.

Figure S15. Distribution and genetic characterization of the TPS gene family in *S. rosmarinus*. (a) Distribution of the TPS gene family in *S. rosmarinus* chromosomes. Red connecting lines represent tandem repeat genes. (b) Analysis of the specific motif and gene structure of the TPS gene family.

Figure S16. The tree of the P450 gene family in *S. rosmarinus*. The four outermost circles show the expression profiles of P450 in different samples; the darker the blue, the higher the expression. The different colors of the fifth circle label the different subfamilies in P450. The red boxes indicate CA-related P450 genes.

Figure S17. Distribution of the P450 gene family in *S. rosmarinus*. The genes marked in red are genes of the CYP76 family. Connecting lines indicate tandem repeat genes, and straight lines indicate covariation between genes.

Figure S18. Sequence alignment of the KSL protein sequences. *S. rosmarinus*: RoKSL1 (Ro012800), RoKSL1.1 (Ro13310), RoKSL2 (Ro208980); *S. officinalis*: SoKSL1 (Saoff1g02120), SoKSL2 (Saoff1g02122); *S. miltiorrhiza*: SmKSL2 (EVM0012226); *S. bowleyan*: SbKSL (GWHTASIU041084).

Figure S19. Sequence alignment of the CPS protein sequences. *S. rosmarinus*: RoCPS (Ro012800), RoCPS1.1 (Ro13310), RoCPS2 (Ro208980); *Salvia officinalis*: SoCPS1 (Saoff1g02127), SoCPS2 (Saoff1g02137); *S. miltiorrhiza*: SmCPS1 (EVM0003235); *S. bowleyan*: SbCPS1 (GWHTASIU041087), SbCPS2 (GWHTASIU041091).

Figure S20. Detection of CA-related gene expression and CA content in different tissues of *S. rosmarinus*. (a) qPCR analysis was conducted to determine the expression levels of CA-related genes in different tissues of *S. rosmarinus*. (b) Determination of CA content in different tissue samples through HPLC. The red arrow indicates the position of the peak of CA.

Table S1. Prediction of *S. rosmarinus* genome size by flow cytometry.

Table S2. The 21-mer statistics of the *S. rosmarinus* genome.

Table S3. PacBio sequencing statistics.

Table S4. Statistics of the *S. rosmarinus* genome assembly.

Table S5. BUSCO statistics for contig-level assembly of the *S. rosmarinus* genome.

Table S6. Characterization of *S. rosmarinus* genes and the Labiataceae and model plants.

Table S7. Statistics of *S. rosmarinus* genome characterization of repeat genes.

Table S8. Determination of CA and CO levels in WT and U2 through LC.

Table S9. Transcriptome sequencing differentially expressed gene statistics.

Table S10. Transcriptomic clustering statistics for *S. rosmarinus*.

Table S11. Summary of *TPS* family genes in *S. rosmarinus*, *S officinalis*, *S. miltiorrhiza*, and *S. bowleyan*.

Table S12. Primers for gene amplification and qRT-PCR.

REFERENCES

Birtić, S., Dussort, P., Pierre, F.X., Bily, A.C. & Roller, M. (2015) Carnosic acid. *Phytochemistry*, **115**, 9–19.

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 Research*, **27**, dsaa055.

Božić, D., Papaefthimiou, D., Brückner, K., de Vos, R.C., Tsoleridis, C.A., Katsarou, D. *et al.* (2015) Towards elucidating carnosic acid biosynthesis in Lamiaceae: functional characterization of the three first steps of the pathway in *Salvia fruticosa* and *Rosmarinus officinalis*. *PLoS One*, **10**, e0124106.

Brückner, K., Božić, D., Manzano, D., Papaefthimiou, D., Pateraki, I., Scheler, U. *et al.* (2014) Characterization of two genes for the biosynthesis of abietane-type diterpenes in rosemary (*Rosmarinus officinalis*) glandular trichomes. *Phytochemistry*, **101**, 52–64.

Brůna, T., Hoff, K.J., Lomsadze, A., Stanke, M. & Borodovsky, M. (2021) BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics*, **3**, lqaa108.

Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. (2021) eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution*, **38**, 5825–5829.

Chan, P.P. & Lowe, T.M. (2019) tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. *Methods in Molecular Biology*, **1962**, 1–14.

Chen, C., Chen, H., Zhang, Y., Thomas, H.R., Frank, M.H., He, Y. *et al.* (2020) TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Molecular Plant*, **13**, 1194–1202.

Chen, F., Tholl, D., Bohlmann, J. & Pichersky, E. (2011) The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *The Plant Journal*, **66**, 212–229.

Doležel, J. & Bartoš, J. (2005) Plant DNA flow cytometry and estimation of nuclear genome size. *Annals of Botany*, **95**, 99–110.

Doležel, J., Greilhuber, J. & Suda, J. (2007) Estimation of nuclear DNA content in plants using flow cytometry. *Nature Protocols*, **2**, 2233–2244.

Dos Reis, M. & Yang, Z. (2013) MCMCTree tutorials.

Drew, B.T., González-Gallegos, J.G., Xiang, C.-L., Kriebel, R., Drummond, C.P., Walked, J.B. *et al.* (2017) Salvia united: The greatest good for the greatest number. *Taxon*, **66**, 133–145.

Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C. *et al.* (2017) De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science*, **356**, 92–95.

Ellinghaus, D., Kurtz, S. & Willhoeft, U. (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, **9**, 1–14.

Emms, D.M. & Kelly, S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, **20**, 1–14.

Gao, W., Hillwig, M.L., Huang, L., Cui, G., Wang, X., Kong, J. *et al.* (2009) A functional genomics approach to tanshinone biosynthesis provides stereochemical insights. *Organic Letters*, **11**, 5170–5173.

González-Minero, F.J., Bravo-Díaz, L. & Ayala-Gómez, A. (2020) Rosmarinus officinalis L.(Rosemary): An ancient plant with uses in personal healthcare and cosmetics. *Cosmetics*, **7**, 77.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K., Jr., Hannick, L.I. *et al.* (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, **31**, 5654–5666.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J. *et al.* (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology*, **9**, 1–22.

Hamilton, J.P., Godden, G.T., Lanier, E., Bhat, W.W., Kinser, T.J., Vaillancourt, B. *et al.* (2020) Generation of a chromosome-scale genome assembly of the insect-repellent terpenoid-producing Lamiaceae species, Callicarpa americana. *GigaScience*, **9**, giaa093.

Huang, S., Kang, M. & Xu, A. (2017) HaploMerger2: Rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. *Bioinformatics*, **33**, 2577–2579.

Ignea, C., Athanasakoglou, A., Ioannou, E., Georgantea, P., Trikka, F.A., Loupassaki, S. *et al.* (2016) Carnosic acid biosynthesis elucidated by a synthetic biology platform. *Proceedings of the National Academy of Sciences of the United States of America*, **113**, 3681–3686.

Jia, K.H., Liu, H., Zhang, R.G., Xu, J., Zhou, S.S., Jiao, S.Q. ***et al.*** (2021) Chromosome-scale assembly and evolution of the tetraploid Salvia splendens (Lamiaceae) genome. *Horticulture Research*, **8**, 177.

Kamli, M.R., Sharaf, A.A.M., Sabir, J.S. & Rather, I.A. (2022) Phytochemical Screening of Rosmarinus officinalis L. as a Potential Anticholinesterase and Antioxidant–Medicinal Plant for Cognitive Decline Disorders. *Plants*, **11**, 514.

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 Research*, **27**, 722–736.

Kumar, V., Marković, T., Emerald, M. and Dey, A. (2016) Herbs: Composition and dietary importance.

Li, C.Y., Yang, L., Liu, Y., Xu, Z.G., Gao, J., Huang, Y.B. ***et al.*** (2022) The sage genome provides insight into the evolutionary dynamics of diterpene biosynthesis gene cluster in plants. *Cell Reports*, **40**, 111236.

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* preprint arXiv:1303.3997. https://arxiv.org/abs/1303.3997

Liu, T., Sui, X., Zhang, R., Yang, L., Zu, Y., Zhang, L. ***et al.*** (2011) Application of ionic liquids based microwave-assisted simultaneous extraction of carnosic acid, rosmarinic acid and essential oil from Rosmarinus officinalis. *Journal of Chromatography A*, **1218**, 8480–8489.

Luis, J. & Johnson, C. (2005) Seasonal variations of rosmarinic and carnosic acids in rosemary extracts. Analysis of their in vitro antiradical activity. *Spanish Journal of Agricultural Research*, **3**, 106–112.

Mendes, F.K., Vanderpool, D., Fulton, B. & Hahn, M.W. (2021) CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics*, **36**, 5516–5518.

Munné-Bosch, S. & Alegre, L. (2001) Subcellular compartmentation of the diterpene carnosic acid and its derivatives in the leaves of rosemary. *Plant Physiology*, **125**, 1094–1102.

Ou, S. & Jiang, N. (2019) LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mobile DNA*, **10**, 1–3.

Parker, M.T., Zhong, Y., Dai, X., Wang, S. & Zhao, P. (2014) Comparative genomic and transcriptomic analysis of terpene synthases in Arabidopsis and Medicago. *IET Systems Biology*, **8**, 146–153.

Ranallo-Benavidez, T.R., Jaron, K.S. & Schatz, M.C. (2020) GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, **11**, 1–10.

Richheimer, S.L., Bernart, M.W., King, G.A., Kent, M.C. & Beiley, D.T. (1996) Antioxidant activity of lipid-soluble phenolic diterpenes from rosemary. *Journal of the American Oil Chemists' Society*, **73**, 507–514.

Robinson, J.T., Turner, D., Durand, N.C., Thorvaldsdóttir, H., Mesirov, J.P. & Aiden, E.L. (2018) Juicebox. js provides a cloud-based visualization system for Hi-C data. *Cell Systems*, **6**, 256–258.

Sasikumar, B. (2012) Rosemary. In: *Handbook of herbs and spices*, vol. **1**, 2nd edition, Sawston, UK: Woodhead Publishing, pp. 452–468.

Scheler, U., Brandt, W., Porzel, A., Rothe, K., Manzano, D., Božić, D. ***et al.*** (2016) Elucidation of the biosynthesis of carnosic acid and its reconstitution in yeast. *Nature Communications*, **7**, 1–11.

Schwarz, K. & Ternes, W. (1992a) Antioxidative constituents of Rosmarinus officinalis and Salvia officinalis. I. Determination of phenolic diterpenes with antioxidative activity amongst tocochromanols using HPLC. *Zeitschrift für Lebensmittel-Untersuchung und -Forschung*, **195**, 95–98.

Schwarz, K. & Ternes, W. (1992b) Antioxidative constituents of Rosmarinus officinalis and Salvia officinalis. II. Isolation of carnosic acid and formation of other phenolic diterpenes. *Zeitschrift für Lebensmittel-Untersuchung und -Forschung*, **195**, 99–103.

Shoji, T., Kajikawa, M. & Hashimoto, T. (2010) Clustered Transcription Factor Genes Regulate Nicotine Biosynthesis in Tobacco. *The Plant Cell*, **22**, 3390–3409.

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**, 3210–3212.

Song, Z., Lin, C., Xing, P., Fen, Y., Jin, H., Zhou, C. ***et al.*** (2020) A high-quality reference genome sequence of Salvia miltiorrhiza provides insights into tanshinone synthesis in its red rhizomes. *Plant Genome*, **13**, e20041.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M. & Paterson, A.H. (2008) Synteny and collinearity in plant genomes. *Science*, **320**, 486–488.

Tholl, D. & Lee, S. (2011) Terpene Specialized Metabolism in Arabidopsis thaliana. *Arabidopsis Book*, **9**, e0143.

Wang, C., Zhang, H., Wang, J., Chen, S., Wang, Z., Zhao, L. ***et al.*** (2020) Colanic acid biosynthesis in Escherichia coli is dependent on lipopolysaccharide structure and glucose availability. *Microbiological Research*, **239**, 126527.

Wang, D., Zhang, Y., Zhang, Z., Zhu, J. & Yu, J. (2010) KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics, Proteomics & Bioinformatics*, **8**, 77–80.

Wenkert, E., Fuchs, A. & McChesney, J.D. (1965) Chemical artifacts from the family Labiatae. *The Journal of Organic Chemistry*, **30**, 2931–2934.

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z. ***et al.*** (2021) clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation*, **2**, 100141.

Xiao, Y., Gao, S., Di, P., Chen, J., Chen, W. & Zhang, L. (2009) Methyl jasmonate dramatically enhances the accumulation of phenolic acids in Salvia miltiorrhiza hairy root cultures. *Physiologia Plantarum*, **137**, 1–9.

Xu, Z., Gao, R., Pu, X., Xu, R., Wang, J., Zheng, S. ***et al.*** (2020) Comparative Genome Analysis of Scutellaria baicalensis and Scutellaria barbata Reveals the Evolution of Active Flavonoid Biosynthesis. *Genomics, Proteomics & Bioinformatics*, **18**, 230–240.

Yao, D., Chen, Y., Xu, X., Lin, Y. & Lai, Z. (2022) Exploring the Effect of Methyl Jasmonate on the Expression of microRNAs Involved in Biosynthesis of Active Compounds of Rosemary Cell Suspension Cultures through RNA-Sequencing. *International Journal of Molecular Sciences*, **23**, 3704.

Yoo, S.D., Cho, Y.H. & Sheen, J. (2007) Arabidopsis mesophyll protoplasts: a versatile cell system for transient gene expression analysis. *Nature Protocols*, **2**, 1565–1572.

Zhang, S., Yan, Y., Wang, B., Liang, Z., Liu, Y., Liu, F. ***et al.*** (2014) Selective responses of enzymes in the two parallel pathways of rosmarinic acid biosynthetic pathway to elicitors in Salvia miltiorrhiza hairy root cultures. *Journal of Bioscience and Bioengineering*, **117**, 645–651.

Zhao, D., Hamilton, J.P., Bhat, W.W., Johnson, S.R., Godden, G.T., Kinser, T.J. ***et al.*** (2019) A chromosomal-scale genome assembly of Tectona grandis reveals the importance of tandem gene duplication and enables discovery of genes in natural product biosynthetic pathways. *Gigascience*, **8**, giz005.

Zheng, X., Chen, D., Chen, B., Liang, L., Huang, Z., Fan, W. ***et al.*** (2021) Insights into salvianolic acid B biosynthesis from chromosome-scale assembly of the Salvia bowleyana genome. *Journal of Integrative Plant Biology*, **63**, 1309–1323.