scientific reports

OPEN

Deciphering the biosynthesis pathway of gamma terpinene cuminaldehyde and para cymene in the fruit of *Bunium persicum*

Mohammad Rasoul Samandari-Bahraseman[1]✉, Mjid Hajibarati[1], Babak Khorsand[2], Narges Soltani[3]✉, Keyvan Esmaeilzadeh-Salestani[4,5] & Evelin Loit[5]

Persian cumin (*Bunium persicum*) is a widely used plant in Iranian cuisine, valued for its aromatic and therapeutic compounds, primarily monoterpenoids such as γ-terpinene, cuminaldehyde, and p-cymene. Many ambiguities surround the exact process by which monoterpenoids are synthesized in plants. This study investigates the key genes and biosynthetic pathways involved in monoterpenoid production during fruit development. To achieve this, RNA sequencing and GC-MS analysis were performed on two tissues (inflorescence and stem) to identify differentially expressed genes associated with volatile compound biosynthesis. Our findings revealed four highly expressed genes, terpinene synthetase (TPS1, and TPS2), and two geraniol hydroxylases—correlated with γ-terpinene and cuminaldehyde production. We also propose that cuminaldehyde biosynthesis from p-cymene likely involves two hydroxylation steps mediated by cytochrome P450 enzymes (CYP76B) family genes. Additionally, we identified other candidate genes, including those from the P450 family, alcohol dehydrogenases, and hydroxylases, that may contribute to the pathway of cuminaldehyde biosynthesis. Ultimately, this study presents a comprehensive analysis of *Bunium persicum*'s transcriptome and proposes potential key genes involved in the biosynthesis of cuminaldehyde and γ-terpinene.

Keywords Monoterpenoids, TPS1, TPS2, Geraniol hydroxylases, De Novo assembly

Abbreviations

Inflor	Inflorescence
GPP	Geranyl pyrophosphate
TPS	Terpene synthases
GTS	γ-Terpinene synthetase
B. persicum	*Bunium persicum*

Bunium persicum (Boiss.) B. Fedtsch. is a perennial medicinal plant of the Apiaceae family with 14 diploid somatic chromosomes (2n = 14). Although it shares the same chromosome number with *Cuminum cyminum* L. (cumin), it belongs to a different genus and exhibits distinct genetic and biochemical characteristics. In contrast, *Carum carvi* L. (caraway), another species with a comparable essential oil composition, has 2n = 20 chromosomes, highlighting its greater genetic divergence from *B. persicum*[1]. Various names have been given to this species, including black cumin, black caraway, Persian cumin, kalazira, and wild cumin[2]. It is typically found in temperate to arid climates and mountainous regions. The plant grows to a height of 40–80 cm, has small leaves, and produces umbrella-shaped flowers with white florets in the third year after fruit planting. It generally matures within three months in both vegetative growth and fruit production[3]. Due to limited knowledge regarding its domestication and commercial cultivation, along with high market demand and uncontrolled wild harvesting,

[1]Varjavand Kesht Kariman, Limited Liability Company, Kerman, Iran. [2]Department of Neurology, University of California, Irvine, CA, USA. [3]Production Engineering and Plant Genetics Department, Faculty of Agriculture, Lorestan University, P.O. Box 465, Khorramabad, Iran. [4]Institute of Technology, University of Tartu, Tartu 50411, Estonia. [5]Chair of Crop Science and Plant Biology, Institute of Agricultural and Environmental Sciences, Estonian University of Life Sciences, Kreutzwaldi 1, Tartu EE51014, Estonia. ✉email: m.r.samandari@gmail.com; Soltani.na@fa.lu.ac.ir

this species faces the risk of extinction[4]. Given the frequent confusion between this plant and other species, we provide an illustrative photograph of the entire plant and its fruits in Fig. 1 to aid in accurate identification.

Iranians, Indians, and other cultures have long incorporated the fruits of this plant into their diet due to its aromatic scent and medicinal properties. In addition to dietary use, the fruits are also essential in the food industry[5]. Studies have shown that black cumin fruits exhibit various medicinal properties, including antidiarrheal, anticancer, and antimicrobial effects[6–8]. Furthermore, black cumin fruits are widely reported to contain a diverse range of phytochemicals, particularly monoterpenoids[4,5].

Plants produce a highly diverse array of monoterpenes, with over 1000 structurally distinct compounds identified so far. This diversity is evident in species from the Apiaceae family, such as *Foeniculum vulgare* Mill. (fennel), as well as other aromatic plants like *Mentha spicata* L. (spearmint), *Bunium persicum* (Persian cumin), and *Thymus vulgaris* L. (thyme), which synthesize various volatile compounds contributing to their characteristic aromas. Final stages of monoterpene production pathways are still unclear and poorly defined. It has been suggested that the production of monoterpenes is divided into four stages, the first and second stages producing the geranyl pyrophosphate (GPP, which is a precursor agent for terpenoid production) through both mitochondrial and cytoplasmic pathways[9,10]. The cytoplasmic The mevalonate (MEV) pathway exhibits greater activity in the production of sesquiterpenes, alkanes, and triterpenes, whereas the plastidial methylerythritol phosphate (MEP) pathway is responsible for synthesizing monoterpenes, diterpenes, and tetraterpenes[11]. Stage three monoterpenoids producing involves the production of intermediate agents of monoterpenoids, which are produced by terpene synthase (TPSs) and CYP family proteins[12]. The enzymes required for this stage have been identified to some extent[13–15]. In the final step, as a result of oxidase and dehydrogenase enzymes like CYP P450, the intermediate compounds are converted into a wide variety of monoterpenes[10]. Despite many uncertainties, some studies have identified the precise mechanism by which monoterpenoid compounds such as carvacrol and linalool are formed[16]. However, in our knowledge, no reports have been made regarding the formation of cuminaldehyde, the most abundant compounds of *B. persicum*.

In this study, we aimed to provide a comprehensive understanding of the monoterpenoid biosynthetic pathway in *B. persicum* by integrating transcriptomic and metabolomics data (GC-MS). First, we introduced the fundamental transcriptomic statistics and provided an overall description of *B. persicum* transcriptome enrichment and differential expression patterns between the stem and inflorescence, with a particular focus on the monoterpenoid biosynthetic pathway. In the second part of the study, we specifically focused on identifying the genes involved in cuminaldehyde production. We conducted a literature review and examined the chemical structure and functional groups of these compounds. Based on this analysis, we hypothesize that cuminaldehyde is the final product of the main monoterpenoid pathway in *B. persicum*. According to this prediction, γ-terpinene undergoes aromatization to form p-Cymene, which is then oxidized in two sequential steps to produce cuminaldehyde. This transformation likely involves hydroxylation or oxidation processes to convert an alkane group into an aldehyde. Finally, by integrating transcriptomic data with metabolomics data (GC-MS), we explored the biosynthetic pathways of monoterpenoids in *B. persicum*, with a specific focus on the synthesis of cuminaldehyde and γ-terpinene.

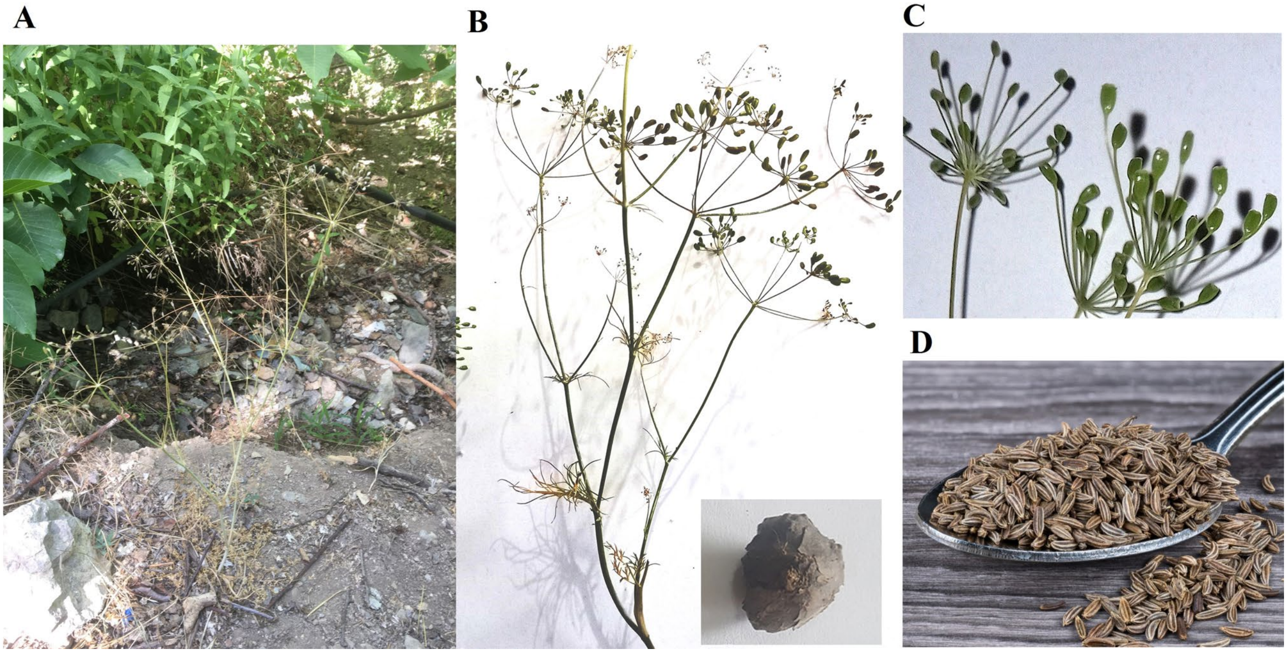


Fig. 1. Morphology of *B. persicum*. (**A**) The plant in its natural habitat. (**B**) Aerial parts and underground tuber. (**C**) Inflorescence of *B. persicum* during the fruit formation stage. (**D**) Final stage of *B. persicum* fruits in the market.

Samples	Number of reads	Number of reads remaining after trimming (%)	GC content (%)	Q20 (%)	Q30 (%)
Inflorescence 1	46,443,507	96.77	44.76	98.59	95.67
Inflorescence 2	44,145,890	97.74	44.12	97.24	94.30
Stem 1	43,562,830	98.55	43.95	98.11	95.98
Stem 2	44,058,732	97.32	44.59	97.87	94.43

Table 1. Statistics and the quality of the samples.

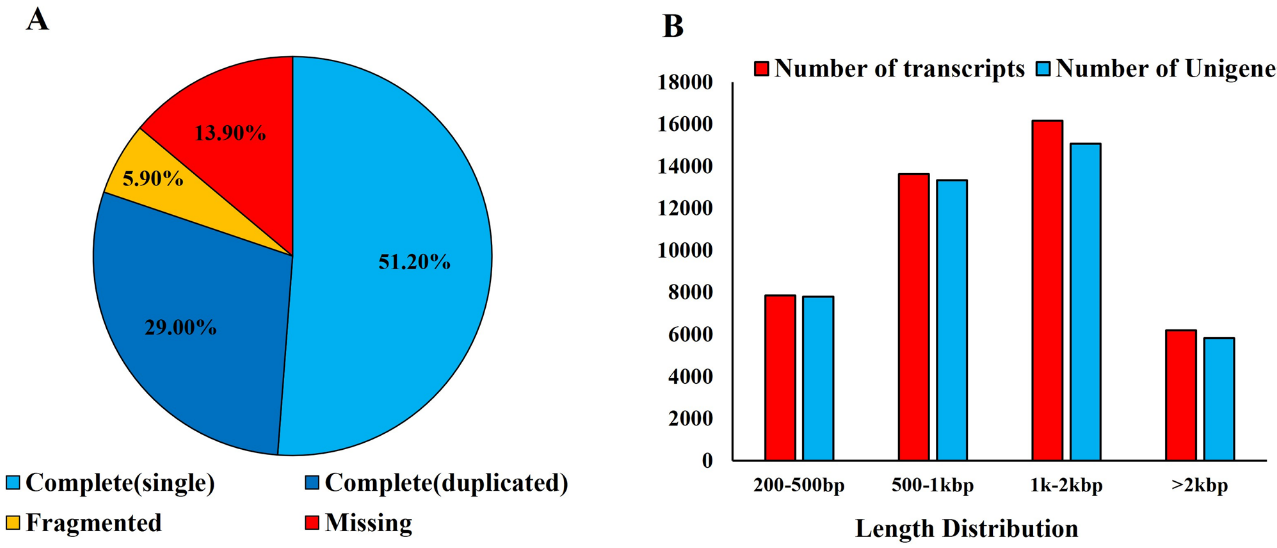


Fig. 2. Overall statistical quality and quality of the samples. (**A**) Results of BUSCO (Benchmarking Universal Single-Copy Orthologs) analysis. (**B**) Length distribution of the assembled transcriptome.

Result

De novo transcriptome assembly and quality evaluation

The sequencing and assembly of raw reads yielded high-quality and reliable data. In total, approximately 178 million raw reads were obtained across all samples. Nearly 94–96% of the reads remained after trimming, with a Q30 quality score ranging from 94.30 to 95.98%. The GC content varied between 43.95% and 44.76%, aligning with expected values for plant transcriptomes. After trimming, approximately 1.45–3.23% of the reads were discarded, primarily due to adapter contamination and low-quality sequences (Table 1). More than 70% of the contigs were longer than 500 bp, while 50% were longer than 1000 bp (Fig. 2B). Transcriptome completeness analysis revealed that 80% of the contigs were of high quality, 6% were fragmented, and the remaining 14% were either missing or of low quality (Fig. 2A). The assembled transcriptome was clustered into 46,156 groups. The N_{50} and mean length of the contigs were 1,450 and 1,191, respectively. Overall, these results suggest that the assembly and resulting unigenes are of good quality and can be relied upon for further analysis.

Functional annotation, gene ontology classification, and enrichment analysis

An analysis of seven databases revealed that 91.7% of unigenes were present in at least one database. A high percentage of unigenes were annotated in the NR (Non-Redundant Protein Database) and NT (Nucleotide Database), with 89.6% and 80%, respectively, while the lowest percentage was associated with KOG at 23.36% (Fig. 3A). Based on an evaluation of annotated unigene terms in NR, 89% showed similarity to *Daucus carota* L., followed by *Vitis vinifera* L. (4.8%) (Fig. s1).

GO analysis indicated that a significant portion of the transcriptome was involved in metabolic and cellular processes within the biological process category. Binding and catalytic activities constituted the majority of the molecular function category. In the cellular component category, the terms “cell,” “cell part,” “membrane,” “organelle,” and “membrane part” had the highest number of unigenes (Fig. 3B).

Differential gene expression between stem and inflorescence tissues

The analysis of gene expression levels between stem and inflorescence samples revealed distinct expression patterns. The mapping of each sample’s reads to the *de novo* assembled reference transcriptome was reliable, with over 85% of each sample successfully mapped (Table 2). Principal Component Analysis (PCA) indicated clear differences in gene expression patterns between the two groups (Fig. 4A). Additionally, overall gene expression levels were higher in the inflorescence compared to the stem (Fig. 4B). Differential expression analysis identified

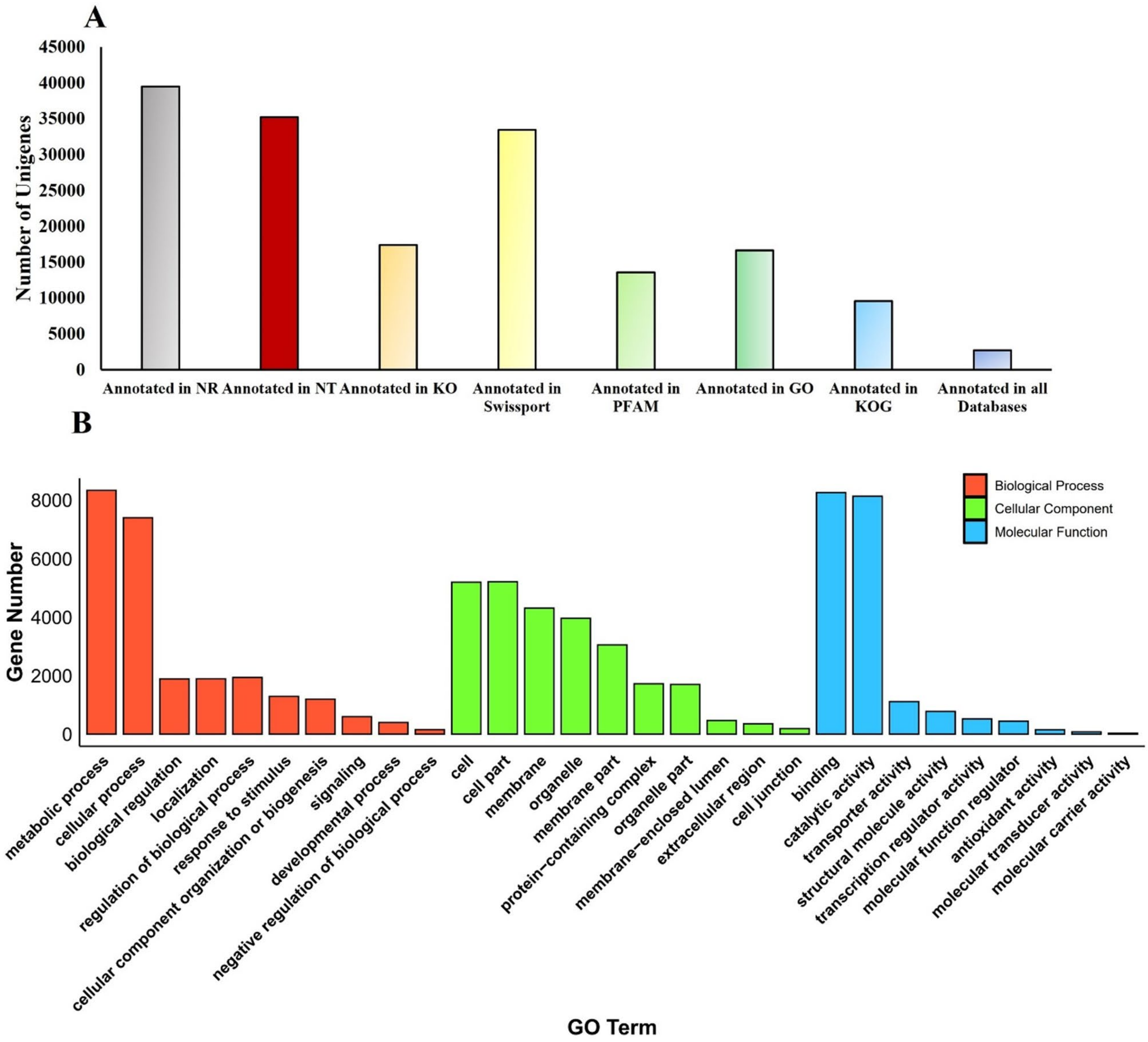


Fig. 3. Overview of the functional annotation and classification of assembled unigenes. (A) Summary of annotation and classification of de novo-assembled unigenes from all *B. persicum* samples in eight databases. (B) GO enrichment analysis of all assembled unigenes from all samples.

Sample name	Total mapped
Inflorescence 1	89.78%
Inflorescence 2	88.68%
Stem 1	87.87%
Stem 2	86.75%

Table 2. Overview of the alignment percentage.

approximately 20,000 genes, with around 10,000 genes significantly more highly expressed in each group (Fig. 4C).

KEGG pathway enrichment of DEGs highlighting monoterpenoid biosynthesis

Through meticulous analysis of both shared and differentially expressed genes, we determined that gene clusters are more abundant in nearly all metabolic processes in the inflorescence compared to the stem (Fig. 5). This disparity is particularly pronounced in metabolic pathways related to secondary metabolites. A detailed

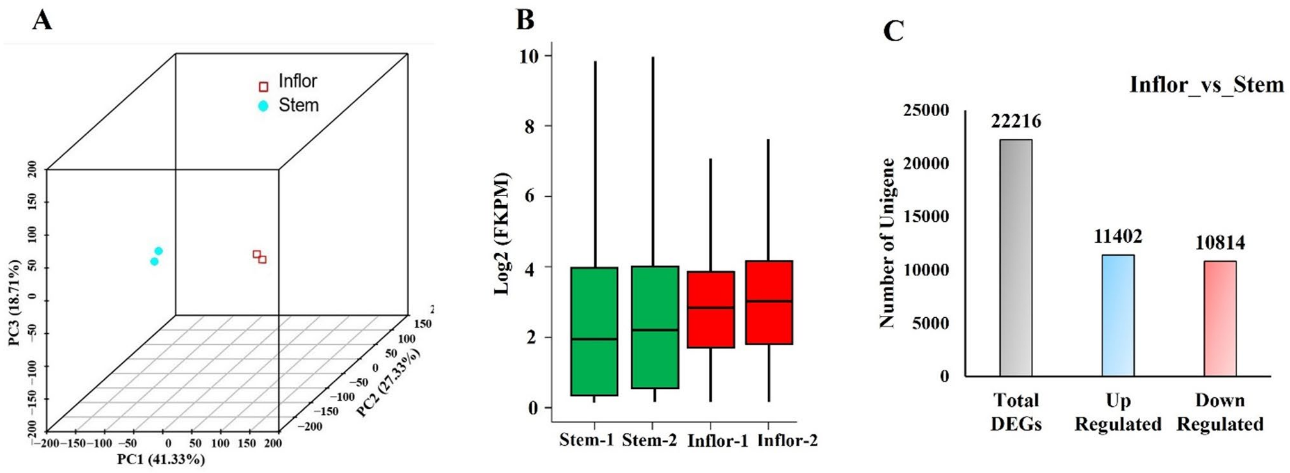


Fig. 4. Differential expression analysis of *B. persicum*. (**A**) PCA of all samples. (**B**) Log2 expression distribution of all samples. (**C**) Statistical representation of differentially expressed genes between inflorescence and stem groups.

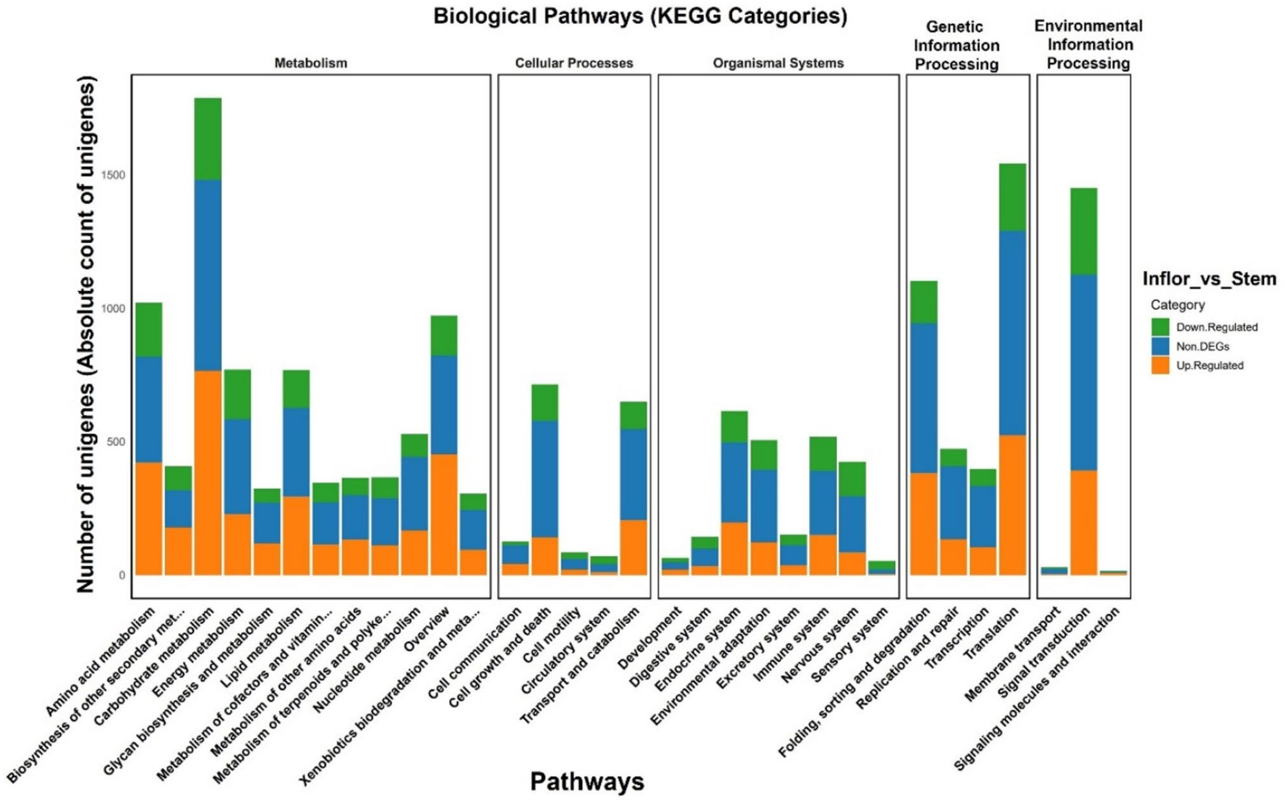


Fig. 5. Enrichment analysis results obtained by BLASTing the assembled unigenes against genome references of *Arabidopsis thaliana* (L.) Heynh., *Daucus carota* L., and *Vitis vinifera* L. using the KAAS tool from the KEGG database.

breakdown of the BLAST processing, focusing on pathways directly or indirectly involved in the production of monoterpenoids, is presented in Table 3. It is observed that, in nearly all processes related to the production of terpenoids, and particularly monoterpenoids, the inflorescence exhibits a greater number of gene clusters. Accordingly, out of approximately 200 blasted clusters (including 163 KO-terms) involved in the production of monoterpenoids and polyketides, 188 clusters (72 KO-terms) were found in both inflorescence and stem tissues. As with other processes, the inflorescence contained more clusters than the stem (122 clusters and 54 KO-terms), while the stem contained 86 clusters (37 KO-terms)[17,18].

KEGG terms related to terpenoids biosynthesis	Non DEGs		Up regulated		Down regulated	
	Number of KO terms	Number of clusters	Number of KO terms	Number of clusters	Number of KO terms	Number of clusters
Metabolism of terpenoids and polyketides	72	188	54	122	37	86
Terpenoid backbone biosynthesis	22	42	24	42	12	15
Monoterpenoid biosynthesis	2	9	3	14	2	10
Sesquiterpenoid and triterpenoid biosynthesis	5	15	5	13	1	5
Diterpenoid biosynthesis	6	11	2	12	3	9
Zeatin biosynthesis	6	32	4	16	2	10
Limonene degradation	1	10	1	5	1	2
Pinene, camphor and geraniol degradation	2	4	2	5	1	1
Carotenoid biosynthesis	17	31	7	7	9	20
Biosynthesis of other secondary metabolites	54	275	75	245	39	143
Biosynthesis of various plant secondary metabolites	14	47	11	39	9	30
Xenobiotics biodegradation and metabolism	38	179	40	139	31	58
Metabolism of xenobiotics by cytochrome P450	3	36	4	23	5	20
Drug metabolism - cytochrome P450	4	39	3	31	5	23
Drug metabolism - other enzymes	10	51	16	44	7	19

Table 3. Selected date of KEGG enrichment analysis related to terpenoids biosynthesis.

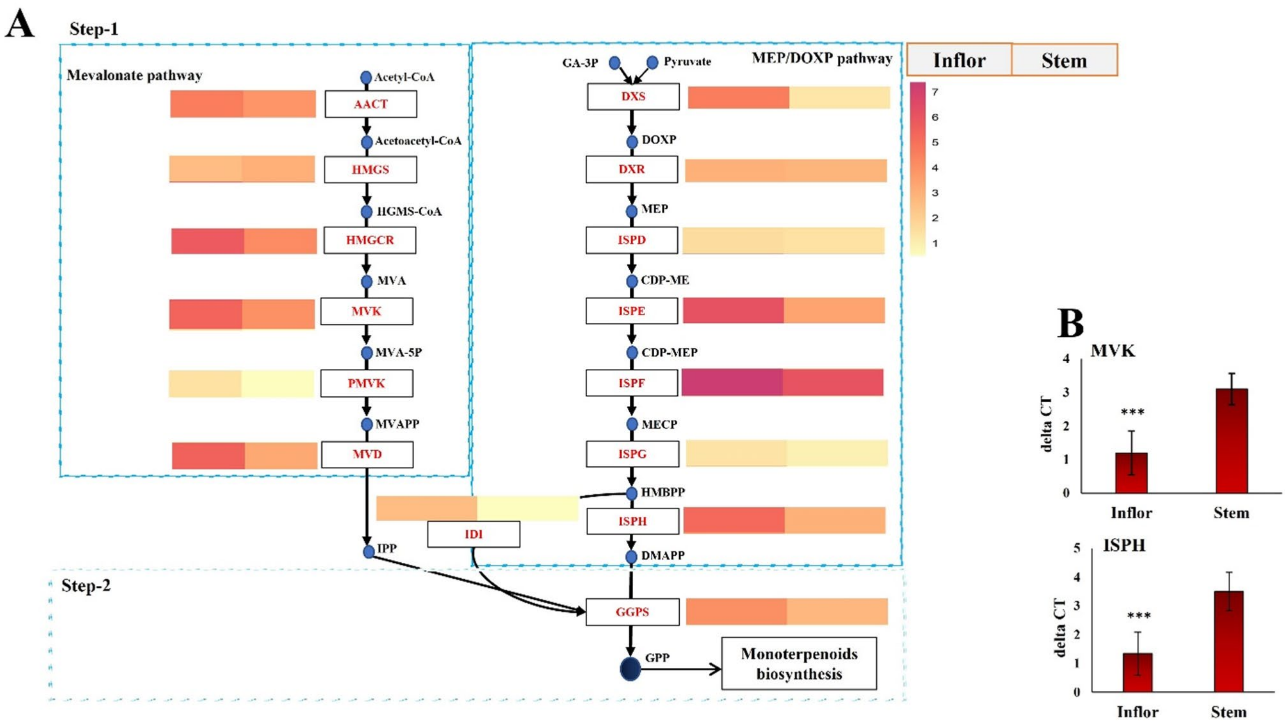


Fig. 6. Terpenoids precursor production pathway expressions of *B. persicum*. (**A**) Illustration cytoplasmic and mitochondrial pathways of terpenoid precursor (GPP) production based on mean of Log2 FPKM (Fragments per kilobase of transcript per million mapped reads) of inflorescences and stems of *B. persicum*. (**B**) The results of real-time PCR for confirmation of the sequencing results. ***, indicates the significant level of 0.001.

Expression analysis and validation of monoterpenoid biosynthesis pathway

Earlier, it was noted that the inflorescence was more active in producing secondary metabolites, especially monoterpenoids, than the stem. The BLAST results revealed that at least one cluster per gene was identified for both the cytoplasmic and mitochondrial pathways of terpenoid backbone biosynthesis (Fig. s2). Upon analyzing the expression levels of unigenes derived from sequencing for the first and second steps of the monoterpenoid pathway, which culminate in GPP production, the average expression levels of the clusters in the inflorescence were significantly higher than in the stem (Fig. 6A). Real-time PCR analysis of genes such as Mevalonate kinase

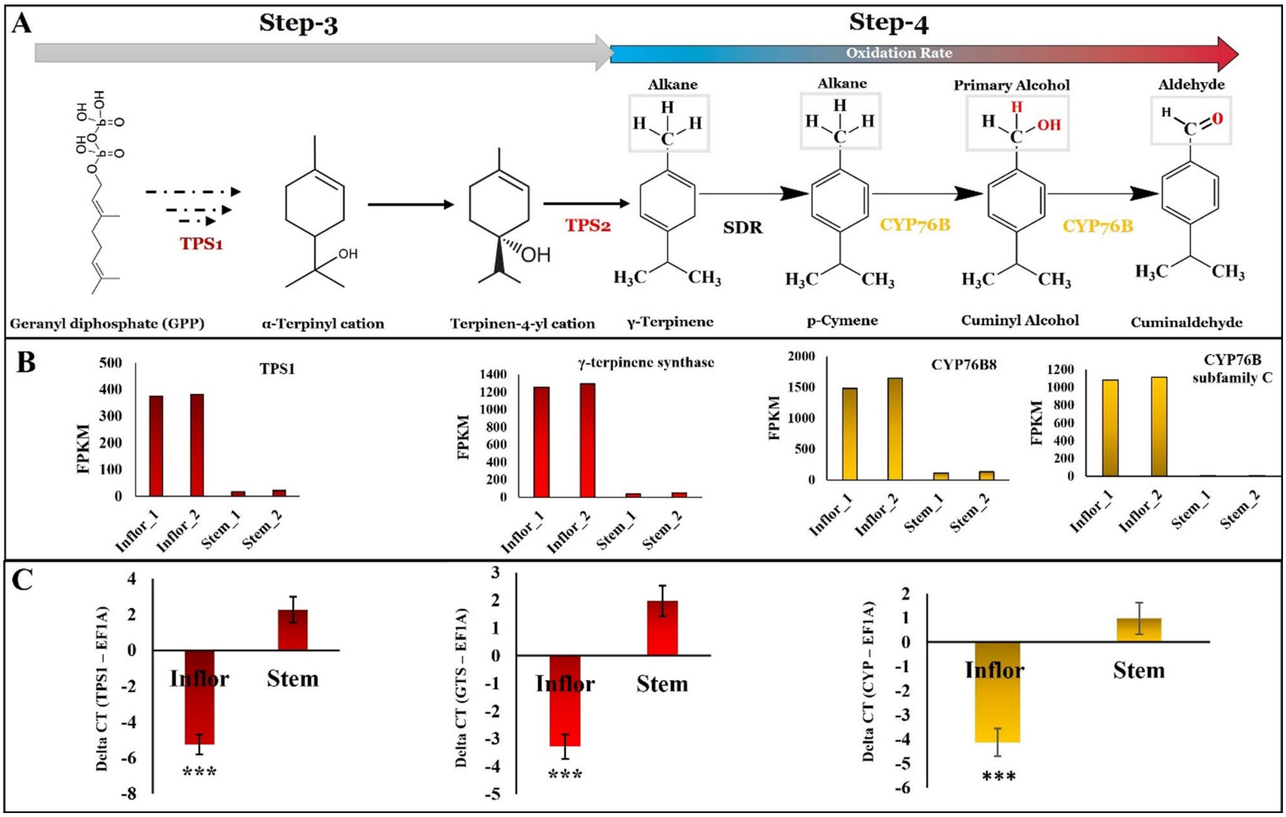


Fig. 7. Overview of the proposed cuminaldehyde biosynthesis pathway and gene expression analysis. (**A**) Proposed pathway of cuminaldehyde production. (**B**) FPKM values of the selected genes. (**C**) Real-time PCR results of the final selected genes. ***, indicates a significance level of 0.001.

Study	γ-terpinene	Cuminaldehyde	γ-Terpinen-7-al	Cuminic-alcohol	*p*-Cymene
[19]	23%	27.8%			13.5%
[20]	46.1%	15%		7.5%	6.7%
[21]	44.2%	16.9%	10.5%		8%
[22]	44.2%	16.9%			8%
[23]	46.1%	23.9%	4.5%		15.9%
[24]	46.1%	15.5%		7.4%	6.7%

Table 4. Summary of GC-MS results of previous research on *B. persicum* fruits.

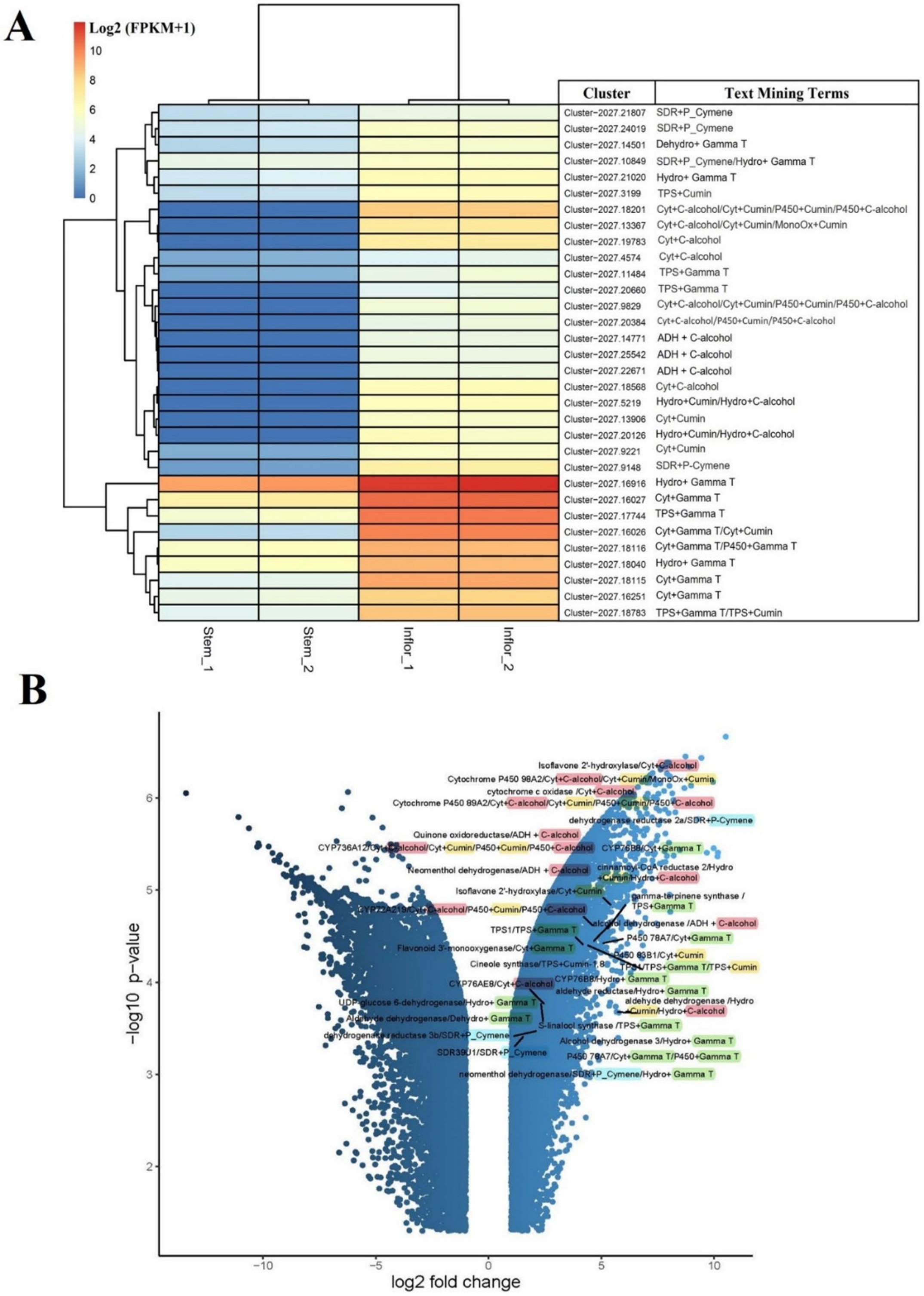
(MVK) and Isopentenyl shikimate kinase (ISPH) further confirmed the sequencing results, showing that the inflorescence exhibited significantly higher gene expression levels compared to the stem (Fig. 6B)[17].

Prediction of cuminaldehyde biosynthesis pathway and essential oil analysis by GC-MS

Reviewing of the previous research indicated that there was a meaningful relation between the *B. persicum* essential compound composition. Based on this relation and considering their chemical structure, we predicted a possible model of cuminaldehyde biosynthesis in *B. persicum* fruit (Fig. 7A; Table 4). In this mean, two steps are proposed to convert p-Cymene to cuminaldehyde, which includes two steps of oxidation on the alkane group to convert this group to an aldehyde group, and as a result, p-Cymene convert to cuminaldehyde.

In the present study, the results of GC-MS showed a significant difference in the composition of compounds in stem and inflorescence essential oils (Table s1a and s1b). Based on the results, it appears that the inflorescence produces significantly elevated levels of cuminaldehyde and γ-terpinene than the stem. In contrast, the stem contains predominantly germacrene D and caryophyllene (Fig. 8). Considering the amount of dry tissue used for essential oil preparation, both stems and inflorescences are active in producing cuminaldehyde, but evidently, inflorescence produce much greater levels than stem.

It should be noted that, based on a review of past studies and the results from the GC-MS analysis in this study (Fig. 9), there is a discrepancy regarding the expected levels of cuminic alcohol, which are lower than anticipated



Cluster	Text Mining Terms
Cluster−2027.21807	SDR+P_Cymene
Cluster−2027.24019	SDR+P_Cymene
Cluster−2027.14501	Dehydro+ Gamma T
Cluster−2027.10849	SDR+P_Cymene/Hydro+ Gamma T
Cluster−2027.21020	Hydro+ Gamma T
Cluster−2027.3199	TPS+Cumin
Cluster−2027.18201	Cyt+C-alcohol/Cyt+Cumin/P450+Cumin/P450+C-alcohol
Cluster−2027.13367	Cyt+C-alcohol/Cyt+Cumin/MonoOx+Cumin
Cluster−2027.19783	Cyt+C-alcohol
Cluster−2027.4574	Cyt+C-alcohol
Cluster−2027.11484	TPS+Gamma T
Cluster−2027.20660	TPS+Gamma T
Cluster−2027.9829	Cyt+C-alcohol/Cyt+Cumin/P450+Cumin/P450+C-alcohol
Cluster−2027.20384	Cyt+C-alcohol/P450+Cumin/P450+C-alcohol
Cluster−2027.14771	ADH + C-alcohol
Cluster−2027.25542	ADH + C-alcohol
Cluster−2027.22671	ADH + C-alcohol
Cluster−2027.18568	Cyt+C-alcohol
Cluster−2027.5219	Hydro+Cumin/Hydro+C-alcohol
Cluster−2027.13906	Cyt+Cumin
Cluster−2027.20126	Hydro+Cumin/Hydro+C-alcohol
Cluster−2027.9221	Cyt+Cumin
Cluster−2027.9148	SDR+P-Cymene
Cluster−2027.16916	Hydro+ Gamma T
Cluster−2027.16027	Cyt+Gamma T
Cluster−2027.17744	TPS+Gamma T
Cluster−2027.16026	Cyt+Gamma T/Cyt+Cumin
Cluster−2027.18116	Cyt+Gamma T/P450+Gamma T
Cluster−2027.18040	Hydro+ Gamma T
Cluster−2027.18115	Cyt+Gamma T
Cluster−2027.16251	Cyt+Gamma T
Cluster−2027.18783	TPS+Gamma T/TPS+Cumin

A
Log2 (FPKM+1)
10
8
6
4
2
0
Stem_1
Stem_2
Inflor_1
Inflor_2
B
Isoflavone 2'-hydroxylase/Cyt+C-alcohol
Cytochrome P450 98A2/Cyt+C-alcohol/Cyt+Cumin/MonoOx+Cumin
cytochrome c oxidase /Cyt+C-alcohol
Cytochrome P450 89A2/Cyt+C-alcohol/Cyt+Cumin/P450+Cumin/P450+C-alcohol
dehydrogenase reductase 2a/SDR+P-Cymene
Quinone oxidoreductase/ADH + C-alcohol
CYP736A12/Cyt+C-alcohol/Cyt+Cumin/P450+Cumin/P450+C-alcohol
CYP76B8/Cyt+Gamma T
Neomenthol dehydrogenase/ADH + C-alcohol
cinnamoyl-CoA reductase 2/Hydro +Cumin/Hydro+C-alcohol
Isoflavone 2'-hydroxylase/Cyt+Cumin
gamma-terpinene synthase / TPS+Gamma T
CYP72A219/Cyt+C-alcohol/P450+Cumin/P450+C-alcohol
alcohol dehydrogenase /ADH + C-alcohol
TPS1/TPS+Gamma T
P450 78A7/Cyt+Gamma T
Flavonoid 3'-monooxygenase/Cyt+Gamma T
P450 83B1/Cyt+Cumin
Cineole synthase/TPS+Cumin-1,8
TPS1/TPS+Gamma T/TPS+Cumin
CYP76B8/Hydro+ Gamma T
CYP76AE8/Cyt+C-alcohol
aldehyde reductase/Hydro+ Gamma T
UDP-glucose 6-dehydrogenase/Hydro+ Gamma T
aldehyde dehydrogenase /Hydro +Cumin/Hydro+C-alcohol
Aldehyde dehydrogenase/Dehydro+ Gamma T
S-linalool synthase /TPS+Gamma T
dehydrogenase reductase 3b/SDR+P_Cymene
Alcohol dehydrogenase 3/Hydro+ Gamma T
SDR39U1/SDR+P_Cymene
P450 78A7/Cyt+Gamma T/P450+Gamma T
neomenthol dehydrogenase/SDR+P_Cymene/Hydro+ Gamma T
−log10 p−value
6
5
4
3
2
−10
−5
0
5
10
log2 fold change

Fig. 8. Final candidate genes identified through keyword-based search and correlation analysis. (A) Clustering and correlation of the selected genes. (B) Positioning of the selected genes on the volcano plot of differentially expressed genes (DEGs). *Cumin* cuminaldehyde, *Dehydro* dehydrogenase, *Hydro* hydrogenase, *C-alcohol* cuminic alcohol, *Cyt* cytochrome, *Gamma T* γ-Terpinene.

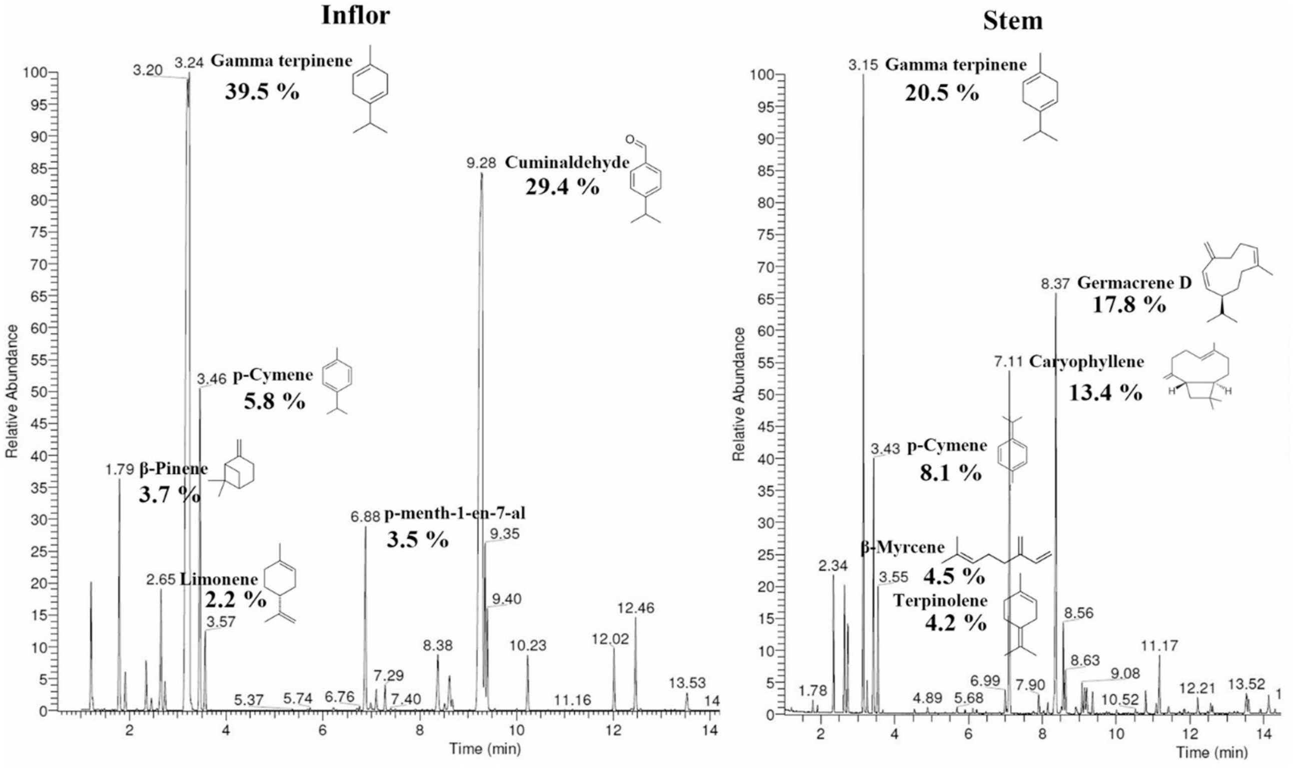


Fig. 9. The volatile compounds identified in GC-MS analysis of inflorescence and stem of *B. persicum*.

(around 10%). However, some studies have reported amounts of this compound that align with the predictions and expectations of this study[19,24]. This discrepancy, however, warrants further detailed investigation.

Keyword-based search in the annotation data file and correlation analysis

Through correlation analysis and keyword-based searches in the annotation data file (Additional File 1), 31 genes were identified as final candidates for the predicted pathway. Correlating compound levels in each tissue with gene expression levels revealed that each cluster had similar term categories (Fig. 8A). The volcano plot (Fig. 8B) showed that all selected genes were positioned in the upper right quadrant, indicating their significant differential expression. Notably, four genes (Cluster-2027.18783, Cluster-2027.16027, Cluster-2027.17744, and Cluster-2027.16026) were identified among the final gene list as the most likely candidates involved in the predicted cuminaldehyde biosynthesis pathway, based on their functional annotation and previous studies (Table 5). Additionally, several other selected genes appear to be promising candidates for facilitating specific reactions within this pathway. For instance, alcohol dehydrogenases (ADHs) and short-chain dehydrogenases/reductases (SDRs) have been previously reported to play a role in monoterpenoid transformation and biosynthesis[16,25].

Final pathway proposal and gene validation in cuminaldehyde biosynthesis

As mentioned above, the final gene list is potentially capable of catalyzing each reaction in the pathway. However, the selection of the best candidate for each reaction was based on a review of the literature, differences in FPKM levels between the inflorescence and stem, and the levels of the compounds of interest (Fig. 7A). In the end, two genes, TPS1 and TPS2 (GTS), were selected as final candidates for catalyzing the conversion of GPP to γ-terpinene. This prediction has also been suggested in previous studies, indicating that this pathway requires two enzymatic reactions. Furthermore, for the downstream pathway following p-Cymene formation, which has not been previously studied, our results identified two genes from the CYP76B family as potential candidates, ultimately leading to the production of cuminaldehyde.

he gene expression results from real-time PCR align well with the sequencing data, showing significantly higher expression in the inflorescence than in the stem (Fig. 7B, C). Additionally, band density analysis confirms both gene expression and the quality of the assembled transcriptome (Fig. s3)[17,26]. Of note, the primers were developed using the shared region of the two CYP76B fragments.

Differential abundance of transcription factors in inflorescence and stem tissues

In the inflorescence tissue of *B. persicum*, a comprehensive analysis of transcription factors (TFs) revealed a total of 2505 TFs. Among these, the MYB, SNF2, C2H2, and C3H families were identified, with respective counts of 18, 12, 11, and 11. Additionally, the AP2, MYB, WRKY, and bZIP families were predominant in stem tissues, with counts of 30, 22, 21, and 20, respectively (Fig. 10).

Genes	NT description	KO description	Swissprot description	keyword-based search terms
Cluster-2027.21807	SDR39U1 homolog	–	SDR39U1 homolog	SDR + p-Cymene
Cluster-2027.24019	–	–	Dehydrogenase reductase 3b	SDR + p-Cymene
Cluster-2027.14501	Aldehyde dehydrogenase family 2 member C4-like	Coniferyl-aldehyde dehydrogenase	Aldehyde dehydrogenase family 2 member C4	Dehydro + Gamma T
Cluster-2027.10849	(+)-Neomenthol dehydrogenase-like	(+)-Neomenthol dehydrogenase	(+)-Neomenthol dehydrogenase	SDR + p-Cymene/Hydro + Gamma T
Cluster-2027.21020	UDP-glucose 6-dehydrogenase 5-like	UDPglucose 6-dehydrogenase	UDP-glucose 6-dehydrogenase 5	Hydro + Gamma T
Cluster-2027.3199	*S*-linalool synthase (LINS)	1,8-cineole synthase	(-)-Alpha-terpineol synthase	TPS + Cumin
Cluster-2027.18201	Cytochrome P450 89A2-like	–	Cytochrome P450 89A2	Cyt + C-alcohol/Cyt + Cumin/P450 + Cumin/P450 + C-alcohol
Cluster-2027.13367	–	Coumaroylquinate(coumaroylshikimate) 3′-monooxygenase	Cytochrome P450 98A2	Cyt + C-alcohol/Cyt + Cumin/MonoOx + Cumin
Cluster-2027.19783	–	–	Isoflavone 2′-hydroxylase	Cyt + C-alcohol
Cluster-2027.4574	Cytochrome P450 (CYP76AE8)	Cytochrome P450 family 76	Cytochrome P450 76A2	Cyt + C-alcohol
Cluster-2027.11484	*S*-linalool synthase (LINS)	1,8-cineole synthase	(-)-Alpha-terpineol synthase	TPS + Gamma T
Cluster-2027.20660	Terpene synthase (TPS1)	1,8-cineole synthase	Monoterpene synthase	TPS + Gamma T
Cluster-2027.9829	Cytochrome P450 CYP736A12	Biphenyl-4-hydroxylase	Cytochrome P450 CYP736A12	Cyt + C-alcohol/Cyt + Cumin/P450 + Cumin/P450 + C-alcohol
Cluster-2027.20384	Cytochrome P450 CYP72A219	–	Cytochrome P450 CYP72A219	Cyt + C-alcohol/P450 + Cumin/P450 + C-alcohol
Cluster-2027.14771	2-Haloacrylate reductase-like	NADPH2: quinone reductase	Quinone oxidoreductase	ADH + C-alcohol
Cluster-2027.25542	Aldo-keto reductase family 4 member C9-like	Alcohol dehydrogenase (NADP+)	NADPH-dependent aldo-keto reductase	ADH + C-alcohol
Cluster-2027.22671	Short-chain dehydrogenase/reductase 2b	(+)-Neomenthol dehydrogenase	Salutaridine reductase	ADH + C-alcohol
Cluster-2027.18568	Cytochrome *c* oxidase subunit	Cytochrome *c* oxidase subunit 6a	Cytochrome *c* oxidase subunit 6a	Cyt + C-alcohol
Cluster-2027.5219	Cinnamoyl-CoA reductase 2	–	Cinnamoyl-CoA reductase 2	Hydro + Cumin/Hydro + C-alcohol
Cluster-2027.13906	–	–	Isoflavone 2′-hydroxylase	Cyt + Cumin
Cluster-2027.20126	Aldehyde dehydrogenase family 2 member B4	Aldehyde dehydrogenase (NAD+)	Aldehyde dehydrogenase family 2 member B4	Hydro + Cumin/Hydro + C-alcohol
Cluster-2027.9221	Cytochrome P450 83B1-like	–	Cytochrome P450 83B1	Cyt + Cumin
Cluster-2027.9148	Dehydrogenase reductase 2a-like	–	Short-chain dehydrogenase reductase 2a	SDR + p-Cymene
Cluster-2027.16916	NADPH-dependent mannose 6-phosphate reductase (m6pr)	Aldehyde reductase	NADP-dependent D-sorbitol-6-phosphate dehydrogenase	Hydro + Gamma T
Cluster-2027.16027	Cytochrome P450 (CYP76B8)	Cytochrome P450 family76	Geraniol 8-hydroxylase	Cyt + Gamma T
Cluster-2027.17744	γ-Terpinene-terpinene synthase (GTS)	(-)-Alpha-terpineol synthase	(-)-Alpha-terpineol synthase	TPS + Gamma T
Cluster-2027.16026	Cytochrome P450 (CYP76B8)	Cytochrome P450 family 76 subfamily C	Geraniol 8-hydroxylase	Cyt + Gamma T/Cyt + Cumin
Cluster-2027.18116	Cytochrome P450 78A7-like	Cytochrome P450 family 78 subfamily A	Cytochrome P450 78A7	Cyt + Gamma T/P450 + Gamma T
Cluster-2027.18040	Alcohol dehydrogenase 1	Alcohol dehydrogenase class-P	Alcohol dehydrogenase 3	Hydro + Gamma T
Cluster-2027.18115	Cytochrome P450 78A7-like	Cytochrome P450 family 78 subfamily A	Cytochrome P450 78A7	Cyt + Gamma T
Cluster-2027.16251	–	–	Flavonoid 3′-monooxygenase	Cyt + Gamma T
Cluster-2027.18783	Terpene synthase (TPS1)	1,8-Cineole synthase	Monoterpene synthase	TPS + Gamma T/TPS + Cumin

Table 5. Annotation summary of the final obtained clusters.

Discussion

Several studies have investigated the compounds in *B. persicum* fruit essential oil, and the majority have identified cuminaldehyde, γ-terpinene, and p-Cymene as the three most abundant constituents (Table 4). However, the biosynthetic pathway of cuminaldehyde has not been proposed until now. In the current study, the essential oil of *B. persicum* stem and inflorescence was analyzed using GC-MS during fruit development, revealing distinct monoterpene compositions between the two tissues. Despite the significant presence of cuminaldehyde in the inflorescence, the stem produced negligible amounts of this metabolite. Instead, it predominantly contained Caryophyllene and Germacrene D. Interestingly, Caryophyllene, which is abundant in the stem, has been reported to have therapeutic benefits[27,28].

As evidenced by qualitative and quantitative analysis of the read lengths and the annotation of contigs in different databases, the raw data and assembled contigs of this study were highly reliable for further transcriptomic investigations of this plant. After de novo transcriptome assembly, approximately 46,000 unigenes were yielded from the transcriptome of this plant. The BUSCO results and quality control analysis conducted prior to assembly demonstrate that the resulting reads are of high quality and reliable for further analysis. In line with previous studies examining the transcriptomes of plants in the Apiaceae family, this study found the transcriptome to be most similar to that of carrot, as carrot is the only plant in the Apiaceae family

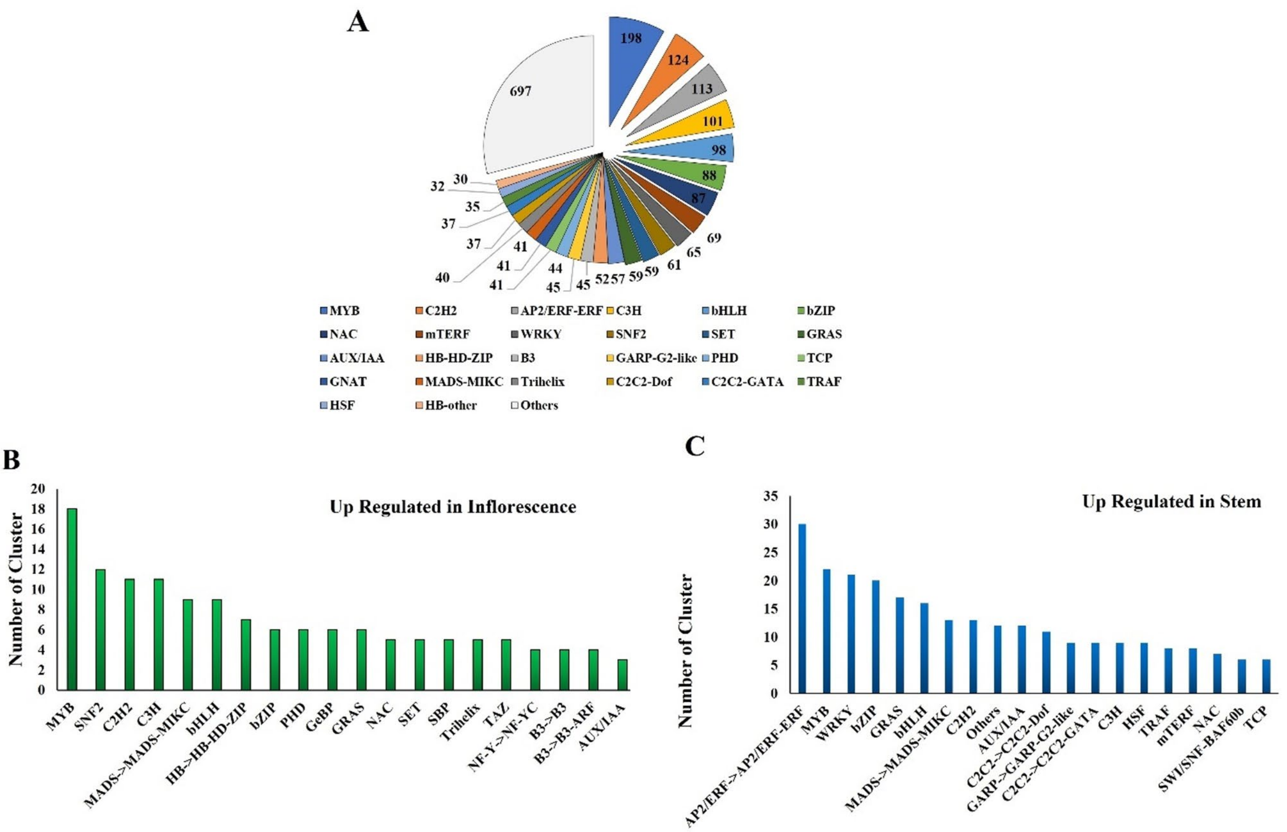


Fig. 10. Transcription factor family analysis. (**A**) Total transcription factors in stem and inflorescence tissues, (**B**) Upregulated transcription factors in inflorescence tissue, and (**C**) Upregulated transcription factors in stem tissue.

with a complete genome sequence[29]. The study of green cumin (*Cuminum cyminum* L.), a species closely related to *B. persicum*, confirmed that cumin has a similar unigene count of 53,103[30].

Our analysis of the cytoplasmic and mitochondrial pathways involved in terpenoid backbone biosynthesis in *B. persicum* revealed that all key genes were expressed in both the stem and inflorescence. Hereupon, study of Zeidabadi et al., has reported that two genes of both mitochondrial and cytoplasmic pathways (HMGR and IPI) were actively expressed during different growth stage of *B. persicum* fruit[31]. However, real-time PCR and sequencing data showed significantly higher expression levels in the inflorescence, suggesting a more active terpenoid biosynthesis in this tissue. This pattern is consistent with findings in other Apiaceae species, where fruits and seeds serve as major sites for terpenoid biosynthesis. For instance, in *Cuminum cyminum* L. (cumin), terpenoid accumulation is significantly higher in seeds compared to vegetative tissues, as these metabolites contribute to seed protection and dispersal[32]. Similarly, in *Foeniculum vulgare* Mill. (fennel), the fruit is the primary site of essential oil production, with key terpenoid biosynthetic genes showing higher expression compared to stems[33]. Our results reinforce the idea that reproductive structures, particularly fruits and seeds, prioritize terpenoid biosynthesis, likely due to their ecological roles in seed dispersal, defense, and germination success.

One of the most abundant compounds found in *B. persicum* and other Apiaceae family is γ-terpinene[21]. As of yet, a suggestion has been made for the production of γ-terpinene from geranyl pyrophosphate (GPP). In this process, the linalyl diphosphate is initially created by re-adding pyrophosphate to the Geranyl cation. Afterward, a linalyl cation is produced through ionization and isomerization of linalyl diphosphate. Next, the α-Terpinyl cation is created through 1,6 ring closure from a linalyl cation[34]. Finally, to produce γ-terpinene via a γ-terpinene synthase activity, the α-Terpinyl cation must first be converted to Terpinen-4-yl cation[35]. The driving genes of this reaction, according to reports and suggestions, are TPS1 and TPS2. While TPS1 proceed the initial part of the pathway, the creation of α-Terpinyl cation[36] γ-terpinene synthase, which is a TPS2, catalyze the conversion of Terpinen-4-yl cation into γ-terpinene[37]. In line with our findings, some researchers have identified the γ-terpinene synthetase (TPS2) responsible for the production of γ-terpinene[37–39]. Our transcriptomic analysis identified two terpene synthase genes, TPS1 and TPS2, that exhibited significant expression in the inflorescence, the tissue with the highest γ-terpinene content. Previous studies have reported similar findings in other plants. The functional roles of these genes were supported by studies on other plants, where γ-terpinene synthase genes have been shown to play crucial roles in the biosynthesis of monoterpenes. For instance, *Nigella sativa* L. (black cumin), a plant known for its medicinal potential, has been found to produce γ-terpinene as a precursor to thymoquinone, the key bioactive compound in its seeds. In this plant, a γ-terpinene synthase

(NsTPS1) was functionally characterized and shown to catalyze the conversion of geranyl diphosphate (GDP) into γ-terpinene[38]. Similarly, in *Thymus vulgaris* L. (thyme), the γ-terpinene synthase gene (Ttps2) has been sequenced and shown to play a key role in the production of thymol, a major essential oil compound in thyme species[37]. The co-expression of TPS1 and TPS2 with γ-terpinene accumulation in *B. persicum* inflorescence in our study supports their functional role in γ-terpinene biosynthesis from GPP, similar to findings in these other species.

Currently, some specific details regarding the biosynthesis of terpenoids, particularly monoterpenoids, remain unknown[40,41]. Since the enzymes responsible for the final stages of monoterpenoid production have not been well characterized[10] still some uncertainty remains for downstream biosynthesis stages of monoterpenoids from GPP[42]. In general, however, cytochromes P450 enzymes and TPSs are the main candidates that could be the most potential candidate to advance the synthesis of these compounds[43]. Studies in the past somehow predicted the formation of p-Cymene from γ-terpinene[16] however, the formation of cuminaldehyde from γ-terpinene in different plants have not been proposed. There has only been limited research reporting the conversion of the C-10 alkane group of p-Cymene to an aldehyde group in the past. However, this pathway was found only in bacteria, which is interestingly similar to our proposed pathway for producing cuminaldehyde from p-Cymene. In this process, which involves two oxidation steps, oxygenation and alcohol dehydrogenation, the alkane group is converted into aldehyde[44,45]. In the current study, four genes were proposed based on their relatively high expression and the high correlation between their differences in expression and differences in γ-terpinene and cuminaldehyde content in two tissues as candidate genes for accomplishing the proposed pathway. For this means, the results indicated that there were two Geraniol hydroxylase, belonging to the CYP76B and CYP76C family highly correlated with cuminaldehyde and γ-terpinene. Previous research showed that these genes could convert geraniol biphosphate into cyclic monoterpenoid intermediates through hydroxylation of an aldehyde group of acyclic compounds like geranyl cation[46,47]. Notably, the proposed reaction in current study needs conversion of an alkane group into an aldehyde of a cyclic compound (p-Cymene). To support this claim, a study has reported that this genes were also able to perform such rection on a cyclic compound a-terpineol and convert it into a 10-hydroxy-a-terpineol[48]. In addition, the carbon number 10 which is the target al.dehyde group of p-Cymene to be converted to cuminaldehyde has been shown to be hydroxylated by CYP76B and CYP76C family[46,48].

Transcription factors (TFs) play a crucial role in regulating gene expression by modulating the expression levels of their target genes. Recent advancements in Next-Generation Sequencing (NGS) techniques have facilitated the identification of TFs and their associated target proteins across diverse species[49,50]. In the context of terpenoid synthesis, the overexpression of specific TFs, such as MYB, bHLH, and NMN, has been observed to affect gene expression patterns, ultimately influencing the production of final terpenoid compounds[51–53]. Notably, our current research highlights a significant upregulation of the MYB family in *B. persicum* fruit, suggesting its potential role as a key enhancer in volatile compound production.

The distinct volatile compound compositions between the stem and inflorescence of *B. persicum* provide a valuable framework for studying the relationship between metabolite production and gene expression. In this study, we proposed candidate genes involved in the biosynthetic pathway of cuminaldehyde based on their expression levels and correlation with γ-terpinene and cuminaldehyde content. While CYP76B and CYP76C family genes emerged as the strongest candidates, it is important to acknowledge that other enzymes, such as neomenthol dehydrogenase and CYP76AE8, which exhibit alcohol dehydrogenase activity, could theoretically contribute to the conversion of cuminic alcohol into cuminaldehyde[25,54]. However, based on prior studies and correlation analyses, CYP76B and CYP76C were identified as the most plausible candidates. Nevertheless, we recognize that our conclusions are based on transcriptomic and metabolomic correlations rather than direct enzymatic assays. Additionally, the limited sample size (four biological replicates) may not fully capture the complexity of monoterpenoid biosynthesis in *B. persicum*. Further experimental validation, including enzyme functional characterization and a broader sampling approach, is necessary to confirm the proposed pathway and refine our understanding of cuminaldehyde biosynthesis.

Conclusions

This study provides the first comprehensive analysis of *B. persicum*'s transcriptome, identifying key genes involved in the biosynthesis of its volatile compounds, including cuminaldehyde. We found that the inflorescence, with higher expression of terpenoid biosynthetic genes, is more active in metabolite production, potentially linked to ecological functions like pollination and defense. Our data suggest that CYP76B and CYP76C genes are likely responsible for the final stages of cuminaldehyde production, while two terpene synthases (TPS1 and TPS2) play a crucial role in the conversion of γ-terpinene from GPP. However, further experimental validation is needed to confirm these findings.

These results offer valuable insights into B. persicum's metabolic pathways and lay the groundwork for future research focused on optimizing its volatile compound production. A more comprehensive study, including larger sample sizes and functional validation, would be crucial not only to confirm these results but also to support the domestication and conservation of this endangered plant, potentially enhancing its productivity and essential oil yield. Additionally, a deeper understanding of the environmental factors that may influence the production of targeted metabolites would further enrich the findings and guide more effective cultivation strategies for this valuable species.

Materials and methods

Plant collection and RNA extraction

In late spring, we collected the aerial parts of Bunium persicum from the foothills of Kerman, while the fruits were in the middle stages of maturation (Fig. 1C). The collected samples were immediately transferred to liquid nitrogen. After brief disinfection, RNA extraction was performed according to the protocol of the Zist-Asia RNA extraction kit. Finally, the quality and quantity of RNA were measured using a 1% agarose gel, and the ratio of OD260/280 and OD260/230 was assessed with a NanoDrop spectrophotometer.

The botanical specimens were authenticated by Dr. Seyed Mansour Mirtadzadini, a botanist at the herbarium within the Shahid Bahonar University of Kerman's biology department. The assigned specimen number is MIR: 4551 (https://biol.uk.ac.ir/). The specimens were obtained in accordance with governmental regulations.

cDNA library construction and sequencing

Four samples of *B. persicum* (two inflorescences and two stems) were sequenced by the Beijing Genomes Institute (China). First, their quality was checked using a Bioanalyzer, and samples with the highest RIN (above 8) were selected for sequencing. cDNA was synthesized using poly-A selection and then sequenced using the standard Illumina kit. Sequencing was performed with a read depth of 6 gigabases (per sample) in paired-end mode with a read length of 150 bp.

RNA sequencing data availability

Raw RNA sequencing data generated in this study have been deposited in the NCBI Sequence Read Archive (SRA) under the BioProject accession number PRJNA1105724. The corresponding SRA accession numbers are SRR28827920, SRR28827918, SRR28827917, and SRR28827919. All data are publicly available and can be accessed at: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1105724.

Data processing and de novo transcriptome assembly

The quality control of the reads was checked before and after trimming by FastQC software (Version 0.11.9)[55] in the Linux operating system environment. The raw reads were trimmed by Trimmomatic[56] software (Version 0.37), and low-quality adapters and reads were removed. Due to the lack of a reference genome for *B. persicum*, the de novo assembled transcriptome for high quality reads was created by Trinity software (v2.13.2)[57] and K-mer 32 was used to assemble the reads and the rest of its parameters were set to default. After that, CORSET[58] software was applied to remove the redundance and them clustering the Trinity results. Hierarchical clustering performs based on multiple mapping events and expression pattern.

Functional annotation of assembled transcriptome

After assembling, to evaluate the quality of the reconstructed transcript, the unigenes were analyzed using BUSCO tool[59]. To perform functional annotation of unigenes, the nucleotide sequences were first converted to amino acid sequences by Trans-Decoder software[60] (an ORF prediction software) with default parameters. Subsequently, for gaining functional annotation of unigenes, all unigenes were blasted against databases such as NR[61] NT[62] SwissProt[63] KOG[64] and KEGG[65] with the BLASTx[66] program (E-value ≤ 1e-5). Data were analyzed utilizingBlast2GO[67] to acquire GO annotation of the unigenes. Metabolic pathways were achieved based on the KEGG (KAAS[68] and BlastKOALA[65]) pathway database. In the study, the PlantTFDB (http://planttfdb.gao-lab.org) was employed to identify and classify transcription factors (TFs) associated with *B. persicum*[69].

Differential gene expression analysis

To evaluate the gene expression, the number of Illumina reads was measured employing Bowtie 2[70] with default parameters, which characterized each unigene expression level of each sample. The quantities of assembled transcripts were calculated by the RSEM software[71]. First, the reads were aligned back to the de novo assembled transcriptome utilizing Bowtie, after that the RSEM was used to calculate the number of aligned reads for each sample. To standardize the expression of genes, the expression values of the unigenes were computed based on fragments per kilobase of transcript per million aligned reads (FPKM > 1) by RSEM software.

Differentially expressed contigs were identified from the count's matrix estimated by RSEM through the edgeR using Rstudio. The TMM normalization were applied to adjust for any differences in sample composition. To obtain the differentially expressed genes (DEGs), a threshold false discovery rate (FDR) of ≤ 0.001 and an absolute value of log2 fold ratio ≥ 2 were used.

Identification of unigenes and gene families related to terpenoid biosynthesis

To elucidate the cuminaldehyde production pathway, we initially reviewed several studies that investigated the quantification of volatile compounds in *B. persicum* fruit. Then, based on the possibility of chemical conversion between compounds, a pathway for producing and converting compounds was suggested. For finding the target genes and due to the fact that there is little information on the function of genes involved in monoterpenoid production, this study attempted to introduce genes associated with the production pathway of black cumin's most important compounds, First, by performing a keyword-based search in the annotation data file (keywords: P450, CYP, Cytochrome, Monooxygenase, Hydrogenases, Hydrolase, Dehydrogenase, TPS, Terpenoid synthase) to identify potential genes related to monoterpenoid biosynthesis. Next, following the selection of the most relevant genes, correlation analysis, and Euclidean relationships were determined between the level of gene expression and the level of metabolite production produced in the stem and inflorescence (using COR function and Pheatmap package in R). Finally, to identify conclusive candidate genes, we considered terminology, the feasibility of advancing the reaction by consulting various databases, the highest correlation, and the closest Euclidean relationship (Fig. 11).

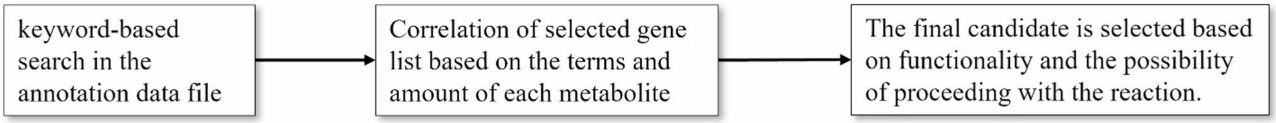

keyword-based search in the annotation data file
Correlation of selected gene list based on the terms and amount of each metabolite
The final candidate is selected based on functionality and the possibility of proceeding with the reaction.

Fig. 11. Processes of selecting most potential clusters for cuminaldehyde production.

Real time PCR

Three biological and three technical replicates were utilized to perform the qRT-PCR assay. The data were analyzed using SPSS software, employing the comparative threshold cycle (ΔCT) method to ascertain the relative gene transcription level. The primer sequences are available in Table s2.

Essential oil extraction and GC-MS analysis

Briefly, a Clevenger apparatus was used to prepare the essential oil. GC-MS analysis was performed with helium gas flowing at a rate of 1 mL/min, and the injector temperature was set at 250 °C. The oven temperature was maintained at 50 °C for 2 min before gradually increasing by 10 °C per minute, reaching 250 °C after 15 min. A Thermo Finnigan Trace DSQ GC-MS system was used for this analysis, with a 70 eV ionization energy, and the column operated using a DB-5 capillary column and helium as a carrier gas. The composition of the compounds was determined by comparing the relative retention times (RT) and mass spectra with those in standards, the NIST, Wiley, and Adam's libraries of the GC-MS system, as well as from literary sources. Notably, to achieve an equivalent quantity of essential oil, varying weights of different tissues were required. Specifically, to yield 1 milliliter of essential oil, approximately 18 g of dry inflorescence tissue were used, while the same amount of essential oil required 100 g of dry stem tissue.

Data availability

Raw RNA sequencing data generated in this study have been deposited in the NCBI Sequence Read Archive (SRA) under the BioProject accession number PRJNA1105724. The corresponding SRA accession numbers are SRR28827920, SRR28827918, SRR28827917, and SRR28827919. All data are publicly available and can be accessed at: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1105724.

Received: 28 May 2024; Accepted: 2 June 2025
Published online: 01 July 2025

References

1. Sheidai, M., Ahmadian, P. & Poorseyedy, S. Cytological studies in Iran zira from three genus: Bunium, carum and cuminum. *Cytologia (Tokyo)*. **61**, 19–25 (1996).
2. Sofi, P. A., Zeerak, N. A. & Singh, P. Kala zeera (Bunium persicum Bioss.): A Kashmirian high value crop. *Turk. J. Biol.* **33**, 249–258 (2009).
3. Bansal, S. et al. A comprehensive review of Bunium persicum: A valuable medicinal spice. *Food Rev. Int.* **39**, 1184–1202 (2023).
4. Singh, S., Kumar, V. & Ramesh. Biology, genetic improvement and agronomy of Bunium persicum (Boiss.) fedtsch.: A comprehensive review. *J. Appl. Res. Med. Aromat. Plants*. **22**, 100304 (2021).
5. Hassanzad Azar, H., Taami, B., Aminzare, M. & Daneshamooz, S. Bunium persicum (Boiss.) B. Fedtsch: An overview on phytochemistry, therapeutic uses and its application in the food industry. *J. Appl. Pharm. Sci.* **8**, 150–158 (2018).
6. Ebada, M. E. Cuminaldehyde: A potential drug candidate. *J. Pharmacol. Clin. Res.* **2**, 1–4 (2017).
7. Rustaie, A. et al. Essential oil composition and antimicrobial activity of the oil and extracts of Bunium persicum (Boiss.) B. Fedtsch.: Wild and cultivated fruits. *Pharm. Sci.* **22**, 296–301 (2016).
8. Jalilzadeh-Amin, G., Nabizadeh, H. & Maham, M. Inhibitory effect of Bunium persicum Boiss essential oil on castor-oil induced diarrhea. *J. Kerman Univ. Med. Sci.* **21**, 139–150 (2014).
9. Kumar, A. et al. Isoprenyl diphosphate synthases of terpenoid biosynthesis in rose-scented geranium (Pelargonium graveolens). *Plant. Physiol. Biochem.* **210**, 108590 (2024).
10. Mahmoud, S. S. & Croteau, R. B. Strategies for transgenic manipulation of monoterpene biosynthesis in plants. *Trends Plant. Sci.* **7**, 366–373 (2002).
11. Davis, E. M. & Croteau, R. Cyclization enzymes in the biosynthesis of monoterpenes, sesquiterpenes, and diterpenes. *Biosynth. Aromat. Polyketid. Isopren. Alkaloids* **209**, 53–95 (2000).
12. Chen, F., Tholl, D., Bohlmann, J. & Pichersky, E. The family of terpene synthases in plants: A mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant. J.* **66**, 212–229 (2011).
13. Liu, L. et al. Integrating RNA-seq with functional expression to analyze the regulation and characterization of genes involved in monoterpenoid biosynthesis in Nepeta tenuifolia Briq. *Plant. Physiol. Biochem.* **167**, 31–41 (2021).
14. Li, M. et al. Genome-wide identification and analysis of terpene synthase (TPS) genes in celery reveals their regulatory roles in terpenoid biosynthesis. *Front. Plant. Sci.* **13**, 1010780 (2022).
15. Weitzel, C. & Simonsen, H. T. Cytochrome P450-enzymes involved in the biosynthesis of mono-and sesquiterpenes. *Phytochem Rev.* **14**, 7–24 (2015).
16. Krause, S. T. et al. The biosynthesis of thymol, carvacrol, and thymohydroquinone in Lamiaceae proceeds via cytochrome P450s and a short-chain dehydrogenase. *Proc. Natl. Acad. Sci.* 118, e2110092118 (2021).
17. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
18. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* **28**, 1947–1951 (2019).
19. Chizzola, R., Saeidnejad, A. H., Azizi, M., Oroojalian, F. & Mardani, H. Bunium persicum: Variability in essential oil and antioxidants activity of fruits from different Iranian wild populations. *Genet. Resour. Crop Evol.* **61**, 1621–1631 (2014).
20. Kareshk, A. T. et al. Efficacy of the Bunium persicum (Boiss) essential oil against acute toxoplasmosis in mice model. *Iran. J. Parasitol.* **10**, 625 (2015).

21. Oroojalian, F., Kasra-Kermanshahi, R., Azizi, M. & Bassami, M. R. Phytochemical composition of the essential oils from three apiaceae species and their antibacterial effects on food-borne pathogens. *Food Chem.* **120**, 765–770 (2010).
22. Jamshidi, A., Khanzadi, S., Azizi, M., Azizzadeh, M. & Hashemi, M. Modeling the growth of Staphylococcus aureus as affected by black zira (Bunium persicum) essential oil, temperature, pH and inoculum levels. *Vet. Res. Forum Int. Q. J.* **5**, 107 (2014).
23. Hajhashemi, V., Sajjadi, S. E. & Zomorodkia, M. Antinociceptive and anti-inflammatory activities of Bunium persicum essential oil, hydroalcoholic and polyphenolic extracts in animal models. *Pharm. Biol.* **49**, 146–151 (2011).
24. Sharififar, F., Yassa, N. & Mozaffarian, V. Bioactivity of major components from the seeds of Bunium persicum (Boiss.) FEDTCH. *Pak J. Pharm. Sci.* **23**, 300–304 (2010).
25. Colonges, K. et al. Two main biosynthesis pathways involved in the synthesis of the floral aroma of the Nacional cocoa variety. *Front. Plant. Sci.* **12**, 681979 (2021).
26. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems database as a model of the real world. *Nucleic Acids Res.* **53**, D672–D677 (2025).
27. Fidyt, K., Fiedorowicz, A., Strządała, L. & Szumny, A. β-caryophyllene and β-caryophyllene oxide—Natural compounds of anticancer and analgesic properties. *Cancer Med.* **5**, 3007–3017 (2016).
28. Scandiffio, R. et al. Protective effects of (E)-β-caryophyllene (BCP) in chronic inflammation. *Nutrients* **12**, 3273 (2020).
29. Soltani Howyzeh, M., Noori, S. A. S., Shariati, V. & Amiripour, M. Comparative transcriptome analysis to identify putative genes involved in thymol biosynthesis pathway in medicinal plant Trachyspermum ammi L. *Sci. Rep.* **8**, 1–19 (2018).
30. Sadeghi, D., Mortazavian, M. M. & Bakhtiyarizadeh, M. R. Transcriptome analysis of Cumin (Cuminum cyminum L.) using RNA-Seq. *Agric. Biotechnol. J.* **9**, 101–116 (2018).
31. Zeidabadi, D. D., Javaran, M. J., Dehghani, H., Monfared, S. R. & Baghizadeh, A. An investigation of the HMGR gene and IPI gene expression in black caraway (Bunium persicum). *3 Biotech* 8, (2018).
32. Bahraminejad, A., Mohammadi-Nejad, G., Abdul Kadir, K., Bin Yusop, M. R. & Samia, M. A. Molecular diversity of Cumin ('Cuminum cyminum' L.) using RAPD markers. *Aust J. Crop Sci.* **6**, 194–199 (2012).
33. Telci, I., Demirtas, I. & Sahin, A. Variation in plant properties and essential oil composition of sweet fennel (Foeniculum vulgare Mill.) fruits during stages of maturity. *Ind. Crops Prod.* **30**, 126–130 (2009).
34. Srividya, N., Davis, E. M., Croteau, R. B. & Lange, B. M. Functional analysis of (4 S)-limonene synthase mutants reveals determinants of catalytic outcome in a model monoterpene synthase. *Proc. Natl. Acad. Sci.* **112**, 3332–3337 (2015).
35. Lima, A. S. et al. Genomic characterization, molecular cloning and expression analysis of two terpene synthases from Thymus caespititius (Lamiaceae). *Planta* **238**, 191–204 (2013).
36. Gao, Y., Honzatko, R. B. & Peters, R. J. Terpenoid synthase structures: A so far incomplete view of complex catalysis. *Nat. Prod. Rep.* **29**, 1153–1175 (2012).
37. Tohidi, B., Rahimmalek, M., Arzani, A. & Trindade, H. Sequencing and variation of terpene synthase gene (TPS2) as the major gene in biosynthesis of thymol in different thymus species. *Phytochemistry* **169**, 112126 (2020).
38. Elyasi, R. et al. Identification and functional characterization of a γ-terpinene synthase in Nigella sativa L (black cumin). *Phytochemistry* **202**, 113290 (2022).
39. Rudolph, K. et al. Expression, crystallization and structure elucidation of γ-terpinene synthase from Thymus vulgaris. *Acta Crystallogr. Sect. F Struct. Biol. Commun.* **72**, 16–23 (2016).
40. Nagegowda, D. A. & Gupta, P. Advances in biosynthesis, regulation, and metabolic engineering of plant specialized terpenoids. *Plant. Sci.* **294**, 110457 (2020).
41. Yang, P. et al. Chromosome-level genome assembly and functional characterization of terpene synthases provide insights into the volatile terpenoid biosynthesis of Wurfbainia villosa. *Plant. J.* **112**, 630–645 (2022).
42. Lu, X. et al. Cyclization mechanism of monoterpenes catalyzed by monoterpene synthases in dipterocarpaceae. *Synth. Syst. Biotechnol.* **9**, 11–18 (2024).
43. Hao, D. L. et al. Function and regulation of ammonium transporters in plants. *Int. J. Mol. Sci.* **21**, 3557 (2020).
44. Eaton, R. W. p-Cymene catabolic pathway in Pseudomonas putida F1: Cloning and characterization of DNA encoding conversion of p-cymene to p-cumate. *J. Bacteriol.* **179**, 3171–3180 (1997).
45. Agullo, L., Romero-Silva, M. J., Domenech, M. & Seeger, M. p-Cymene promotes its catabolism through the p-cymene and the p-cumate pathways, activates a stress response and reduces the biofilm formation in Burkholderia xenovorans LB400. *PLoS One.* **12**, e0169544 (2017).
46. Zhu, X., Zeng, X., Sun, C. & Chen, S. Biosynthetic pathway of terpenoid Indole alkaloids in Catharanthus roseus. *Front. Med.* **8**, 285–293 (2014).
47. Ilc, T., Parage, C., Boachon, B., Navrot, N. & Werck-Reichhart, D. Monoterpenol oxidative metabolism: Role in plant adaptation and potential applications. *Front. Plant. Sci.* **7**, 509 (2016).
48. Höfer, R. et al. Dual function of the cytochrome P450 CYP76 family from Arabidopsis thaliana in the metabolism of monoterpenols and phenylurea herbicides. *Plant. Physiol.* **166**, 1149–1161 (2014).
49. Liu, Z. et al. Identification of novel regulators required for early development of vein pattern in the cotyledons by single-cell RNA-sequencing. *Plant. J.* **110**, 7–22 (2022).
50. Soltani, N., Firouzabadi, F. N., Shafeinia, A., Shirali, M. & Sadr, A. S. De novo transcriptome assembly and differential expression analysis of catharanthus roseus in response to salicylic acid. *Sci. Rep.* **12**, 17803 (2022).
51. Dong, Y. et al. The transcription factor LaMYC4 from lavender regulates volatile terpenoid biosynthesis. *BMC Plant. Biol.* **22**, 289 (2022).
52. Van Moerkercke, A. et al. The bHLH transcription factor BIS1 controls the iridoid branch of the monoterpenoid indole alkaloid pathway in Catharanthus roseus. *Proc. Natl. Acad. Sci.* **112**, 8130–8135 (2015).
53. Zhang, H. et al. The basic helix-loop-helix transcription factor CrMYC2 controls the jasmonate-responsive expression of the ORCA genes that regulate alkaloid biosynthesis in Catharanthus roseus. *Plant. J.* **67**, 61–71 (2011).
54. Ikeda, H. et al. Acyclic monoterpene primary alcohol: NADP + oxidoreductase of Rauwolfia serpentina cells: The key enzyme in biosynthesis of monoterpene alcohols. *J. Biochem.* **109**, 341–347 (1991).
55. Andrews, S. FastQC: A quality control tool for high throughput sequence data. *Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/* (2010).
56. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
57. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
58. Davidson, N. M. & Oshlack, A. Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. *Genome Biol.* **15**, 1–14 (2014).
59. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness. *Methods Mol. Biol.* **1962**, 227–245 (2019).
60. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
61. Benson, D. A. et al. GenBank. *Nucleic Acids Res.* **41**, D36–D42 (2012).
62. Wheeler, D. L. et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **36**, D13–D21 (2007).

63. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement trembl in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
64. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36 (2000).
65. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and ghostkoala: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* **428**, 726–731 (2016).
66. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
67. Conesa, A. et al. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
68. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, W182–W185 (2007).
69. Jin, J., Zhang, H., Kong, L., Gao, G. & Luo, J. PlantTFDB 3.0: A portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acid Res.* **42**, D1182–D1187. (2014). https://doi.org/10.1093/nar/gkt1016
70. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with bowtie 2. *Nat. Methods.* **9**, 357–359 (2012).
71. Li, B. & Dewey, C. N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **12**, 1–16 (2011).

Author contributions

E.L and M.R.S.B: The experiments were designed and planned. M.R.S.B, B.K, M.H and K.E.S: performed the data analysis experiments. M.R.S.B, M.H and N.S: performed dry-lab experiment. M.R.S.B, K.E.S and E.L: wrote the manuscript. Each author made contributions to the manuscript's final version.

Funding

The authors state that they did not receive any financial support for conducting the research, writing the article, or publishing it.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-05415-0.

Correspondence and requests for materials should be addressed to M.R.S.-B. or N.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.