Genet Resour Crop Evol (2024) 71:4171–4181
https://doi.org/10.1007/s10722-024-01900-z

RESEARCH ARTICLE

Whole-genome resequencing identifies exonic single-nucleotide variations in terpenoid biosynthesis genes of the medicinal and aromatic plant common sage (*Salvia officinalis* L.)

Fatima Ceylan · Ayse Ozgur Uncu · Aysenur Soyturk Patat · Ali Tevfik Uncu

Received: 26 December 2023 / Accepted: 24 January 2024 / Published online: 18 February 2024

Abstract Common sage (*Salvia officinalis* L.), the type species of the genus *Salvia*, is a historically acknowledged medicinal and aromatic plant that is utilized in several different industries for manufacturing diverse end products, including food, pharmaceuticals, cosmetics, personal hygiene products and insect repellants. The medical uses of sage essential oil terpenoids have made these secondary metabolites a focus of medical/pharmaceutical chemistry research. In the present work, the common sage genome was resequenced and assembled, and the protein-encoding gene content was annotated. The terpenoid biosynthesis gene repertoire, which includes 75 terpene synthase and 67 terpenoid backbone biosynthesis pathway genes, was predicted and located on assembly scaffolds, revealing tandem duplication blocks on the chromosomes. Variant analysis identified 188 variable single-nucleotide loci in the coding sequences of sage terpenoid biosynthesis genes. A total of 24,570 single-nucleotide polymorphisms were identified in the common sage total exome, representing a database of potential variable loci for targeted genotyping research. Given that terpene synthase activity is highly prone to modulation by point mutations and that the genotype plays an important role in the complex traits of terpenoid composition, single-nucleotide polymorphisms located in coding sequences constitute candidate functional markers that can be associated with terpenoid compositional traits in future research.

Keywords Genome assembly · Lamiaceae · Mint · Secondary metabolite · Variant calling

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10722-024-01900-z.

F. Ceylan · A. O. Uncu
Department of Biotechnology, Necmettin Erbakan University, 42090 Meram, Konya, Turkey

A. Soyturk Patat · A. T. Uncu (✉)
Department of Molecular Biology and Genetics, Necmettin Erbakan University, 42090 Meram, Konya, Turkey
e-mail: atuncu@erbakan.edu.tr

Introduction

Common sage (*Salvia officinalis* L.) (2n = 2x = 14) is the type species for the genus *Salvia* L., which constitutes the largest genus of sages (Lamiaceae), with approximately 1000 species (Kriebel et al. 2019). This genus is distributed worldwide and includes several species historically known and used for their medicinal properties and are currently utilized in the pharmaceutical, food and cosmetic industries. Common sage, native to the Mediterranean, is among those *Salvia* species of industrial importance and is rich in essential oil constituents with potent bioactive compounds. In addition to its prominent medicinal potential, this aromatic plant has long been used as

a spice and in herbal beverage (Sarrou et al. 2017; Li et al. 2022). The medicinal properties of *S. officinalis*, including anticancer, antimicrobial, anti-inflammatory and antioxidant effects, have been reported in several investigations, and these health-beneficial biological activities are commonly attributed to terpenoids extracted from sage essential oils (Chalchat et al. 1998; Ali et al. 2017; Li et al. 2022).

Common sage plants are rich in terpenoids, and species specific, 'specialized' terpenoids accumulate throughout the plant (Li et al. 2022). Plant terpenoids are produced in both vegetative tissues and reproductive organs (as well as roots) and constitute the largest class of plant secondary metabolites (Dudareva et al. 2004). The number of distinct terpenoids, including specialized terpenoids and terpenoid-linked moieties, reaches thousands. Plant terpenoids have numerous functions. They act as hormones, protein modification agents, and antioxidants; take part in electron transport systems; determine membrane fluidity; attract pollinators and mycorrhizae; and initiate defense responses. As a component of traditional medicine used as a remedy since ancient times, plant terpenoids are used today in mouth washes, cough medicines and disinfectants. Several terpenoids (e.g., Taxol and vinblastine) are used as chemotherapy agents, for the treatment of malaria and parasitic worm infections (e.g., artemisinin) and for the synthesis of cortisone, progesterone and medicinal steroids (e.g., diosgenin) (Pichersky and Raguso 2018). Thus, terpenoids have long been a focus of medicinal chemistry and the pharmaceutical industry. The gene repertoire involved in terpenoid biosynthesis in the *Salvia* genus is being studied and identified through transcriptome and genome analyses (Xu et al. 2016; Ali et al. 2017; 2018; Wimberly et al. 2020; Li et al. 2022; Bryson et al. 2023; Han et al. 2023). In related work, a draft genome of *Salvia miltiorrhiza* was assembled and used for the identification of 82 terpene synthase genes (Xu et al. 2016). In their work, the authors reported that interspecific, nonsynonymous single-nucleotide polymorphisms were highly abundant in terpenoid biosynthesis pathway genes. A total of 69 terpene synthase-encoding genes were identified by Li et al. (2022) by sequencing the common sage genome, and 67 terpene backbone biosynthesis genes were identified as a result of work in which the *Salvia rosmarinus* genome was analyzed (Han et al. 2023). These investigations also highlighted the importance of gene duplication events in diversifying the genes involved in terpenoid biosynthesis. A de novo leaf transcriptome assembly of common sage was constructed and used to identify terpenoid biosynthesis genes. Sixty-five putative terpene synthases (monoterpenes, sesquiterpenes, diterpenes and triterpenes) and 70 genes putatively involved in the mevalonate (MVA) and methyl-erythritol phosphate (MEP) pathways and terpenoid backbone biosynthesis were identified (Ali et al. 2017). Similar work with the blue anise sage (*Salvia guaranitica* L.) leaf transcriptome resulted in the identification of 143 candidate genes, 69 of which were terpene synthases (monoterpenes, sesquiterpenes, diterpenes and triterpenes) and 74 of which were involved in the mevalonate (MVA) and methyl-erythritol phosphate (MEP) pathways and terpenoid backbone biosynthesis (Ali et al. 2018). A total of 30 terpene synthase-encoding assembled transcripts were identified from chia (*Saliva hispanica* L.) leaf and root transcriptome analysis (Wimberly et al. 2020).

The essential oil composition constitutes a key trait in breeding and genetic characterization research of medicinal and aromatic plant species such as *S. officinalis*. *S. officinalis* has a high level of diversity in its essential oil composition, as reflected by the differential expression of the major terpenoid constituents, namely, cineole, thujones, pinene, camphor, camphene, borneol and bornyl acetate (Jug-Dujakovic´ et al. 2012; Sarrou et al. 2017; Schmiderer et al. 2023). Despite the environmental influence and growth-stage/tissue specificity of metabolite production, the sage terpenoid composition is determined mainly by genotypic variation (Jug-Dujakovic´ et al. 2012; Sarrou et al. 2017). The genotype dependency of terpenoid composition manifests itself by the feasibility of discriminating genotypes grown in the same environment according to their essential oil compositions. Thus, determining allelic variation in terpenoid biosynthesis-associated loci is important for functional genomics of the *Salvia* genus and would be useful for the targeted genetic improvement of metabolic characters of industrial significance. Accordingly, resequencing of the common sage (*S. officinalis*) genome was performed in the present study with the aim of defining terpenoid biosynthesis genes and identifying variable nucleotide loci in the gene set as potential functional markers. A reference-guided genome assembly comprising seven scaffolds

was constructed, and putative terpenoid biosynthesis genes were identified through structural and functional gene annotation analyses. The predicted genes were assigned to their locations on *S. officinalis* assembly scaffolds. Further work involved the analysis of single-nucleotide variants within the terpenoid biosynthesis gene set as well as the identification of variations in the total predicted *S. officinalis* exome sequence.

Materials and methods

Genome sequencing and assembly

Common sage (*S. officinalis*) leaf tissue (researchers' material) was obtained from Selcuk University, Department of Horticulture (Konya, Turkey). A CTAB protocol modified from Doyle and Doyle (1990) and described below was used for extracting DNA from the leaf material. Accordingly, 200 mg of leaf tissue was ground in liquid nitrogen, mixed with 800 µL of CTAB extraction buffer [100 mM Tris–HCl (pH 8.0), 20 mM EDTA (pH 8.0), 1.4 M NaCl, 2% (w/v) CTAB, 1% PVP] and 5 µL of RNase A (0.01 mg/µL), and incubated at 65 °C for 1 h. Then, 600 µL of chloroform:isoamyl alcohol (24:1) was added to the lysed sample, and the sample was centrifuged at 20,900 × g for 10 min. The supernatant was then incubated at 4 °C for 1 h with 200 µL of isopropanol for DNA precipitation. Following incubation, the precipitated DNA was pelleted by centrifugation (4 °C for 10 min at 20,900 × g) and washed in 100 µL of 70% ethanol. The DNA was resuspended in 100 µL of Tris–EDTA buffer (pH 8.0). Three replicate extractions were performed and combined for sequencing. Paired-end sequencing with an Illumina NovaSeq 6000 platform was obtained from Gen Era Diagnostics (Ankara, Turkey). A TruSeq DNA PCR-Free Kit was used according to the TruSeq DNA PCR-Free Sample Preparation Guide for preparing a sequencing library with a median insert size of 450 bp. A total of 151 bp paired-end reads were produced by sequencing. k-mer analysis was performed prior to the assembly process using KmerGenie (Chikhi and Medvedev 2014). Paired-end reads were assembled into contigs with the Minia assembly algorithm version 3 (Chikci and Rizk 2013). The contig assembly was uploaded to the Galaxy platform (https://usegalaxy.eu/) (Afgan et al. 2018) for further analyses with built-in bioinformatic tools. Contigs were scaffolded using the RagTag software toolset version 2.1.0 (Alonge et al. 2022). Sage chromosome sequences accessed at https://ngdc.cncb.ac.cn/gwh under the ID GWHBJVP00000000 were used as references for placing and ordering contigs while scaffold building. RepeatMasker version 4.1.2 (Smit et al. 2015) was used for softmasking the scaffold assembly. The gene content completeness of the assembly was analyzed with BUSCO (Benchmarking Universal Single-Copy Orthologs) version 5.3.2 (Simão et al. 2015). The software was run using the lineage dataset embryophyta_odb10 (creation date: 2020–09-10; number of genomes: 50; number of BUSCOs: 1614).

Structural and functional gene annotation

Identification of putative protein-encoding gene structures in the scaffolds was performed using AUGUSTUS version 3.4.0 (Stanke et al. 2008). Gene structures on opposite strands were predicted independently to avoid neglecting overlapping genes. The evolutionary genealogy of genes: Nonsupervised Orthologous Groups (eggNOG) Mapper (database version 5.0.2) (Huerta-Cepas et al. 2017) was used with taxonomic scope 71,274 (asterids) to verify the predicted gene models and assign functions to the putative gene structures. The resulting gene structures were further analyzed to identify terpenoid biosynthesis genes. Homology analysis was performed with the terpenoid metabolism coding sequence records for *Salvia* species (GenBank Accession 869,088–894017). The blastx algorithm was used with an E-value threshold of 1E-10. The InterProScan tool (version 95.0) (Paysan-Lafosse et al. 2023) was used for functional classification and verification of the putative homologs. Graphical visualization of the physical locations of the identified genes was performed using MapChart (version 2.32) (Voorrips 2002).

Calling exonic single-nucleotide polymorphisms (SNPs) from the *S. officinalis* exome and terpenoid biosynthesis genes

Variant calling was performed with the DNAdiff wrapper, which includes the nucmer aligner (Marçais et al. 2018). The entire exome of *S. officinalis*

and the coding sequences of the set of putative terpenoid biosynthesis genes identified through structural and functional annotation were compared to available common sage genome sequences accessed at https://ngdc.cncb.ac.cn/gwh under the ID GWHBJVP00000000. The identified single-nucleotide variants were filtered for loci flanked by at least 20 bp of nonvariable, perfectly aligned bases on each side of the variable nucleotide. To visualize the locations of the single-nucleotide variations within peptide domains and evaluate potential amino acid alterations, domain searches were performed on peptide sequences encoded by the open reading frames (ORFs) of the putative terpenoid biosynthesis genes using the InterProScan tool (Paysan-Lafosse et al. 2023).

Results and discussion

Genome sequencing and assembly

Genome resequencing has led to a significant paradigm shift in plant breeding and functional genomics. Identification of single-nucleotide variants at the whole-plant genome scale has become the most efficient approach for developing genome-wide SNP maps with physical coordinates, improving the efficiency and precision of marker-assisted breeding, and increasing the quantity and accuracy of genotype calls for population genetic research (Lu et al. 2020; Song et al. 2023). In the present work, resequencing was performed in an attempt to identify and locate terpenoid biosynthesis genes in the common sage genome and define single-nucleotide polymorphisms within the coding sequences as potential functional markers for further work toward understanding the genotype-based differences in terpenoid composition and possible fine tuning of metabolic phenotypes. Accordingly, a total of 118.8 Gb of raw data were produced as a result of paired-end sequencing. The reads are deposited as sequence read archive (SRA) files in GenBank under BioProject PRJNA995944. After adapter trimming and filtering the bases according to a phred quality score of 30 (Q30), 106 Gb of data were retained, corresponding to 89.2% of the raw sequence reads. k-mer analysis of the sequence data identified k = 117 as the optimal value for operating a de Bruijn graph assembler (Fig. S1). Adapter-trimmed reads that passed the Q30 filter were initially assembled into contigs, and the subsequent reference-based scaffolding process assembled the contigs into seven scaffolds (Table 1). A scaffold assembly of seven pseudomolecules encompassing 431.5 Mb was produced, corresponding to 90.8% of the estimated genome size (475.3 Mb) (Maksimović et al. 2007) (Table 1). The basic assembly statistics are listed in Table 1. The scaffolds can be accessed in FASTA format at https://doi.org/https://doi.org/10.6084/m9.figshare.24099927 (Scaffolds_1-7.fasta) (Ceylan 2023). Prior to the annotation of protein-encoding gene loci, the completeness of the gene content was assessed quantitatively. Because genome resequencing was performed to identify potential variable loci in terpenoid biosynthesis coding sequences, it was important to achieve a high degree of gene completeness. Thus, quality assessment based on the evolutionarily expected set of single-copy orthologous genes was essential for evaluating the completeness of the gene repertoire and ensuring that false segmental duplications are minimized (Kelley and Salzberg 2010). According to the BUSCO analysis results, 92.7% of the 1614 BUSCO groups in the Embryophyta dataset were present in the assembly scaffolds, with 1413 complete (87.6%) and 82 fragmented (5.1%) BUSCOs. A total of 98.1% of the complete BUSCOs were identified as

Table 1 Basic assembly statistics

Assembly statistics	
Number of scaffolds	7
Pseudomolecule 1 (bp)	88,498,997
Pseudomolecule 2 (bp)	77,009,490
Pseudomolecule 3 (bp)	60,691,335
Pseudomolecule 4 (bp)	53,566,787
Pseudomolecule 5 (bp)	52,731,730
Pseudomolecule 6 (bp)	54,576,965
Pseudomolecule 7 (bp)	44,470,283
Total assembly size (bp)	431,545,587
GC (%)	35.36
AT (%)	64.6
N50 value (bp)	60,691,335
NG50 value (bp)	54,576,965
Complete BUSCOs	1413
Complete and single-copy BUSCOs	1386
Complete and duplicated BUSCOs	27
Fragmented BUSCOs	82

being single-copy (Table 1). High rates of duplicated BUSCOs imply misassembly-associated false duplications in gene content. Nevertheless, the BUSCO analysis results indicated high gene completeness and a minimal number of duplicated BUSCOs (Table 1). The BUSCO IDs, descriptions and locations of the identified BUSCOs in the assembly scaffolds are provided in Supplementary Table S1.

Protein-encoding gene content prediction and identification of terpenoid biosynthesis genes

HMM-based (hidden Markov model-based) protein-encoding gene content identification was performed using AUGUSTUS (Stanke et al. 2008). Putative gene structures and features in gff format can be accessed at https://doi.org/https://doi.org/10.6084/m9.figshare.24100032 under the file name CodingGeneStructures.gff3. Functional annotation via eggNOG (Huerta-Cepas et al. 2017) verified both the HMM-predicted gene structures and the transferred functions from matching orthologous groups. The functional annotation yielded 21,443 putative protein-encoding genes in the sage assembly scaffolds. Of the 21,443 orthologous genes, 20,191 were assigned to 124 Clusters of Orthologous Genes (COG) categories, with 948 genes assigned to multiple categories (Table S2). A total of 9421 genes were assigned to the COG category 'S', which defines genes identified as members of an orthologous group with unknown function (Table S2). The distribution of the most common COG assignments is shown in Fig. 1. The coding sequences and encoded peptides are deposited at https://doi.org/https://doi.org/10.6084/m9.figshare.24100083 under the file names Gene_Structures_Nucleotide.fasta and Gene_Structures_Peptide.fasta, respectively.

Putative terpenoid biosynthesis genes were specifically identified in the set of protein-encoding orthologous genes via homology analysis with the collection of available *Salvia* terpenoid biosynthesis genes (terpenoid backbone biosynthesis pathway genes and terpene synthase genes) and functional analysis of the homologous peptides based on peptide domain signatures using InterProScan (Paysan-Lafosse et al. 2023). A total of 142 putative terpenoid biosynthesis genes were identified, including 75 terpene synthase and 67 terpenoid backbone biosynthesis pathway genes

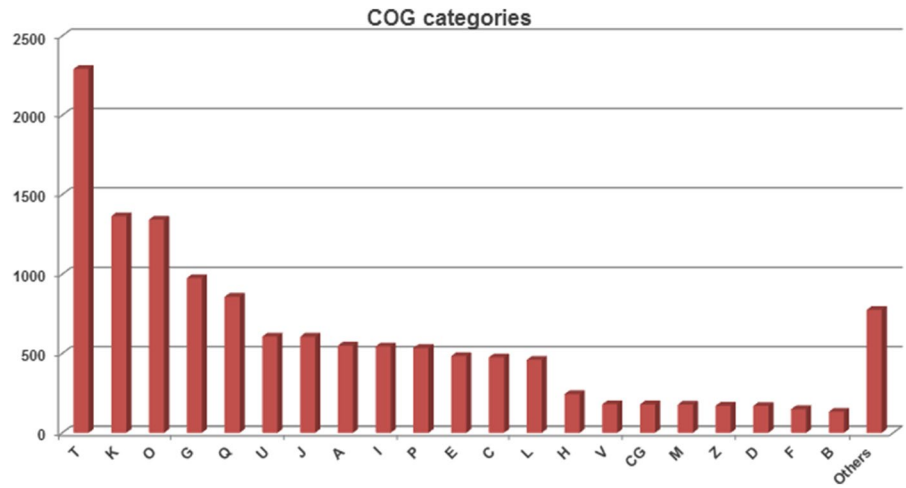

Fig. 1 Clusters of Orthologous Genes (COG) category assignment distribution of functionally annotated gene structures. COG categories with defined functions are included. T, Signal transduction mechanisms; K, Transcription; O, Posttranslational modification, protein turnover, chaperones; G, Carbohydrate transport and metabolism; Q, Secondary metabolite biosynthesis, transport and catabolism; U, Intracellular trafficking, secretion, and vesicular transport; J, Translation, ribosomal structure and biogenesis; A, RNA processing and modification; I, Lipid transport and metabolism; P, Inorganic ion transport and metabolism; E, Amino acid transport and metabolism; C, Energy production and conversion; L, Replication, recombination and repair; H, Coenzyme transport and metabolism; V, Defense mechanisms; M, Cell wall/membrane/envelope biogenesis; Z, Cytoskeleton; D, Cell cycle control, cell division, chromosome partitioning; F, Nucleotide transport and metabolism; B, Chromatin structure and dynamics. COG assignments representing at least 100 genes are presented in the graph as individual bars. COG assignments with fewer than 100 genes are corepresented as the 'Others' bar

(MVA and MEP pathways and other terpenoid backbone biosynthesis genes). The functional descriptions, pathway and gene ID information, and locations of the 142 putative terpenoid biosynthesis genes on the assembly scaffolds are provided in Table S3. The spliced coding sequences of the putative gene set and encoded peptide sequences can be accessed at https://doi.org/https://doi.org/10.6084/m9.figshare.24100134 under the file names TerpenoidBiosynthesis_Coding.fasta and TerpenoidBiosynthesis_Peptide.fasta, respectively. The detailed coordinates of the identified terpenoid biosynthesis genes in the assembly scaffolds, including the coordinates of the splice junctions and coding strand (±) information, are available in the structural gene annotation file (CodingGeneStructures.gff3) https://doi.org/https://doi.org/10.6084/m9.figshare.24100032.

In the course of the present work, genome-wide identification of terpenoid biosynthesis genes was performed using structural gene models located on genomic scaffolds, and open reading frames, including the locations of exon−intron junctions, were extracted. As a result, it was also feasible to determine the interchromosomal distribution of the putative terpenoid biosynthesis genes.

Terpene synthases constitute one of the most significantly expanded secondary metabolism gene families over the course of Embryophyta evolution. Both tandem duplications and whole-genome duplications are responsible for the expansion of the terpene synthase gene family in plants (Li et al. 2021). There are approximately 30 to 100 terpene synthase-encoding genes in the plant genome (Pichersky and Raguso 2018). In the present work, terpenoid biosynthesis-associated genes were found to be distributed across all seven chromosomes, and there were tandem blocks of genes encoding both terpene synthases and terpenoid backbone biosynthesis pathway components (Fig. 2 & Table S3). The interchromosomal distribution of the genes, as shown in Fig. 2, was not uniform, as the majority of the genes were aggregated on four chromosomes, Chr 1, 2, 3 and 6 (Fig. 2).

In a recent work, 342 terpenoid biosynthesis-associated genes were identified as 126 tandem blocks in the genome of lavender (L*avandula angustifolia*), an important Lamiaceae species that belongs to the Nepetoideae subclade together with the members of the *Salvia* genus; that study reported a nonuniform distribution of terpenoid biosynthesis genes among the chromosomes (Li et al. 2021), similar to the findings of the present work. In the present study, tandem monoterpene synthase and diterpene synthase blocks were found to be most common on chromosome 2 and on chromosomes 1 and 7, respectively

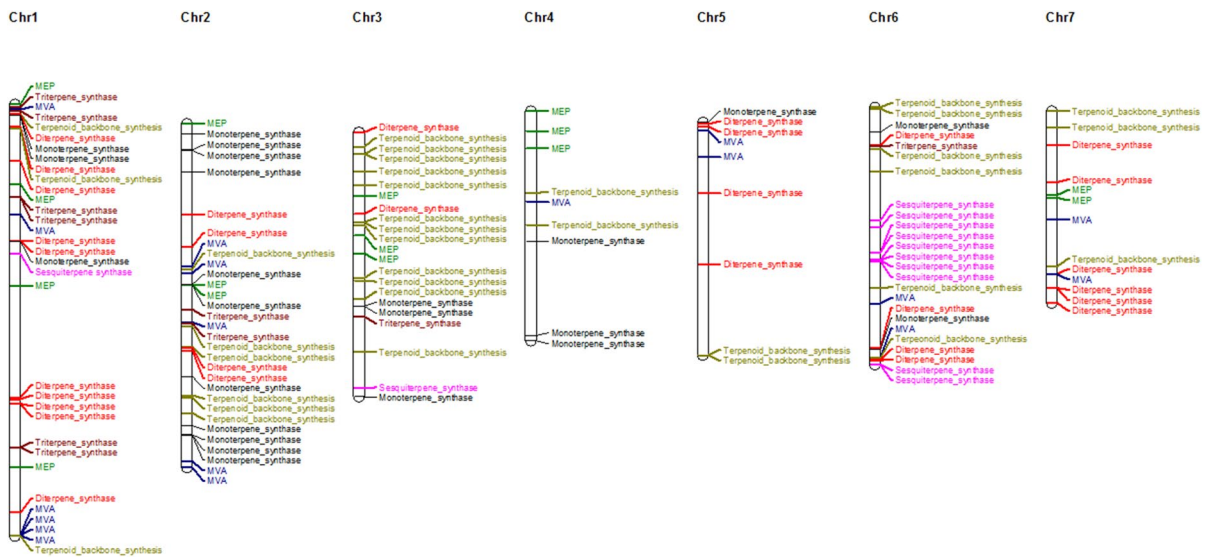

Fig. 2 Genomic distribution of the common sage putative terpenoid biosynthesis gene repertoire. The exact locations of the loci on the assembly scaffolds are provided in Table S3. The detailed gene features can be accessed at https://doi.org/10.6084/m9.figshare.24100032 under the file name CodingGeneStructures.gff3

(Fig. 2). Blocks of terpenoid backbone biosynthesis genes, including MVA and MEP pathway genes, were most commonly found on chromosomes 1–4 (Fig. 2). Clustering of diterpene synthesis-associated genes on chromosome 1 (Fig. 2) was in concordance with previous work in which the common sage genome sequence was analyzed for diterpene biosynthesis genes (Li et al. 2022). Similar to the results obtained in the present work (Table S3), analysis of the *Salvia rosmarinus* genome by Han et al. (2023) revealed 67 terpenoid backbone biosynthesis genes. The results of the work with the *S. officinalis* and *S. rosmarinus* genomes emphasized the effect of gene duplications as a major driver of the evolution of genes associated with terpenoid production in the *Salvia* genus.

SNP variants in the predicted total exome and terpenoid biosynthesis genes of *S. officinalis*

The functional diversification of duplicate copies within tandem gene blocks via mutation and natural selection significantly contributes to the tremendous diversity of plant terpenoids, leading to the most extensive class of plant secondary metabolites (Trindade et al. 2018; Li et al. 2021). Certain features of terpene synthase enzymes are effective at determining the scale of the effect of mutations on plant terpene diversity. Terpene synthases usually produce multiple products from the same substrate due to stochastic charge migrations in the enzyme active site, and the effect of a single amino acid change in a terpene synthase is usually reflected by the production of multiple new terpene end products. As a result, mutations that alter even a single amino acid in a terpene synthase frequently have significant outcomes (Pichersky and Raguso 2018). Modification of the terpene skeleton by a variety of reactions certainly further expands the diversity of the end products generated by terpene synthase. Nevertheless, it should be noted that the effect of a mutation in an enzyme-encoding gene may also manifest itself at the level of expression control, resulting in changes in the spatial or temporal production rate of the enzyme (Cartegni et al. 2002; Nielsen et al. 2007; Komar 2007; Tuller et al. 2010). Consequently, identifying sequence variations in terpene synthase coding sequences as potential tools for associating certain alleles with certain favorable metabolic phenotypes and/or with increased production of desirable terpene end products is valuable. In a relevant study, analysis of the genome sequence of *S. miltiorrhiza* for single-nucleotide polymorphisms revealed a high rate of nonsynonymous substitutions in diterpenoid biosynthesis-associated genes (Xu et al. 2016). To develop molecular markers for linkage mapping, genetic association and marker-assisted selection of rubber production traits in rubber tree (*Hevea brasiliensis*), Mantello et al. (2014) identified 78 SNPs in MVA and MEP pathway genes. Külheim et al. (2011) identified SNPs in MVA and MEP pathway genes and found significant associations between these SNPs and the terpene production traits of *Eucalyptus globulus*. Identification of the potential variability in the terpenoid biosynthesis gene repertoire of *S. officinalis* has not been previously reported; therefore, the availability of an *S. officinalis* genome sequence (Li et al. 2022) was critical to the purpose of the present work, as we were able to perform comparative sequence analysis for the identification of genomic nucleotide variations.

In the course of the present work, a total of 447 substitutions were identified with respect to the reference sequence in the exons of the putative terpenoid biosynthesis genes. After filtering for loci with at least 20 bp of perfect alignment (100% match with no sequence divergence) on each side, a total of 188 single-nucleotide polymorphisms were retained in 82 unique terpenoid biosynthesis genes. The detailed results of variant calling are provided in Table S4. Transition mutations were predominant, representing 64.9% of the total number of SNPs, as expected. Transversions constituted 35.1% of the SNPs. Purine/purine transitions (A/G, G/A) were the most common type of substitution, representing 35.1% of the 188 SNP loci (Table 2). Single-nucleotide polymorphisms in the entire predicted *S. officinalis* exome were also determined by variant calling. In total, 53,143 variable nucleotide loci were identified (https://doi.org/https://doi.org/10.6084/m9.figshare.24100155, TotalExome_substitutions.tabular), and after filtering for perfect alignment of the flanking 20 bases, 24,570 single-nucleotide polymorphisms remained (https://doi.org/https://doi.org/10.6084/m9.figshare.24100155, TotalExome_gSNPs). Similar to the substitutions in the terpenoid biosynthesis genes, in the total exome of *S. officinalis*, transition mutations constituted almost 65% of the total number of single-nucleotide substitutions, as expected (Table 2). InterProScan (Paysan-Lafosse et al. 2023) analysis was carried

Table 2 Variant calling statistics for the *Salvia officinalis* total exome and the predicted set of terpenoid biosynthesis genes

Total exome			Terpenoid		
variant[a]	Query[c]	Reference[d]	biosynthesis variant[b]	Query[c]	Reference[d]
GA	3969 (16.15%)	3985 (16.22%)	AC	12 (6.03%)	8 (4.02%)
GT	1193 (4.86%)	1129 (4.59%)	AT	8 (4.02%)	7 (3.52%)
GC	997 (4.06%)	1024 (4.17%)	AG	36 (19.12%)	30 (15.96%)
AC	1155 (4.70%)	1223 (4.98%)	GA	30 (15.96%)	36 (19.12%)
AG	3985 (16.22%)	3969 (16.15%)	GT	9 (4.79%)	4 (2.13%)
AT	952 (3.87%)	993 (4.04%)	GC	13 (6.91%)	5 (2.66%)
TA	993 (4.04%)	952 (3.87%)	TG	4 (2.13%)	9 (4.79%)
TG	1129 (4.59%)	1193 (4.86%)	TA	7 (3.52%)	8 (4.02%)
TC	3932 (16.00%)	4018 (16.35%)	TC	24 (12.77%)	32 (17.02%)
CG	1024 (4.17%)	997 (4.06%)	CT	32 (17.02%)	24 (12.77%)
CA	1223 (4.98%)	1155 (4.70%)	CG	5 (2.66%)	13 (6.91%)
CT	4018 (16.35%)	3932 (16.00%)	CA	8 (4.02%)	12 (6.03%)
Total	24,570	24,570	Total	188	188

[a]Substitutions are identified in the total collection of exonic sequences identified via structural gene modeling in the *S. officinalis* genome assembly scaffolds

[b]Substitutions are identified in the exonic sequences of the predicted set of terpenoid biosynthesis genes

[c]Query represents the sequence data produced in the present study

[d]Reference represents the *S. officinalis* genome assembly sequence accessed at https://ngdc.cncb.ac.cn/gwh under the ID GWHBJVP00000000

out individually for randomly selected terpenoid biosynthesis genes that harbor single-nucleotide polymorphisms. Among the analyzed variable nucleotide loci in these genes, more than half (9 out of 16) were nonsynonymous and caused amino acid alterations. The N-terminal hydroxymethylglutaryl-CoA synthase domain of a mevalonate pathway gene identified on chromosome 1 (assigned gene ID: 01_RagTag.g10549) is located between the 13th and 163rd amino acids of the encoded peptide. A single-nucleotide polymorphism identified in the hydroxymethylglutaryl-CoA synthase (*hmgs*) gene of *Eucalyptus globulus* was previously shown to be strongly associated with sesquiterpene production traits, including the overall quantity and composition of sesquiterpenes in *Eucalyptus* leaves (Külheim et al. 2011). In the present work, variant analysis identified two single-nucleotide polymorphisms in the exons of the *hmgs* gene, and the SNP (G/A) identified at the 91st position of the coding sequence alters amino acid residue 31 (asparagine) of the peptide, changing the residue to aspartate. Similarly, the sequence encoding the C-terminal terpene cyclase like 1 domain of a monoterpene synthase (myrcene/ocimene synthase) identified on chromosome 2 (assigned gene ID: 02_RagTag.g24179) harbors four SNPs, one of which causes a threonine/arginine change at residue 455. There are four SNPs located in the coding sequence of a geranylgeranyl diphosphate reductase identified as a putative terpenoid backbone biosynthesis gene (assigned gene ID: 03_RagTag.g36721). Of the four SNPs, two are nonsynonymous, altering residues 219 and 283 in the reductase domain. Five amino acid-changing SNPs out of the six variable loci identified in the MEP pathway gene *hds1* (assigned gene ID: 04_RagTag.g42863) alter residues at positions 184, 276, 301, 507 and 525 of the 4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase domain. Interestingly, multiple nonsynonymous SNP loci in the *hds* gene have been found to be significantly associated with monoterpene production in *Eucalyptus* leaves. As a result of that study, the authors indicated that further work is needed to determine which SNP in the *hds* locus is the primary determinant of the altered accumulation of monoterpenes in *Eucalyptus* (Külheim et al. 2011).

In the present work, exons were specifically targeted for variant calling to identify SNPs with high potential for tuning the overall activity of the encoded

enzymes. When considering the effects of variable loci in protein-encoding sequences, it is important to note that the outcome of a substitution in an open reading frame is not restricted by the synonymous/nonsynonymous distinction, and the term "silent" does not actually correspond to "with no consequence" when the complexity of eukaryotic cotranscriptional and posttranscriptional gene regulation is considered (Cartegni et al. 2002; Nielsen et al. 2007; Komar 2007; Tuller et al. 2010). The effect of a mutation in a coding sequence is not restricted to an altered protein structure but can be exerted through differential gene regulation. When a substitution corresponds to a critical residue of the encoded peptide, even a synonymous substitution in an exon can change the transcription factor binding efficiency, therefore altering the rate of transcription. In addition, alternative cleavage and polyadenylation of the encoded mRNA may occur, affecting the length and/or splicing of the final 3'-UTR sequence. Splice variants can be produced due to synonymous mutations during mRNA maturation (Cartegni et al. 2002). Despite no change in the encoded peptide sequence, synonymous codons may alter translation efficiency and therefore may alter the level of protein expression, as tRNAs that recognize synonymous codons in a cell are usually not produced in equal quantities. Thus, codon bias becomes an important determinant of the outcome of a substitution mutation (Komar 2007; Tuller et al. 2010). Differential tRNA trafficking to an mRNA during translation may also cause local variations in protein folding, changing the yield of correctly folded proteins (Komar 2007). Since all of the abovementioned effects are related to gene expression control rather than protein function, they can be exerted regardless of the location of the substitution within an open reading frame. These effects may also vary among tissues, developmental stages or conditions with which the organism must cope, creating ways to fine-tune adaptive responses. As a result, either residing in a domain unit or not, synonymous/nonsynonymous polymorphisms in genes controlling metabolite production may be associated with differentially abundant metabolite composition and/or cause conditional/tissue-specific phenotypes. Diversification in metabolite yield, composition and stress responses would not create discrete phenotypic classes, as observed for qualitative characters, but would explain the genotype dependency of a metabolic feature, as observed and reported for terpenoid production.

Conclusion

SNPs seem to have replaced other marker systems in plant functional genomics and breeding research due to several different advantages, such as high genome coverage, abundance in coding sequences, and superior genotyping accuracy and reproducibility. Another highly beneficial feature of SNPs for plant genomics and molecular breeding is the amenability of genotyping by state-of-the-art high-throughput sequencing approaches, allowing automated allele calling in large populations by targeted, short-amplicon sequencing (Eriksson et al. 2020). In the course of the present work, it was feasible to identify single-nucleotide variants in the coding sequences of common sage terpenoid biosynthesis genes via genome resequencing. The genotype dependency of the main essential oil components of *S. officinalis* has been investigated previously (Jug-Dujakovic´ et al. 2012), indicating that despite the complex nature of terpenoid production, it is feasible to apply genomic selection rather efficiently for the targeted improvement of cultivars with desired compositional traits. Thus, novel cultivars can be tailored to the needs of specific industries that use sage terpenoids as raw materials. Given that extremely diverse terpenoid compounds have a wide range of utilization as raw materials for different industries, from the production of anticancer drugs to the generation of biofuels, allelic variations in biosynthetic genes constitute a valuable potential molecular toolkit. In addition, associating compositional modulations with allelic variations would provide insights into the molecular mechanisms that give rise to the extreme diversity of this metabolite class, which constitutes the highest portion of plant natural products.

Author contributions A.O.U. and A.T.U. designed the study. F.C. prepared the plant and DNA material. F.C., A.S.P., A.O.U. and A.T.U. performed the data analysis. A.O.U. and F.C. wrote the main manuscript text. A.T.U. supervised the entire study.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK). The authors declare that no funds, grants, or other support was received during the preparation of this manuscript.

Data availability The Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under BioProject PRJNA995944. Sequence Read Archive (SRA) files have been deposited under the BioProject with the accession number SRR25318905. Genome assembly scaffolds in FASTA format can be accessed at https://doi.org/https://doi.org/10.6084/m9.figshare.24099927. The gene feature file describing the de novo predicted genes in gff format can be accessed at https://doi.org/https://doi.org/10.6084/m9.figshare.24100032. Coding sequences and encoded peptides of the identified protein-encoding genes are deposited at https://doi.org/https://doi.org/10.6084/m9.figshare.24100083. Coding sequences and encoded peptides of the terpenoid biosynthesis gene content are deposited at https://doi.org/https://doi.org/10.6084/m9.figshare.24100134. The data for variant calling in the total exome are accessible at https://doi.org/https://doi.org/10.6084/m9.figshare.24100155. Assembly quality assessment and additional functional characterization and variant identification data generated during this study are included in this published article as Supplementary Information files.

Declarations

Competing interests The authors declare that they have no relevant financial or nonfinancial competing interests to disclose.

References

Ali M, Li P, She G, Chen D, Wan X, Zhao J (2017) Transcriptome and metabolite analyses reveal the complex metabolic genes involved in volatile terpenoid biosynthesis in garden sage (*Salvia officinalis*). Sci Rep 7:16074. https://doi.org/10.1038/s41598-017-15478-3

Ali M, Hussain RM, Rehman NU, She G, Li P, Wan X et al (2018) DNA Res 25(6):597–617. https://doi.org/10.1093/dnares/dsy028

Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S et al (2022) Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biol 23:258. https://doi.org/10.1186/s13059-022-02823-7

Bryson AE, Lanier ER, Lau KH, Hamilton JP, Vaillancourt B, Mathieu D et al (2023) Uncovering a miltiradiene biosynthetic gene cluster in the Lamiaceae reveals a dynamic evolutionary trajectory. Nat Commun 14(1):343. https://doi.org/10.1038/s41467-023-35845-1

Cartegni L, Chew S, Krainer A (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3:285–298. https://doi.org/10.1038/nrg775

Ceylan F (2023) Adaçayı (*Salvia officinalis* L.) genomunun de novo sekanslanması ve fonksiyonel anotasyonu. (Ph.D. thesis)

Chalchat JC, Michet A, Pasquier B (1998) Study of clones of *Salvia officinalis* L. yields and chemical composition of essential oil. Flavour Fragr J 13:68–70

Chikci R, Rizk G (2013) Space-efficient and exact de Bruijn graph representation based on a Bloom filter. Algorithms Mol Biol 8:22. https://doi.org/10.1186/1748-7188-8-22

Chikhi R, Medvedev P (2014) Informed and automated k-mer size selection for genome assembly. Bioinformatics 30(1):31–37. https://doi.org/10.1093/bioinformatics/btt310

Doyle JJ, Doyle JL (1990) Isolation of plant DNA from fresh tissue. Focus 12:13–15

Dudareva N, Pichersky E, Gershenzon J (2004) Biochemistry of plant volatiles. Plant Physiol 135:1893–1902. https://doi.org/10.1104/pp.104.049981

Eriksson CE, Ruprecht J, Levi T (2020) More affordable and effective noninvasive single nucleotide polymorphism genotyping using high-throughput amplicon sequencing. Mol Ecol Resour 20:1505–1516. https://doi.org/10.1111/1755-0998.13208

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

Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, Von Mering C et al (2017) Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol Biol Evol 34:2115–2122. https://doi.org/10.1093/2Fmolbev/2Fmsx148

Jug-Dujaković M, Ristić M, Pljevljakušić D, Dajić-Stevanović Z, Liber Z, Hančević K et al (2012) High diversity of indigenous populations of dalmatian sage (*Salvia officinalis* L.) in essential-oil composition. Chem Biodivers 9:2309–2323. https://doi.org/10.1002/cbdv.201200131

Kelley DR, Salzberg SL (2010) Detection and correction of false segmental duplications caused by genome mis-assembly. Genome Biol 11:R28. https://doi.org/10.1186/gb-2010-11-3-r28

Komar AA (2007) SNPs, silent but not invisible. Science 315(5811):466–467. https://doi.org/10.1126/science.1138239

Kriebel R, Drew BT, Drummond CP, Gonzáles-Gallegos JG, Celep F, Mahdjoub MM et al (2019) Tracking temporal

shifts in area, biomes, and pollinators in the radiation of *Salvia* (sages) across continents: leveraging anchored hybrid enrichment and targeted sequence data. Am J Bot 106(4):573–597. https://doi.org/10.1002/ajb2.1268

Külheim C, Yeoh SH, Wallis IR, Laffan S, Moran GF, Foley WJ (2011) The molecular basis of quantitative variation in foliar secondary metabolites in *Eucalyptus globulus*. New Phytol 191:1041–1053. https://doi.org/10.1111/j.1469-8137.2011.03769.x

Li J, Wang Y, Dong Y, Zhang W, Wang D, Bai H (2021) The chromosome-based lavender genome provides new insights into Lamiaceae evolution and terpenoid biosynthesis. Hortic Res 8:53. https://doi.org/10.1038/s41438-021-00490-6

Li CY, Yang L, Liu Y, Xu ZG, Gao J, Huang YB et al (2022) The sage genome provides insight into the evolutionary dynamics of diterpene biosynthesis gene cluster in plants. Cell Rep 40:111236. https://doi.org/10.1016/j.celrep.2022.111236

Lu J, Hou J, Ouyang Y, Luo H, Zhao J, Mao C et al (2020) A direct PCR–based SNP marker–assisted selection system (D-MAS) for different crops. Mol Breed 40:9. https://doi.org/10.1007/s11032-019-1091-3

Maksimović M, Vidic D, Miloš M, Šolić ME, Abadžić S, Siljak-Yakovlev S (2007) Effect of the environmental conditions on essential oil profile in two Dinaric *Salvia* species: *S. brachyodon* Vandas and *S. officinalis* L. Biochem Syst Ecol 35(8):473–478. https://doi.org/10.1016/j.bse.2007.02.005

Mantello CC, Cardoso-Silva CB, da Silva CC, de Souza LM, Scaloppi Junior EJ, de Souza GP et al (2014) De novo assembly and transcriptome analysis of the rubber tree (*Hevea brasiliensis*) and SNP markers development for rubber biosynthesis pathways. PLoS ONE 9(7):e102665. https://doi.org/10.1371/journal.pone.0102665

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A (2018) MUMmer4: a fast and versatile genome alignment system. PLOS Comput Biol 14(1):e1005944. https://doi.org/10.1371/journal.pcbi.1005944

Nielsen KB, Sørensen S, Cartegni L, Corydon TJ, Doktor TK, Schroeder LD, Reinert LS et al (2007) Seemingly neutral polymorphic variants may confer immunity to splicing-inactivating mutations: a synonymous SNP in exon 5 of MCAD protects from deleterious mutations in a flanking exonic splicing enhancer. Am J Hum Genet 80(3):416–432. https://doi.org/10.1086/511992

Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA et al (2023) InterPro in 2022. Nucleic Acids Res 51(D1):D418–D427. https://doi.org/10.1093/nar/gkac993

Pichersky E, Raguso RA (2018) Why do plants produce so many terpenoid compounds? New Phytol 220:692–702. https://doi.org/10.1111/nph.14178

Sarrou E, Ganopoulos I, Xanthopoulos A, Masuero D, Martens S, Madesis P et al (2017) Genetic diversity and metabolic profile of *Salvia officinalis* populations: implications for advanced breeding strategies. Planta 246:201–215. https://doi.org/10.1007/s00425-017-2666-z

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Smit A, Hubley R, Green P (2015) RepeatMasker Open-4.0

Song B, Ning W, Wei D, Jiang M, Zhu K, Wang X et al (2023) Plant genome resequencing and population genomics: current status and future prospects. Mol Plant 16:1252–1268. https://doi.org/10.1016/j.molp.2023.07.009

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

Trindade H, Pedro LG, Figueiredo AC, Barroso JG (2018) Chemotypes and terpene synthase genes in Thymus genus: state of the art. Ind Crops Prod 124:530–547. https://doi.org/10.1016/j.indcrop.2018.08.021

Tuller T, Waldman YY, Kupiec M, Ruppin E (2010) Translation efficiency is determined by both codon bias and folding energy. PNAS 107(8):3645–3650. https://doi.org/10.1073/pnas.0909910107

Voorrips RE (2002) MapChart: software for the graphical presentation of linkage maps and QTLs. J Hered 93(1):77–78. https://doi.org/10.1093/jhered/93.1.77

Wimberly J, Cahill J, Atamian HS (2020) *De novo* sequencing and analysis of *Salvia hispanica* tissue-specific transcriptome and identification of genes involved in terpenoid biosynthesis. Plants 9(3):405. https://doi.org/10.3390/plants9030405

Xu H, Song J, Luo H, Zhang Y, Li Q, Zhu Y et al (2016) Analysis of the genome sequence of the medicinal plant *Salvia miltiorrhiza*. Mol Plant 9:949–952. https://doi.org/10.1016/j.molp.2016.03.010