

## Research Article

\*Authors contributed equally.

**Cite this article:** Thulasiram HV, Karegaonkar SJ, Sharma P, Kumar A, Ramkumar S, Pandreka A (2023). Targeted metabolite profiling and *de novo* transcriptome sequencing reveal the key terpene synthase genes in medicinally important plant, *Couroupita guianensis* Aubl. *Plant Genetic Resources: Characterization and Utilization* **21**, 558–570. <https://doi.org/10.1017/S1479262123000953>

Received: 13 September 2022

Revised: 8 November 2023

Accepted: 8 November 2023

First published online: 4 January 2024


### Keywords:

*Couroupita guianensis* Aubl; flower; metabolite profiling; terpene synthases; transcriptomics

### Corresponding author:

Hirekodathakallu V. Thulasiram;  
Email: [hv.thulasiram@ncl.res.in](mailto:hv.thulasiram@ncl.res.in)

# Targeted metabolite profiling and *de novo* transcriptome sequencing reveal the key terpene synthase genes in medicinally important plant, *Couroupita guianensis* Aubl

Hirekodathakallu V. Thulasiram<sup>1,2,3</sup> , Shrikant Jagannathrao Karegaonkar<sup>1,2,\*</sup>, Poojadevi Sharma<sup>1,\*</sup>, Ashish Kumar<sup>1,2</sup>, Sudha Ramkumar<sup>1</sup> and Avinash Pandreka<sup>1</sup>

<sup>1</sup>Chemical Biology Unit, Division of Organic Chemistry, CSIR-National Chemical Laboratory, Pashan, Pune-411008, Maharashtra, India; <sup>2</sup>Academy of Scientific & Innovative Research (AcSIR), CSIR-Human Resource Development Centre Campus, Ghaziabad, Uttar Pradesh-201002, India and <sup>3</sup>CSIR-Institute of Genomics and Integrative Biology, New Delhi-110007, India

## Abstract

The *Lecythidaceae* family tree, *Couroupita guianensis* Aubl, popularly known as Nagpushpa, is a widely cultivated ornamental tree with several uses in traditional medicine. The tree is revered as highly sacred in Indian traditional culture due to its uniquely shaped, fragrant flowers. Considering the significance, we were prompted to carry out the metabolite and transcriptome analysis of Nagapushpa. The flower, petals, stamen, stem and leaf of *C. guianensis* were metabolically profiled, and it was discovered that the flower tissue contained the highest terpenoid reservoir. A number of terpenoid pathway transcripts were also found in the flower tissue after transcriptome profiling. KEGG pathway mapping was carried out to correlate transcript sequences with the biosynthesis of different types of terpenes. We were able to clone three full-length terpene synthase gene candidates, i.e. monoterpene ocimene synthase, diterpene ent-kaurene synthase and sesquiterpene farnesene synthase. The transcript expression of selected terpene synthase genes was also verified in flower tissue. These cloned sequences were used for *in silico* structural investigations and protein function prediction at the level of 3D structure. The data presented in this study provide a comprehensive resource for the metabolic and transcriptomic profiles of *C. guianensis*. The study paves the way towards the elucidation of terpene biosynthetic pathway in *C. guianensis* and heterologous production of useful terpenoids in the future.

## Introduction

*Couroupita guianensis* Aubl is a large deciduous tropical tree that belongs to the *Lecythidaceae* family (Shekhawat and Manokari, 2016). Peculiar features of its flower and fruit make it a distinguished tree, and it is widely planted as an ornamental tree in botanical gardens around the globe. Its flower shape gives a distinct impression that, in Indian traditional culture, is interpreted as a snake hood-like stamen structure guarding a stigma in the shape of a Shiva lingam (an Indian holy symbol) at the flower's centre. This feature has given rise to many Indian common names, such as 'Kailashpati' in Hindi, 'Mallikarjuna' in Telugu and 'Nagalingapushpam or Nagpushpa' in Tamil. The tree is considered sacred in India and Sri Lanka, as the flowers of the tree are offered in holy ceremonies in these countries (Lim, 2012; Shekhawat and Manokari, 2016). The tree is also commonly referred to as the 'Cannonball tree' in English due to its fruit shape and size. The fruits are globular brown woody with size of a human head or a cannon and are used for feeding animals (Lim, 2012; Shekhawat and Manokari, 2016).

This sacred tree has been used in traditional medicine in India as well as worldwide. Flowers, leaves and barks of *C. guianensis* are used to treat hypertension, tumours, pain, inflammatory processes, malaria and many other health issues (Sanz-Biset *et al.*, 2009). Juice made from the leaves is used to cure skin diseases. The fruit has been used for disinfecting the wounds and young leaves are used in curing the toothache (Al-Dhabi *et al.*, 2012). Further, it is shown to possess pharmacologically relevant biological properties, such as anti-bacterial, anti-biofilm, anti-oxidant, ovicidal, larvicidal, anti-ulcer, anti-arthritic, anti-platelet, anti-diarrhoea, analgesic, anti-inflammatory, anti-fertility, anti-cancer, neuropharmacological, anxiolytic, anti-plasmodial, anti-depressant, anti-nociceptive, immunomodulatory, anti-quorum sensing, anti-malarial and wound healing (Sanz-Biset *et al.*, 2009; Al-Dhabi *et al.*, 2012; Shekhawat and Manokari, 2016; Kaneria *et al.*, 2017).



Few reports are available in literature on metabolite profiling of *C. guianensis*, which indicated the presence of eugenol, linalool, nerol, tryptanthrine, indigo, indirubin, isatin, linoleic acid, carotenoids, sterols and (E,E)-farnesol metabolites (Khan *et al.*, 2014; Kaneria *et al.*, 2017).

Despite multiple available reports on metabolite profiling and pharmacological activities, we realized that *C. guianensis* is still a rather uncharacterized plant in terms of transcriptomics and tissue-specific metabolite analysis. With the advent of high-throughput mass spectral analytical techniques and nucleotide sequencing in the past decade, significant efforts have been made by researchers to carry out metabolomics and transcriptomics approaches on unexplored medicinal plants for their detailed characterization (Guo *et al.*, 2021; Alami *et al.*, 2022). Techniques such as RNA-Seq and mass spectrometry act as modern lenses through which we can characterize traditional medicinal plants in detail at the molecular level. These efforts are specifically focused on identifying genes and metabolites of secondary metabolism because it is primarily these molecules that directly or indirectly give rise to unique characteristics such as fragrance, flavour, colour, pharmacological activity, plant defence against abiotic and biotic stresses, disease resistance, etc., in medicinal plants (A. Kumar *et al.*, 2023; S. Kumar *et al.*, 2023). Such knowledge offers the possibility of further biotechnological interventions, such as plant breeding or genetic manipulation for trait improvement, optimization of plant cultivation and more recently heterologous gene expression for the production of desirable secondary metabolites in bacteria and yeast (Rai *et al.*, 2017; Navale *et al.*, 2019; Guo *et al.*, 2021; A. Kumar *et al.*, 2023; S. Kumar *et al.*, 2023).

Considering the lacunae, we ventured into metabolite and transcriptome profiling of *C. guianensis* to identify the range of its secondary metabolites and decipher relevant secondary metabolite pathway genes. Such molecular data may provide us the opportunity to assess whether there may be some scientific reasoning underlying the age old traditional wisdom, that the flower should be used for sacred offerings. These genes elucidated in this study may further be used in numerous biotechnological applications in the future.

For the study, initially, we carried out metabolic profiling of the whole flower, petals, stamen, stem and leaf of *C. guianensis* and also screened these tissues for their antimicrobial activity. Flower tissue stood out among all other plant parts for having a diverse and large terpenoid repertoire and potent antibacterial action. These results led us to concentrate our efforts on a thorough examination of floral tissue and construct a flower transcriptome. A cDNA library generated from the RNA of flower tissue was sequenced, and transcriptomic analysis was carried out. We successfully screened out terpenoid pathway transcripts from flower tissue and correlated them with terpenoid biosynthesis. Then, using three full-length terpene synthase gene sequences, we performed structural investigations to predict gene architecture and 3D protein structure for protein function prediction. This work is the first study of the secondary metabolite biosynthesis pathway of the hitherto underexplored plant *C. guianensis*.

## Materials and methods

### Plant materials

*Couroupita guianensis* tissue, i.e. the whole flower, flower petal, stamen, stem and leaf, was collected in liquid nitrogen from a

tree at the NCL commercial complex near the National Chemical Laboratory in Pune. The plant used in this study was confirmed as *C. guianensis* and authenticated by a botanist at Agharkar Research Institute, Pune (the herbarium accession number allotted to the plant is AHMA: 32430).

### Phytochemical extraction and GC-MS analysis

Metabolite analysis was carried out in five tissues, including the whole flower, petal, stamen, leaf and stem of *C. guianensis*. Tissues were crushed to powder under liquid nitrogen and extracted with TBME (10 ml × 3) by continuous stirring for 3 h for a total of three times. The pooled TBME layer was passed through anhydrous sodium sulphate, concentrated under reduced pressure to obtain crude triterpenoid extract, and reconstituted to 500 µl in TBME. For analysis of TBME extract, GC-MS was performed on an Agilent 7890A GC coupled with a 5975C mass detector, and the conditions used were as follows: Restek Rtx-5 ms (30 m × 0.25 mm × 0.25 µm) capillary column was used; helium was used as carrier gas flow with a flow rate of 1.0 ml/min. The column was initially maintained at 150 °C for 2 min, then the temperature was raised from 150 to 250 °C at a 5 °C/min rate with a hold of 11 min, and finally the temperature was maintained at 270 °C for 15 min. Injector and detector temperatures were 230 and 280 °C, respectively. Then 1 µl of plant extract was injected into the column. Compounds were identified by comparison with the mass spectra reference library NIST MS and by using retention time matches with reference standards wherever possible (Eugenol and Linalool). The data were processed by MSD ChemStation Data Analysis (Agilent Technologies, USA).

### RNA extraction from the flower tissue of *C. guianensis*

Total RNA was isolated from flower tissue (pre-treated with an acetone wash) by a spectrum kit (Total RNA Isolation Kit, Sigma-Aldrich, USA) and treated with DNase to remove DNA contamination. Any residual contamination and integrity of total RNA were checked by electrophoresis on 1% agarose made in 0.1% DEPC containing TAE buffer. Further concentrations and impurities of salt and proteins were analysed on the Nanodrop (Thermo Fisher, USA). Isolated and high-quality RNA from flower tissue was sent for sequencing.

### Transcriptome de novo assembly and functional annotation

Isolated RNA from the flower of *C. guianensis* tissues was sent to Genotypic Technology in Bengaluru, India. NextSeq500 (Illumina, USA) was used for the sequencing of RNA to generate processed reads. These reads were assembled to generate unigenes by Trinity software for the generation of the k-mers (25 base pairs). To assign molecular function, biological processes and cellular components of the transcript, functional annotation of unigenes was performed using KEGG-KAAS analysis, Pfam domain analysis and MEGA blast search against the NCBI database, SwissProt/Uniprot database and Protein Data Bank (PDB) with an *E*-value  $\leq 10^{-5}$ .

### cDNA synthesis and semi-quantitative PCR of selected terpene synthases

The RNA isolated from the floral tissue of *C. guianensis* was used to produce cDNA according to the SuperScript® III First-Strand

Synthesis System (Invitrogen, USA) kit instructions. Semi-quantitative RT-PCR was used to verify the expression of the cloned genes (details can be found in SI Material and Methods) obtained from the study. The total RNA of the flower was extracted, as mentioned earlier. After reverse transcription, semi-quantitative RT-PCR was carried out with GADPH as the internal control reference gene. The semi-quantitative RT-PCR experiment on tissue was performed with three repetitions. The final PCR programme used for amplification of all three transcripts was: 1 cycle of 95 °C (5 min); 30 cycles of 95 °C (30 s); 56 °C (30 s); 72 °C (2.5 min); 72 °C (5 min). The amplified fragment was resolved on a 1% agarose gel, visualized by staining with Gel Red dye (Sigma-Aldrich, USA), and imaged digitally. ImageJ was used for densitometry analysis of amplified PCR products in gel images. The intensity of cloned genes was normalized against that of internal control, and the expression ratio of three cloned genes was represented as arbitrary units in flower tissue along with the standard deviation.

### Physicochemical characterization and phylogenetic analysis

ExPasy's ProtParam server was used for the primary structure analysis of the three sequenced genes. The biophysical and biochemical properties such as isoelectric point (pI), molecular weight, aliphatic index, extinction coefficient and GRAVY were computed using this programme (Gasteiger *et al.*, 2005). The nucleotide sequences of three full-length ORFs of putative terpene synthases were subjected to blastx analysis. Conserved domain searches were performed using the Clustal Omega tool and the Conserved Domain Database (CDD) available at NCBI for the identification of conserved motifs. Further, matching-reviewed terpene synthases were screened out from the UniProt database. All these sequences were subjected to NgPhylogeny.Fr analysis for phylogenetic analysis in a one-click workflow (Lemoine *et al.*, 2019).

### Protein 3D-structure and gene function prediction analysis

Homology protein modelling was carried out using the Swiss modeller in Automated mode, and validation of the protein models was carried out using PDBSum by evaluating the Ramachandran Plot (Schwede *et al.*, 2003; Laskowski *et al.*, 2005). The PROCHECK programme was used to check the stereochemical excellence and the overall structural geometry of the homology model at both 2D and 3D levels (Laskowski *et al.*, 2018). The ProFunc web server tool was used to predict the biochemical function of homology-modelled *C. guianensis* terpene synthase proteins at the 3D structure level (Laskowski *et al.*, 2005). All the protein structures were visualized using UCSF Chimera software (Pettersen *et al.*, 2004).

## Results

### Metabolite profiling

The stem, leaf and flower tissues of *C. guianensis* showed inherent metabolite variety when metabolite profiling was done using GC-MS (Fig. 1 and Fig. S1). The metabolite composition in these plant tissues ranged from volatiles such as phenylpropanoids, benzenoids and terpenoid groups; straight chain hydrocarbons; and non-volatiles such as high molecular weight terpenoids, steroids, straight chain hydrocarbons, etc. Phenylpropanoids/

benzenoids and terpenoid group volatiles were discovered to be highly concentrated in flower tissue, accounting for 83 and 16%, respectively (Fig. 1). Although high in percentage, phenylpropanoids and benzenoids diversity was low, and only nine metabolites of the group were found in flower tissue (Table S1). Of the nine metabolites, eugenol (29.54%), isatin (21.96%) and phenylethyl alcohol (0.79%) were the main metabolites of the group in flower tissue.

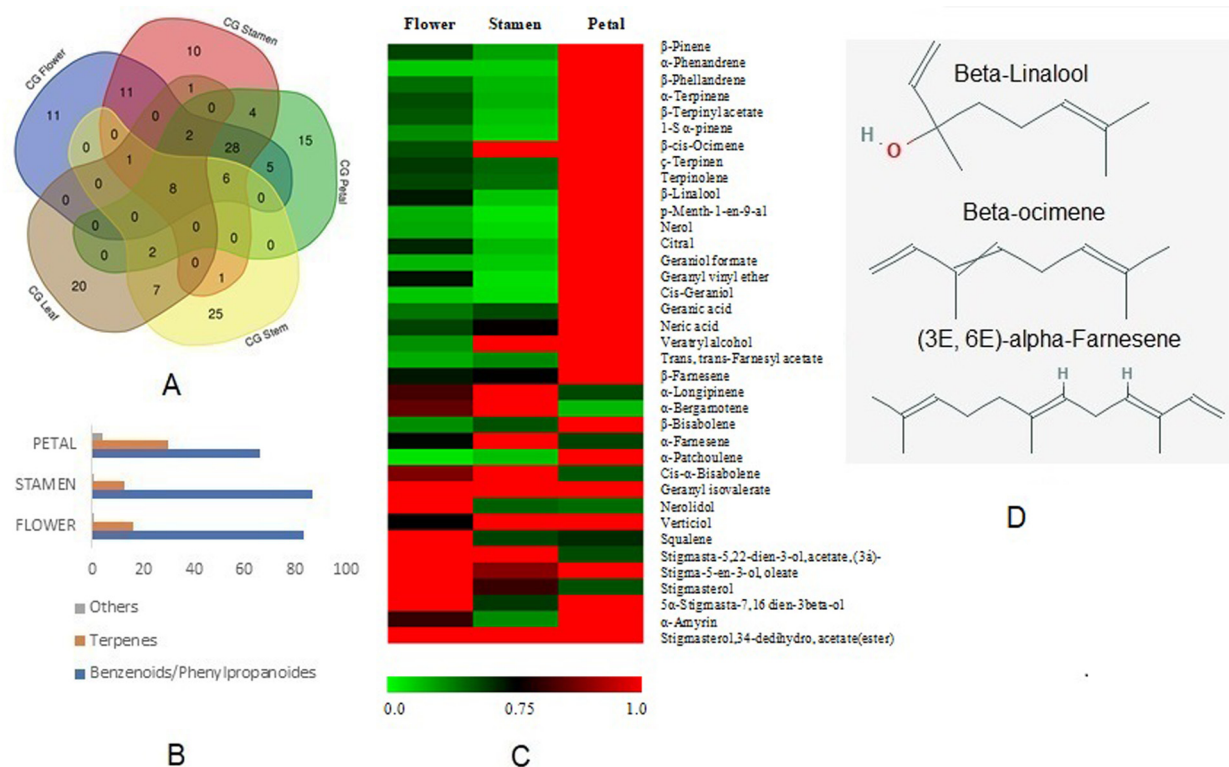
Although low in percentage, terpene diversity was high, and 28 different terpenes were found in flower tissue (Table S1). Among the 28 terpenes, beta-linalool (a monoterpene) (2.22%), geranic acid (a monoterpene) (3.20%) and alpha-farnesene (a sesquiterpene) (0.49%) were major metabolites detected in the flower tissue. Further, monoterpenes and sesquiterpenes were the dominant types of terpenoids in flower tissue, with an overall monoterpene to sesquiterpene ratio of 4:1. A heat map comparing terpenoid content variation in flower, petal and stamen is also shown in Fig. 1. We found that petals had a high terpene content compared to stamen (Fig. 1, Table S1). These terpenoids and phenylpropanoids with benzenoids may contribute to the flower's scent as well as various biological activities.

The flower, petal, stamen, stem and leaf tissue of *C. guianensis* were screened for antimicrobial activity (as described in SI Material and Methods). Among them, flower tissue and its sub-parts, petal and stamen, showed the most potent anti-microbial activity against the bacterial cultures used in this study (Fig. S1). Flower, petal and stamen tissue extracts inhibited bacterial growth for bacterial strains *Klebsiella pneumoniae*, *Salmonella typhi* and *Staphylococcus aureus* at MIC  $\leq$  0.0039 mg/ml and *Pseudomonas aeruginosa* at MIC 0.5 mg/ml, respectively. Evidently, the secondary metabolite profile containing highly diverse terpenes and high antibacterial activity of flower tissue made it stand out among other tissues. We focused on performing transcriptomics on diverse terpenoid-containing flower tissue to better understand the terpene production pathway.

### Transcriptome generation and analysis

The transcriptome of any tissue reflects its biosynthetic machinery, and therefore, could reveal key genes involved in secondary metabolite production as well. For this purpose, good quality of RNA was isolated from flowers and used for transcriptome sequencing (Fig. S2). The *de novo* transcriptomic assembly of sequenced RNA was carried out using Trinity software to generate 32.94 million high-quality reads. The clustering of these reads resulted in 55,995 unique putative transcripts of an average length of 1208 bp and an N50 of 1808 bp (Table 1). Further, around 23,474 proteins were found with an average length ranging from 1000 to 5000 bp, transcript indicating the presence of functional proteins (Table 1).

Using the BLAST2.5.03 version, a homology search was conducted against the Viridiplantae dataset from the Uniprot database, which contains 4,269,328 protein sequences, to annotate transcripts. At least 64.73% of the transcripts were functionally annotated with high confidence (e1–5). *Couroupita* unigenes were functionally classified into different Gene Ontology (GO) terms (Fig. 2). Classification showed that 14.38% of the annotated genes were involved in biological processes, 42.54% in cellular components and 43.08% in molecular function (Fig. 2). Within the biological process, regulation of transcription (20.79%) and transcription (18.6%) were the two dominant GO terms, followed by terms such as metabolic processes, defence responses,



**Figure 1.** *Couroupita guianensis* flower targeted metabolite profiling and analysis: (A) Venn diagram representing differences and similarities among GC-MS metabolite profile of different plant tissues. (B) Major metabolite class composition of flower tissue. (C) Heatmap of terpene content variation of flower, stamen and petal. (D) Major terpenes in flower tissue.

translation, protein folding and transmembrane transport. Defence responses (4.6%) in the GO term suggest that *Couroupita* flowers are probably an active tissue for secondary metabolism. The majority of 53.54, 17.51 and 7.9% of the annotated genes fell into the GO terms of integral components of the membrane, nucleus and cytoplasm, respectively, under the category of cellular components. In the group of molecular functions, ATP binding, zinc binding, nucleic acid DNA binding and metal ion binding were the principal GO terms of molecular function, comprising 26.69, 15.2, 11.51 and 10.5% of annotated genes, respectively.

Pathway analysis was done using the KAAS4 Server. Different plants, namely, *Arabidopsis thaliana* (thale cress), *Arabidopsis lyrata* (lyrate rockcress), *Brassica napus* (rapeseed), *Brassica rapa* (field mustard), *Capsella rubella*, *Eutrema salsugineum*, *Fragaria vesca* (woodland strawberry), *Theobroma cacao* (cacao) and *Vitis vinifera* (wine grape), were taken as reference organisms for pathway analysis using the KAAS server. Then, KO\_ID assignment of transcripts using KEGG pathway analysis was carried out to identify genes of different secondary metabolite pathways (Figs. S4–S6). We focused on screening of genes involved in terpenoid biosynthetic pathway. During the process, 45 KO\_IDs related to the terpenoid pathway were assigned to a total of 67 transcripts. For the terpenoid pathway, KEGG pathway analysis indicated the presence of several terpene synthases, such as monoterpene pathway-related transcripts, namely, terpineol synthase, linalool synthase, ocimene synthase and myrcene synthase; sesquiterpene pathway-related genes, namely, germacrene D synthase and farnesene synthase; and diterpenoid pathway genes, namely, geranyl-linalool synthase and ent-kaurene synthase. In addition,

phenylpropanoid pathway transcripts were also mined for KEGG pathway analysis. Relevant information can be found in the Supplementary material.

Virtual Ribosome, a web-based server, was also used for finding the Open Reading Frame (ORF) of transcripts. The virtual ribosome technique was used to convert a total of 55,995 clustered transcript sequences into 48,320 peptide sequences, which were then subjected to Pfam analysis (Table S2). Of these, 43,483 peptides had lengths between 100 and 500 amino acids, which is the ideal range for proteins involved actively in cellular processes. These 48,320 submitted peptides yielded 40,745 predicted proteins belonging to different protein domains and families. Proteins belonging to the terpene synthase family were screened out by searching for two essential domains: PF01397 (the N-terminal domain) and PF03936 (the C-terminal domain or metal binding domain). A total of 24 transcripts contained these conserved domains. Among these, eight transcripts had both domains but were missing a few bases towards the N- and C-terminal ends; 13 transcripts were missing the N-terminal end; and three transcripts were missing the C-terminal end. Further, blastx studies for these total 24 transcript sequences were carried out for homology-based annotation of putative gene function. Monoterpene pathway-related genes, namely alpha-terpineol synthase, linalool synthase, geranyl linalool synthase, beta-ocimene synthase, and myrcene synthase, were identified. Sesquiterpene pathway-related genes, i.e. germacrene D synthase and farnesene synthase, were identified. The diterpenoid pathway gene, namely, ent-kaurene synthase, was identified. These results are in agreement with terpenoid metabolite profiling of flower tissues. Among the transcripts, we were able to clone

**Table 1.** *Couroupita guianensis* Aubl flower transcriptomic analysis statistics

Parameters	Assembled_transcripts	Clustered_transcripts
Number of transcripts identified	89,588	55,995
Maximum contig length	15,339	15,339
Minimum contig length	201	300
Average contig length	981.1 ± 1029.8	1208.3 ± 1050.3
Median contig length	1330	1479
Total contigs length	8,78,97,713	6,76,56,968
Total number of non-ATGC characters	0	0
Percentage of non-ATGC characters	0	0
Contigs ≥ 100 bp	89,588	55,995
Contigs ≥ 200 bp	89,588	55,995
Contigs ≥ 500 bp	47,081	39,342
Contigs ≥ 1 Kbp	29,483	24,035
Contigs ≥ 10 Kbp	19	11
N50 value	1740	1808

and sequence three candidate full-length terpene synthase genes (Table 2).

#### Primary structure analysis of full-length ORF of three terpene synthases

Three candidate terpene synthase genes, **A\_c43359\_g3\_i1**, **A\_c38347\_g2\_i1** and **A\_c45679\_g1\_i3**, with full-length ORFs, were each successfully cloned in pET-28a as well as pET-32a expression vectors and validated through sequencing and restriction digestion studies (Fig. S2). Further full-length ORFs of the three sequenced genes were subjected to Blastx analysis. These results indicated that **A\_c38347\_g2\_i1** is putative  $\alpha$ -farnesene synthase (sesquiterpene synthase), and henceforth **A\_c38347\_g2\_i1** is termed as *Cg\_Fs*; **A\_c43359\_g3\_i1** is putative  $\beta$ -ocimene synthase (monoterpene synthase), and henceforth **A\_c43359\_g3\_i1** is termed as *Cg\_Os*; **A\_c45679\_g1\_i3** is putative ent-kaurene synthase (diterpene synthase), and henceforth **A\_c45679\_g1\_i3** is termed as *Cg\_Ks*, respectively.

Physiochemical properties play an important role in determining protein functions. ExPASy's ProtParam tool was used for computing the physiochemical properties of all three genes. The molecular weight for cloned genes encoding terpene synthase proteins fell in the range 63.0–90.0 kD and pI fell in the range of 5.4–6.4, respectively. The three terpene protein sequences showed higher values of the aliphatic index (86.2–96.0) and lower values of Grand Average Hydropathy (GRAVY) (−0.186 to −0.284). The computed instability index for the three terpene protein sequences was >40. The estimated half-lives of these three proteins in different cell systems were predicted to be 30 h (mammalian reticulocytes, *in vitro*), >20 h (yeast, *in vivo*) and >10 h (*E. coli*, *in vivo*).

The multiple sequence alignment of amino acid sequences of cloned terpene synthases *Cg\_Fs*, *Cg\_Os* and *Cg\_Ks* was carried out with that of amino acid sequences of functionally characterized terpene synthases from the UniProt database using CLUSTALW. The results revealed the presence of two highly

conserved motifs, DDxxD and NSE/DTE motifs, in all three candidate genes (Fig. S3). Further, two more motifs, i.e. SAYDITAW and QxxDGSW, were also found in the putative ent-kaurene synthase of *C. guianensis* *Cg\_Ks* terpene sequences (Fig. S3).

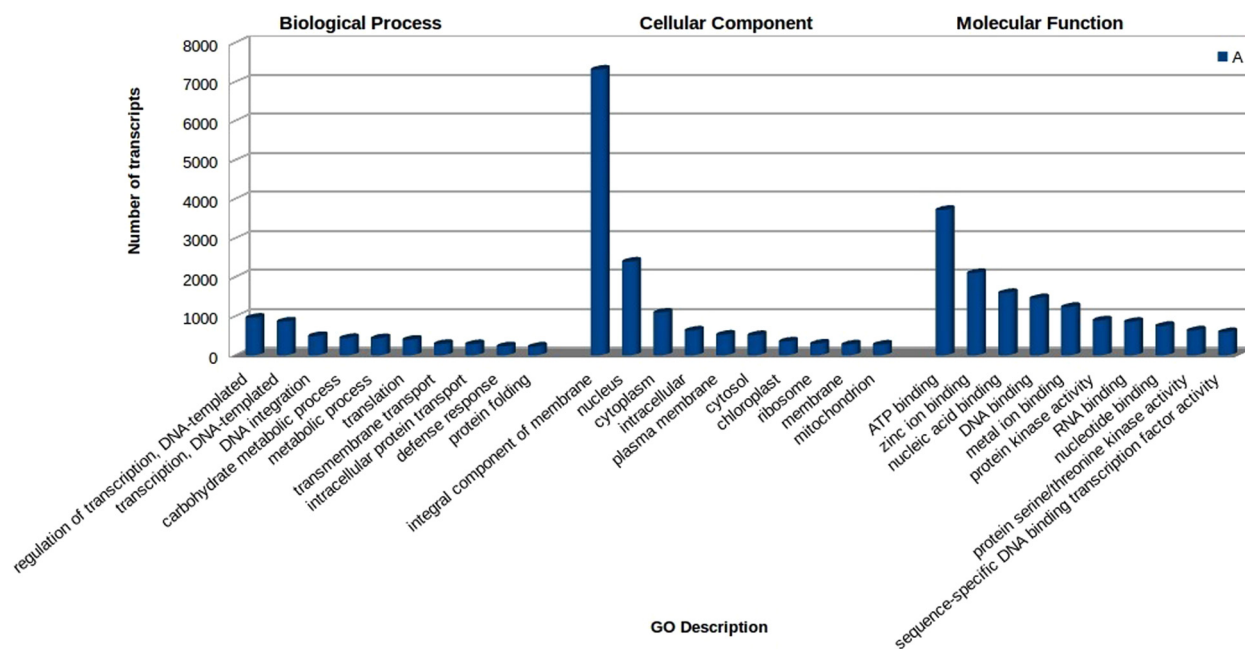
A phylogenetic tree of these three genes and similar proteins from different plant species was constructed using NgPhylogeny.fr to investigate the evolutionary relations. *Cg\_Fs* is grouped with other farnesene synthase genes. The *Cg\_Os* is grouped with ocimene synthases in a phylogenetic tree. Both sequences belonged to the TPS-b family group and shared a common ancestor. *Cg\_Ks* is grouped with other ent-kaurene synthase sequences in family groups TPS-e, f. Both TPS-b and TPS-e, f share a common evolutionary origin (Fig. 3).

#### Gene expression analysis in *C. guianensis* tissue

The three terpene synthase genes, i.e. *Cg\_Os* (putative monoterpene tricyclene/ $\beta$ -ocimene synthase), *Cg\_Fs* (putative sesquiterpene  $\alpha$ -farnesene synthase) and *Cg\_Ks* (putative diterpene ent-kaurene synthase), were analysed for their expression in flower tissue. Semi-quantitative PCR analysis verified their expression in flower tissue. Among the three terpene synthase genes, *Cg\_Os* showed the highest expression in flower tissues. In comparison, *Cg\_Fs* had an expression level half that of *Cg\_Os*. Further, *Cg\_Ks* showed the least expression, which was around 18-fold less than that of *Cg\_Os* in flower tissues (Fig. 4). Semiquantitative RT-PCR analysis verified the expression of three cloned terpene synthases in flower tissue with the biosynthetic potential to produce terpenes.

#### Protein structure-based function prediction

After cloning terpene synthase genes, we generated 3D protein structures to predict protein function and conduct *in silico* structural studies. Homology-based protein models of the three terpene synthases were constructed and validated (Fig. 4). Based on the best fit, the crystal structure of limonene synthase from *Citrus sinensis*



**Figure 2.** Analysis of *C. guianensis* unigenes. Top 10 GO terms for three categories.

(PDBID: 5uv0.1A) was used as a template for modelling *Cg\_Fs* and *Cg\_Os* terpene synthases. The crystal structure of abietadiene synthase from *Abies grandis* (PDBID:3s9v.1A) was used as a template for modelling *Cg\_Ks* terpene synthase. The models generated were validated using PDB Sum, which generated a Ramachandran plot and evaluated all its constraints (Fig. 4). For *Cg\_Fs*, *Cg\_Os* and *Cg\_Ks*, respectively, 92.8, 93.3 and 90.4% of residues were observed in the favoured regions, whereas 6.6, 6.1 and 8.9% of residues were observed in the allowed regions. The protein models were deemed to be of good quality when 90% or more of the residues were found in the Ramachandran plot's preferred regions. The *G*-factor was in the optimal range for high-quality protein models, which was between  $-1.0$  and  $0.1$ .

In the homology models of putative *Cg\_Os* and *Cg\_Fs* proteins constructed in the study,  $\alpha\beta$  domains with a DDxxD motif in the  $\alpha$  domain can be seen, which is a characteristic feature of proteins of the type I TPS terpene synthase gene family. In the case of a homology model of *Cg\_KS*, all three  $\alpha\beta\gamma$  domains can be seen. The DDxxD motif was found in the  $\alpha$  domain, whereas the DxDD motif was absent.

The validated homology models of terpene synthases were further analysed by the ProFunc web server for protein function prediction from the 3D structure. The results of the homology search are summarized in Table S2. ProFunc predicted GO terms associated with the three terpene synthases indicate that all three proteins have metal-binding capacities and take part in cellular metabolite processes. 'Enzyme active site template'-based homology search of 3D structure by ProFunc analysis revealed that *Cg\_Fs* and *Cg\_Os* terpene synthases had high similarity with sesquiterpene synthase, namely, homo5-epi-aristolochene synthase from *Nicotiana tabacum*. Further, as shown in Table S2, a reverse template 3D structure-based search by ProFunc indicated that *Cg\_Fs* and *Cg\_Os* had high similarity with a hemiterpene synthase. A 'protein 3D structure enzyme active site template'-based homology search by ProFunc for *Cg\_Ks* predicted it to

have pentalene synthase activity. Reverse template 3D structure-based ProFunc search revealed *Cg\_Ks* to have high similarity with ent-copalyl diphosphate synthase (diterpene synthase) from *A. thaliana*.

## Discussion

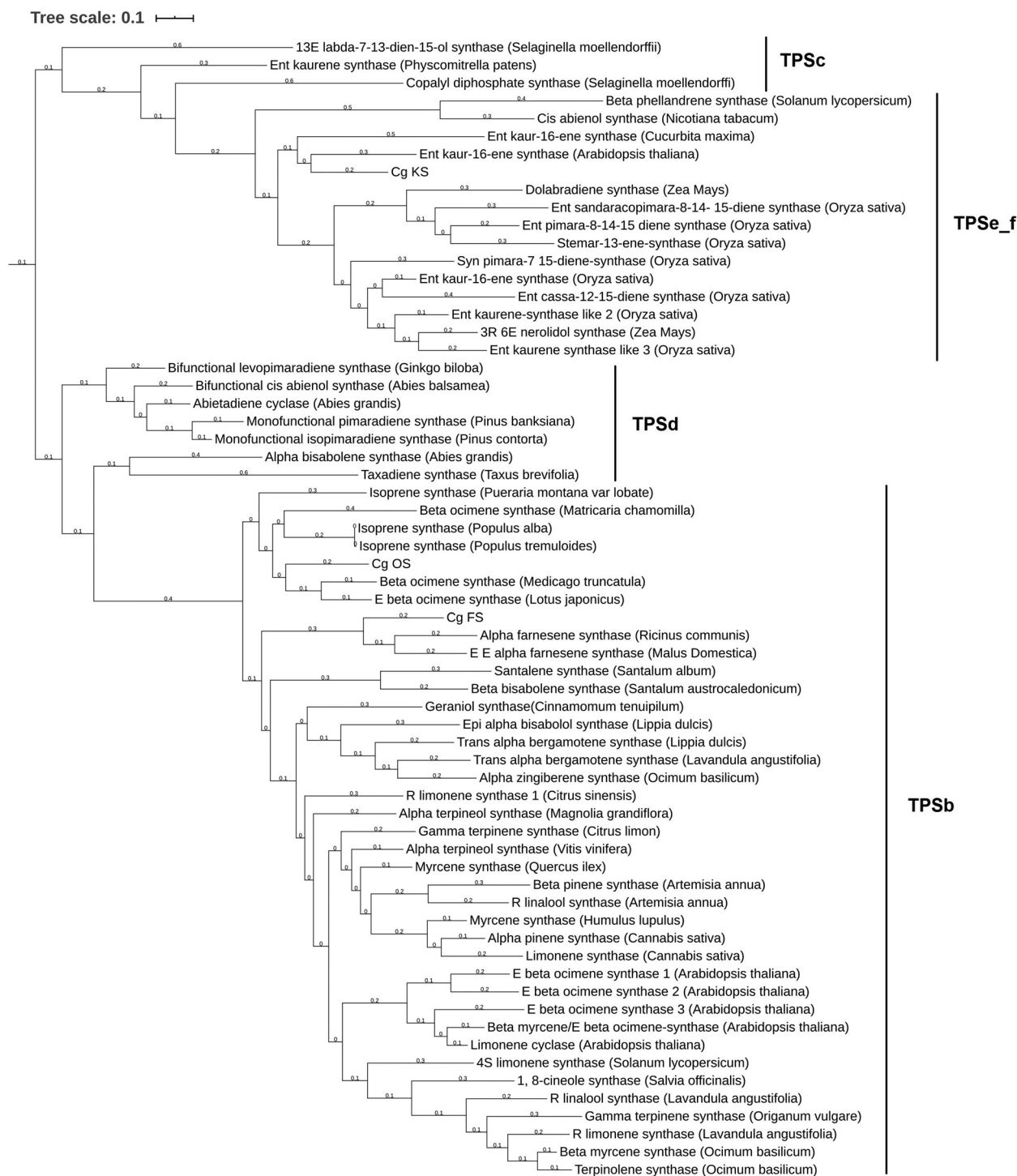
### Metabolite profiling

Metabolite profiling of different plant parts showed that flower tissue contained the highest terpene and phenolic content. The terpene volatiles linalool, ocimene and farnesene, as well as the phenylpropanoid/benzenoid volatiles eugenol, isatin and phenylethyl alcohol, were the most abundant metabolites in flower tissue. These volatiles could contribute to the fragrance as well as the different biological activities of a flower. Terpenoids and phenylpropanoid/benzenoid compounds have been found as major constituents of the flower tissue of many different plants (Knudsen *et al.*, 2006; Dhandapani *et al.*, 2021). Further, many reports have also substantiated the terpene linalool as a ubiquitous floral volatile. It is implicated in diverse functions, from a toxin involved in plant defence to long distance pollinator attraction (Raguso, 2016). In agreement, the *C. guianensis* flower showed the presence of phenylpropanoid and terpenoid groups as major volatile constituents.

It is specifically the flower of the plant that is used in sacred ceremonies in Indian and Asian cultures, and no other parts like the stem and leaf. To comprehend this traditional wisdom, we decided to assess the anti-microbial activity of flower, petal, stamen, stem and leaf tissues of *C. guianensis* in order to acquire a sense of their comparative bioactive potential & (Wiegand *et al.*, 2008; Mann and Markham, 1998). In our study, the flower showed higher bioactive potential compared to the stem and leaf. Previously, many studies have reported the antimicrobial

**Table 2.** The general statistics of assembled terpene transcript analysis for the generation of Pfam\_ID

Sr.	PF01397 (N-terminal domain)	PF03936 (C-terminal domain)	Missing end	Blastx studies for respective transcripts	Query cover, %	E value	Per. ident, %	Accession
1	A_c31106_g1_i1	ND	Both ends missing	(-)-Germacrene D synthase [ <i>Vitis vinifera</i> ]	89	2.00E-39	59.86	RWW61713.1
2	A_c43359_g2_i1	ND	Both ends missing	Isoprene synthase [ <i>Populus deltoides</i> ]	60	1.00E-14	50.00	AEK70966.1
3	A_c43429_g2_i4	ND	Both ends missing	P(E)-nerolidol/(E,E)-geranyl linalool synthase [ <i>Vitis vinifera</i> ]	91	1.00E-24	55.43	NP_001268004.1
4	A_c43663_g1_i1	ND	C-ter and middle	(E)-beta-ocimene/myrcene synthase [ <i>Vitis vinifera</i> ]	63	7.00E-145	52.30	ADR74206.1
5	A_c43663_g1_i2	ND	C-ter and middle	(-)-Alpha-terpineol synthase [ <i>Vitis vinifera</i> ]	61	2.00E-149	51.57	RWW71174.1
6	A_c43663_g1_i3	ND	C-ter and middle	(-)-Alpha-terpineol synthase [ <i>Vitis vinifera</i> ]	62	0	51.57	RWW71174.1
7	A_c43663_g1_i5	ND	C-ter and middle	(-)-Alpha-terpineol synthase [ <i>Vitis vinifera</i> ]	60	3.00E-108	52.45	NP_001268216.1
8	A_c43663_g1_i6	ND	C-ter and middle	(-)-Alpha-terpineol synthase [ <i>Vitis vinifera</i> ]	63	7.00E-176	52.45	NP_001268216.1
9	A_c43663_g1_i7	ND	C-ter and middle	(-)-Alpha-terpineol synthase [ <i>Vitis vinifera</i> ]	51	1.00E-151	51.57	RWW74233.1
10	A_c43663_g1_i8	ND	C-ter and middle	(-)-Alpha-terpineol synthase [ <i>Vitis vinifera</i> ]	54	7.00E-148	53.52	RWW22702.1
11	A_c43663_g1_i9	ND	seems full length	Linalool synthase [ <i>Gossypium hirsutum</i> ]	60	3.00E-124	44.89	AJT59543.1
12	A_c43663_g1_i10	ND	C-ter and middle	(-)-Alpha-terpineol synthase [ <i>Vitis vinifera</i> ]	56	6.00E-143	54.42	RWW76571.1
13	A_c43663_g1_i11	ND	C-ter and middle	(-)-Alpha-terpineol synthase [ <i>Vitis vinifera</i> ]	64	0	51.57	RWW71174.1
14	ND	A_c991_g1_i1	N-ter missing	Beta-caryophyllene synthase [ <i>Vitis vinifera</i> ]	99	3.00E-40	60.75	QBL52480.1
15	ND	A_c37933_g1_i1	Both ends missing	(-)-Alpha-terpineol synthase [ <i>Vitis vinifera</i> ]	99	2.00E-56	65.71	RWW80387.1
16	ND	A_c43359_g3_i2	N-ter missing	(E)-beta-ocimene synthase [ <i>Malus domestica</i> ]	65	1.00E-125	62.88	AGB14628.1
17	A_c10978_g1_i1	A_c10978_g1_i1	Both ends missing	Terpene synthase [ <i>Camellia sinensis</i> ](germacrene D synthase)	100	1.00E-62	67.24	AFE56211.1
18	A_c31106_g2_i1	A_c31106_g2_i1	N-ter missing	(-)-Germacrene D synthase [ <i>Vitis vinifera</i> ]	90	3.00E-147	50.86	RWW94686.1
19	A_c38347_g2_i1	A_c38347_g2_i1	N-ter missing	(E)-beta-ocimene/(E, E)-alpha-farnesene synthase [ <i>Vitis vinifera</i> ]	85	0	57.69	ADR74207.1
20	A_c40990_g1_i1	A_c40990_g1_i1	Both ends missing	Myrcene synthase, chloroplastic-like [ <i>Vitis vinifera</i> ]	77	0	59.33	NP_001268009.1
21	A_c40990_g1_i2	A_c40990_g1_i2	Both ends missing	3R-linalool synthase, putative isoform 1 [ <i>Theobroma cacao</i> ]	76	1.00E-145	57.75	EOY18953.1
22	A_c43359_g3_i1	A_c43359_g3_i1	N-ter missing	Terpene synthase 2 [ <i>Camellia sinensis</i> ]	70	0	65.12	ANB66347.1
23	A_c43429_g2_i2	A_c43429_g2_i2	N-ter missing	(E,E)-geranylinalool synthase [ <i>Vitis vinifera</i> ]	92	0	58.67	RVX08505.1

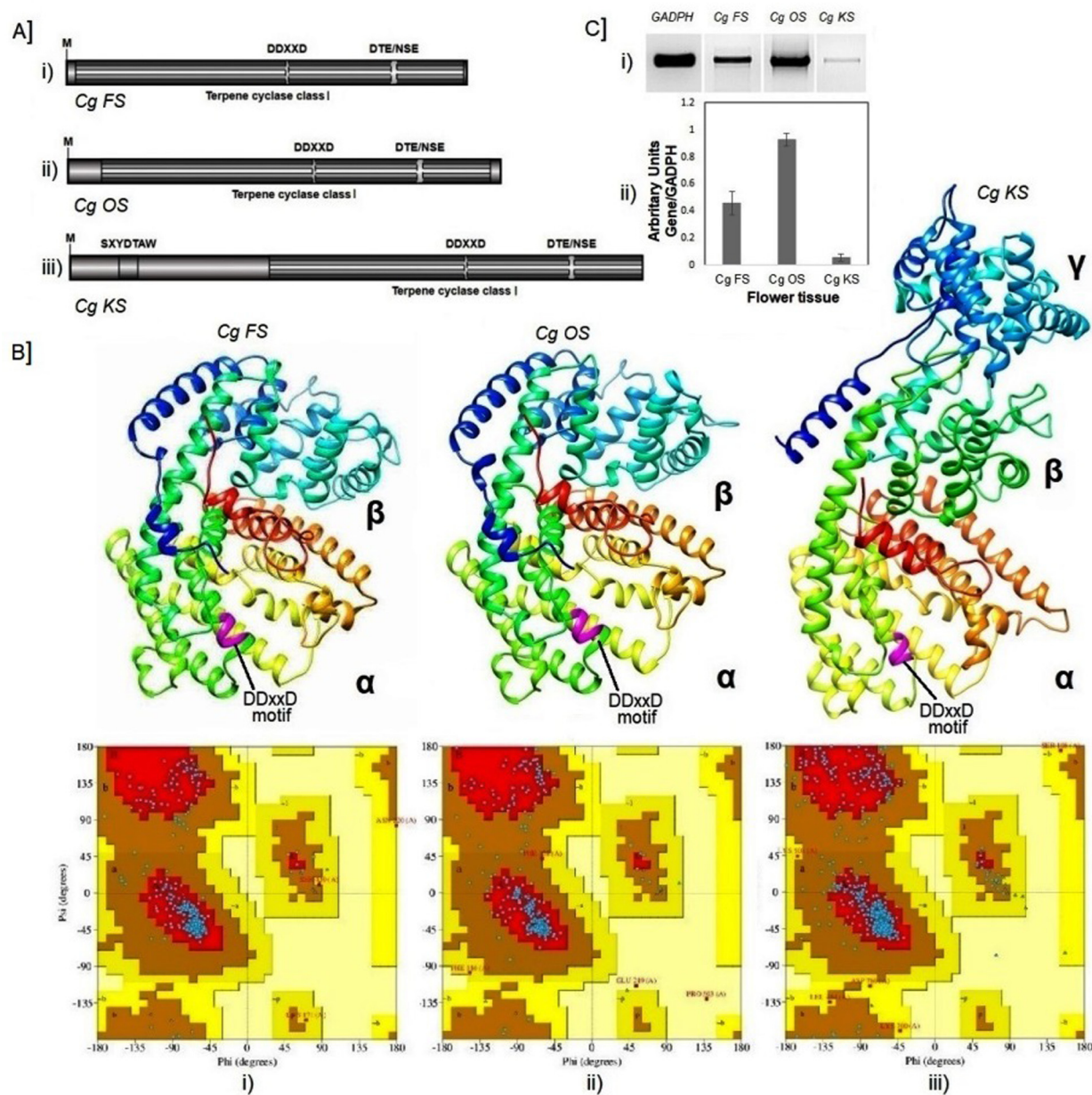


**Figure 3.** Phylogenetic analysis of three cloned *C. guianensis* terpene synthase genes with reviewed terpene synthase sequences in UNIPROT database using NgPhylogeny.Fr.

activity of *C. guianensis* plant extracts against many Gram-positive and Gram-negative bacteria. In one such study, methanol extracts of leaves, flowers, fruit, stem and roots of the plant inhibited the growth of the microorganisms (Khan *et al.*, 2003). In another study, chloroform extracts from flowers also showed antimicrobial activity (Al-Dhabi *et al.*, 2012). However, no metabolite profiling was reported for any of these tissues. Our analysis of metabolite profiles identified the main volatile components in the *C.*

*guianensis* flower. These metabolites have previously been found to have potent antimicrobial properties (Pauli and Kubeczka, 2010; Chouhan *et al.*, 2017; Caulier *et al.*, 2019; Khameneh *et al.*, 2019). Accordingly, metabolites could be connected to the bioactive potential of flower, which could result in a variety of documented pharmacological effects (Sanz-Biset *et al.*, 2009; Al-Dhabi *et al.*, 2012; Shekhawat and Manokari, 2016; Kaneria *et al.*, 2017).





**Figure 4.** Gene structure analysis of selected terpene synthase genes along with predicted protein 3D structure and RNA expression analysis in flower tissue: (A) key features of cloned and sequenced terpene synthase genes: (i) *Cg FS*: putative  $\alpha$ -farnesene synthase (sesquiterpene synthase); (ii) *Cg OS*: putative  $\beta$ -ocimene synthase (monoterpene synthase); (iii) *Cg KS*: putative *ent*-kaurene synthase (diterpene synthase). (B) Predicted protein 3D structure of three cloned terpene synthases using homology modelling along with Ramachandran plots. (i) *Cg FS*; (ii) *Cg OS*; (iii) *Cg KS*. (C) Semi-quantitative RT-PCR analysis of selected terpene synthase genes in flower tissue: (i) agarose gel electrophoresis of RT-PCR products (lanes [1] GADPH; [2] *Cg FS*; [3] *Cg OS*; [4] *Cg KS*); (ii) densitometry quantification of RT-PCR products depicted as arbitrary units.

### Transcriptomic analysis

A great diversity of volatile terpenes was identified to make up flower tissue, compared to phenylpropanoid/benzenoids that were shown to be less diverse. Thus, we focused on the transcriptome profiling of the terpenoid pathway in flower tissue. The transcriptome of *C. guianensis* flower revealed the occurrence of many terpene pathway-related genes, which strongly correlated with the terpenoid profile of *C. guianensis* flower tissues in the study. In plants, the biosynthesis of terpenoids arises from the methylerythritol 4-phosphate (MEP) pathway in

plastids and/or the mevalonate (MVA) pathway in the cytosol. The first committed step is the condensation of IPP and DMAPP into geranyl diphosphate (GPP, C10), farnesyl diphosphate (FPP, C15) and geranylgeranyl diphosphate (GGPP, C20), which are precursors for the production of mono-, sesqui- and diterpenes, respectively. Then the final cyclization and oxidation steps are carried out by the terpene synthases (TPS) and cytochrome P450s (CYP450) to generate diverse terpene structures (Srivastava *et al.*, 2015).

The transcripts of terpenoid pathway enzymes were found to be expressed in the *C. guianensis* flower. These transcripts were

monoterpene pathway-related transcripts, namely terpineol synthase, linalool synthase, ocimene synthase and myrcene synthase; sesquiterpene pathway-related genes, namely germacrene D synthase and farnesene synthase; and diterpenoid pathway-related genes namely geranyl linalool synthase and ent-kaurene synthase. These transcripts may be involved in the production of terpenoids confirmed in the flower tissue.

Thus, we created a transcriptomic resource for *C. guianensis* in this study, enabling us to mine numerous nucleotide and protein sequences implicated in the biosynthesis of terpenoids in flower tissue.

### Primary structure analysis of full-length ORF of three terpene synthases

After screening out potential terpene synthase gene candidates that may be part of the terpene biosynthetic pathway, we carried out on detailed analysis of *Cg\_Os* (putative monoterpene tricyclene/beta-ocimene synthase), *Cg\_Fs* (putative sesquiterpene alpha-farnesene synthase) and *Cg\_Ks* (putative diterpene ent-kaurene synthase) terpene synthase gene candidates, which serve as entry point enzymes for several terpenoid biosynthesis routes.

Multiple sequence alignments of *Cg\_Fs*, *Cg\_Os* and *Cg\_Ks* genes with known terpene synthase genes revealed the presence of two highly conserved terpene synthase motifs, DDxxD and NSE/DTE motifs. The DDxxD motif is involved in the coordination of divalent metal ions ( $Mg^{2+}$ ) for substrate binding. NSE/DTE is also reported to be a consensus sequence (L, V) (V, L, A) (N, D) D (L, I, V) x (S, T) x x x (E) and a second divalent cation ( $Mg^{2+}$ ) binding site in terpenoid synthases in all three sequences (Bohlmann *et al.*, 1998; Gao *et al.*, 2012). Further, two more motifs, i.e. SAYDTAW and QxxDGSW, were found in the putative ent-kaurene synthase of *C. guianensis*. These motifs are found to be highly conserved among ent-kaurene synthase proteins (Kim *et al.*, 2009; Alqazar *et al.*, 2017). The functional role of these conserved motifs in ent-kaurene synthases remains elusive, although QxxDGSW motifs in a bacterial squalene-hopene cyclase are involved in the stabilization of the whole protein (Wendt *et al.*, 1997). Another important motif worth mentioning is the DxDD motif, which also mediates the initial protonation of the substrate in coordination with divalent cations ( $Mg^{2+}$ ) (Zhou and Pichersky, 2020). Multiple sequence alignments revealed that the *Cg\_Ks* transcript does not possess a conserved DxDD motif; thus, it can be annotated as a monofunctional ent-kaurene synthase.

Physicochemical properties play an important role in determining protein functions. ExPasy's ProtParam tool helped predict the molecular weight, pI, aliphatic index and GRAVY score of three cloned sequences. Recently, general terpene synthase structure and function were reviewed (Tholl, 2006; Rafiqi *et al.*, 2019). In general, terpene synthase cDNAs encode proteins of 550–850 aa, leading to molecular masses of 50–100 kDa. The pI is the pH value at which a protein is neutral, i.e. it has zero net charge. Terpene synthases bear zero net charges at pH 5–6. An aliphatic index is an indicator of the thermostability of proteins. The three terpene proteins showed thermostability within a wider temperature range. Our calculated parameters for cloned genes encoding terpene synthase proteins are in agreement with consensus values for terpene synthases (Tholl, 2006). The GRAVY value of a protein is a measure of the interaction of a particular protein with water. The lower values of GRAVY of these three terpene synthases indicate the possibility of better interaction with

water. The instability index evaluates the stability of a protein *in vitro*. Our three terpene protein sequences were predicted to be highly unstable proteins *in vitro*. The estimated half-lives of these three proteins in different cell systems indicated that their expression was stable for many hours. Such information on the physicochemical properties of predicted proteins is useful when utilising and characterising proteins for bioinformatics, biochemistry and biotechnology analysis.

Phylogeny analysis helps us accurately represent how molecular function evolved for any particular set of protein, and is thus often used for function predictions supported by evolutionary principles (Eisen, 1998). Phylogenetic tree analysis helped us gain insight into the evolutionary history of these three cloned terpene synthases compared to previously known terpene synthases. *Cg\_Fs* was grouped with other farnesene synthase genes. *Cg\_Os* was grouped with ocimene synthases in a phylogenetic tree. Both sequences belonged to the TPS-b family group and shared a common ancestor. *Cg\_Ks* is grouped with other ent-kaurene synthase sequences in family groups TPS-e, f. Both TPS-b and TPS-e, f share a common evolutionary origin. Phylogenetic analysis results were in agreement with Blastx and multiple sequence alignment results.

### Gene expression analysis in *C. guianensis* tissue

The three terpene synthase genes, i.e. *Cg\_Os* (putative monoterpene tricyclene/beta-ocimene synthase), *Cg\_Fs* (putative sesquiterpene alpha-farnesene synthase) and *Cg\_Ks* (putative diterpene ent-kaurene synthase), were examined for their expression in flower tissue using semi-quantitative PCR. Among the three terpene synthase genes, *Cg\_Os* showed the highest expression in flower tissues. In comparison, *Cg\_Fs* had an expression level half that of *Cg\_Os* in flower tissues. The expression of *Cg\_Ks* was the lowest in flower tissues, almost 18-fold lower than that of *Cg\_Os*.

Metabolite profiling results reveal ocimene and  $\alpha$ -farnesene to be present in flower tissue at 0.16 and 0.4%, respectively. Metabolite profiling did not confirm the presence of the kaurene metabolite. Many studies have suggested that for some metabolites, high gene expression may or may not translate to high metabolite content (Iijima *et al.*, 2004; Redestig and Costa, 2011). This could partly be due to transcriptional or post-translational regulatory factors limiting enzyme activity and, therefore, metabolite biosynthesis at the levels determined. Finally, semiquantitative RT-PCR analysis verified the expression of *Cg\_Os*, *Cg\_Fs* and *Cg\_Ks* cloned terpene synthases in terpene-producing flowers of *C. guianensis*.

### Protein structure-based function prediction

Proteins are linear chains of amino acids that fold into exceedingly complex three-dimensional structures, depending on the sequence and physical interactions within the chain. The structure, in turn, determines the ultimate biological function of proteins as well as their interactions. Homology-based protein models of the three terpene synthases were constructed and validated. The Ramachandran plot score suggested the refined models were of good quality (Greener *et al.*, 2017). The *G*-factor provides a measure of how 'normal', or 'unusual', a given stereochemical property, i.e. bonds, is in protein structure. If a protein has many residues with low *G*-factors, it indicates a less stereochemically valid structure. Ideally, *G* values should be above  $-0.5$

(Rising *et al.*, 2020). For the three predicted models, the G-factor value indicated satisfactory geometry.

Generally, the plant terpene synthase TPS family consists of two types of domains, i.e.  $\alpha\beta$  or  $\alpha\beta\gamma$  (Zhou and Pichersky, 2020). These domains can be traced from the N-terminus to the C-terminus in the forward direction as  $\gamma$ ,  $\beta$  and  $\alpha$ . Type I TPSs have the conserved DDxxD motif in the  $\alpha$  domain, while type II TPSs have the conserved DxDD motif in the  $\beta$  domain. A recent review by Zhou and Pichersky (2020) provides a detailed understanding of the 3D structure of proteins in the terpene synthase gene family. Based on homology models of putative *Cg\_Os*, *Cg\_Fs* and *Cg\_Ks* proteins, it can be predicted that they belong to the type I TPS terpene synthase gene family.

The validated homology models of terpene synthases were further analysed by the ProFunc web server for protein function prediction from their 3D structures. ProFunc is a web server for predicting the likely function of proteins using predicted homology models of 3D protein structure. ProFunc makes use of the protein sequence alignment, conserved motif features, enzyme active site and ligand-binding site comparisons in the 3D structure of known proteins, etc., to functionally characterize proteins. All three predicted protein structures, *Cg\_Os*, *Cg\_Fs* and *Cg\_Ks*, had metal-binding capacities and took part in cellular metabolite processes. 'Protein 3D structure enzyme active site template'-based homology search compares against manually curated residues in PDB known from the literature to be catalytic. This search analysis gave a strong prediction for sesquiterpene synthase capability for both *Cg\_Fs* and *Cg\_Os*. Reverse 3D structure template-based search uses hundreds of small residue reverse templates generated by breaking down the target structure. These are then scanned against a representative set of the structures in the PDB. The approach tends to match functionally important sites. This search gave a prediction for hemiterpene synthase activity for both *Cg\_Fs* and *Cg\_Os*.

Recently, many studies have highlighted biochemical reaction similarities between isoprene synthase (hemiterpene synthase) and farnesene synthase (sesquiterpene synthase) (Koksal *et al.*, 2010). A study involving Poplar isoprene synthase expression revealed that the chemistry of the elimination step yielding isoprene is identical to that yielding farnesene from farnesyl diphosphate (Pazouki and Niinemets, 2016). In an earlier study, it was reported that isoprene synthase and  $\beta$ -ocimene synthase formed a monophyletic group within the TPS-b clade of terpene synthases (Sharkey *et al.*, 2013). In agreement, we also found *Cg\_Os* and several isoprene synthases in the TPS-b group in our phylogeny analysis. The chemistry of isoprene synthase and ocimene synthase is reported to be similar and likely affects the phylogenetic relationships among TPS-b enzymes (Koksal *et al.*, 2010; Faraldos *et al.*, 2012).

'Protein 3D structure enzyme active site template'-based homology search for *Cg\_Ks* predicted pentalene synthase activity. The enzyme pentalene synthase catalyses the cyclization of farnesyl diphosphate into pentalene, a tricyclic sesquiterpene that is the hydrocarbon precursor of the pentalenolactone family of antibiotics (Irmisch *et al.*, 2015). A study dealing with detailed bioinformatics and crystalized 3D structure analysis of bacterial *Bradyrhizobium japonicum* kaurene synthase found that the protein structure had high homology with *epi*-aristolochene synthase from the plant *N. tabacum*, 1,8-cineole synthase from the plant *Salvia fruticose* and pentalene synthase from the bacterium *Streptomyces* (Liu *et al.*, 2015). The homology analysis revealed the DDxxD motif and ND(x)<sub>6</sub>(D/E) sequence to be conserved

in active sites in all of them (Liu *et al.*, 2015). The crystal structure of this pentalene synthase revealed that the active site is present in the  $\alpha$ -barrel active site and is proposed as a minimal terpenoid synthase fold preserved among a majority of terpenoid synthases in  $\alpha$  domain (Lesburg *et al.*, 1997). A reverse template-based ProFunc search revealed *Cg\_Ks* to have high similarity with *ent*-copalyl diphosphate synthase (diterpene synthase) from *A. thaliana*. The biosynthesis of diterpenoids starts with the conversion of GGPP into *ent*-copalyl diphosphate, catalysed by a type II enzyme, *ent*-copalyl diphosphate synthase. Subsequently, a class I enzyme, *ent*-kaurene synthase, converts *ent*-copalyl diphosphate to *ent*-kaurene (Zhou *et al.*, 2012). Type II terpene synthase enzymes are characterized by highly conserved DxDD motif. Type I diterpene synthases possess characteristic DDxxD and NSE/DTE motifs (Cho *et al.*, 2004). Bifunctional copalyl diphosphate and kaurene synthase also occur in nature, containing both DxDD and DDxxD motifs. During multiple sequence alignment of *Cg\_Ks*, it was confirmed that it does not possess the DxDD motif; thus, *Cg\_Ks* cannot be an *ent*-copalyl diphosphate synthase and is most likely a monofunctional *ent*-kaurene synthase of type I terpene synthase.

After taking into account the results of both the Blastx investigation and the ProFunc analysis, it was determined that putative *C. guianensis* *Cg\_Os*, *Cg\_Fs* and *Cg\_Ks* terpene synthases may exhibit a diverse range of catalytic properties. Earlier, many studies on several plant TPS genes revealed the existence of remarkable plasticity in terpenoid biosynthesis in higher plants (Yang *et al.*, 2022). There is a growing body of proof that many TPSs are multi-substrate enzymes capable of producing terpenes of different chain lengths depending on corresponding substrate availability, i.e. TPSs can form monoterpenes with GDP as the substrate and sesquiterpenes with FDP as the substrate (Gao *et al.*, 2012). Therefore, accurate prediction of the enzymatic products of terpene synthases solely based on the protein similarity of terpene synthases is often difficult. However, structural studies do offer insight into the possible range of catalytic activities that may exist in terpene synthases. In our case, we can predict *Cg\_Fs* to have isoprene or sesquiterpene (farnesene) synthase-like catalytic activity, *Cg\_Os* to have isoprene, monoterpene (ocimene) or sesquiterpene synthase-like catalytic activity, and *Cg\_Ks* to have pentalene or *ent*-kaurene synthase-like activity.

To better understand the conventional wisdom behind the revered status of *C. guianensis* flowers, the entire flower, petals, stamens, stem and leaf of *C. guianensis* were metabolically profiled in the beginning, and these tissues were also tested for antibacterial activity. The findings made it evident that flower tissue stood out from other tissues like stem and leaf due to its diverse terpenoid repertoire and strong antibacterial property. Encouraged by the findings, we concentrated our efforts on a thorough examination of floral tissue and constructed a flower transcriptome by RNA sequencing to reveal terpene metabolite pathway genes in flower.

Finally, three full-length terpene synthase gene candidates representing a putative monoterpene synthase, a putative diterpene synthase and a putative sesquiterpene synthase were cloned and sequenced. These candidates are entry point enzymes for several terpenoid biosynthesis routes. The transcript expression of three cloned terpene synthase genes was also verified in flower tissue. Furthermore, we used a variety of fast and accessible bioinformatics methods for rapid terpene synthase gene function prediction. With the use of these three gene sequences, we were able to predict protein function at the level of the 3D structure and conduct *in silico* structural

investigations to better understand the range of terpene synthesising catalytic capabilities.

To the best of our knowledge, *C. guianensis* is an underexplored medicinal plant in terms of transcriptomics for secondary metabolite biosynthetic pathway studies. Our study was carried out with an exploratory perspective to characterize previously unstudied *C. guianensis* at metabolite and transcriptome level in detail. We have generated a transcriptomic resource for the plant to unravel hidden gene sequences involved in terpene production in flower tissue. The study can pave the way for translational work in the fields of protein engineering and metabolic engineering, where potential terpene synthase genes can be functionally validated and heterologous production of *C. guianensis* terpenes can be attempted in industrially friendly host systems in future.

**Supplementary material.** The supplementary material for this article can be found at <https://doi.org/10.1017/S1479262123000953>.

**Availability of data.** Raw reads generated from RNA sequencing of *C. guianensis* Aubl flower tissue were deposited at NCBI's SRA database with accession number PRJNA715623.

**Acknowledgements.** S. J. K. would like to thank ICMR for fellowship. P. S. would like to thank CSIR-Research Associate Fellowship (31/11(953)/2017-EMRI). We would also like to thank CSIR-National Chemical Laboratory for funding the research work through projects CSC0130 and CSC0106. We are grateful to Director, CSIR-National Chemical Laboratory, India for infrastructure and research facility. The authors declare no conflict of interest.

**Author contributions.** S. J. K. and P. S. performed all major experiments, data analysis and manuscript writing. S. J. K. handled transcriptomics, gene cloning and gene expression analysis. P. S. carried out RNA isolation, transcriptomics, metabolite profiling and protein structural bioinformatics. A. K. and A. P. helped in transcriptomics along with S. J. K. S. R. helped RNA isolation and antimicrobial activity along with P. S. Work was planned, supervised and critically analysed by H. V. T.

## References

- Alami MM, Ouyang Z, Zhang Y, Shu S, Yang G, Mei Z and Wang X (2022) The current developments in medicinal plant genomics enabled the diversification of secondary metabolites' biosynthesis. *International Journal of Molecular Sciences* **23**, 15932.
- Al-Dhabi NA, Balachandran C, Raj MK, Duraipandiyar V, Muthukumar C, Ignacimuthu S, Khan IA and Rajput VS (2012) Antimicrobial, antimycobacterial and antibiofilm properties of *Couroupita guianensis* Aubl. fruit extract. *BMC Complementary and Alternative Medicine* **12**, 1–8.
- Alqazur B, Rodre-guez A, de la Pena M and Pena L (2017) Genomic analysis of terpene synthase family and functional characterization of seven sesquiterpene synthases from *Citrus sinensis*. *Frontiers in Plant Science* **8**, 1481.
- Bohlmann J, Meyer-Gauen G and Croteau R (1998) Plant terpenoid synthases: molecular biology and phylogenetic analysis. *Proceedings of the National Academy of Sciences* **95**, 4126–4133.
- Caulier S, Nannan C, Gillis A, Licciardi F, Bragard C and Mahillon J (2019) Overview of the antimicrobial compounds produced by members of the *Bacillus subtilis* group. *Frontiers in Microbiology* **10**, 302.
- Cho EM, Okada A, Kenmoku H, Otomo K, Toyomasu T, Mitsuhashi W, Sassa T, Yajima A, Yabuta G and Mori K (2004) Molecular cloning and characterization of a cDNA encoding entcassa-12, 15-diene synthase, a putative diterpenoid phytoalexin biosynthetic enzyme, from suspension cultured rice cells treated with a chitin elicitor. *The Plant Journal* **37**, 1–8.
- Chouhan S, Sharma K and Guleria S (2017) Antimicrobial activity of some essential oils present status and future perspectives. *Medicines* **4**, 58.
- Dhandapani S, Jin J, Sridhar V, Sarojam R, Chua NH and Jang IC (2021) Integrated metabolome and transcriptome analysis of *Magnolia champaca* identifies biosynthetic pathways for floral volatile organic compounds. *BMC Genomics* **18**, 1–18.
- Eisen JA (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research* **8**, 163–167.
- Faraldos JA, Gonzalez V, Li A, Yu F, Koksai M, Christianson DW and Allemann RK (2012) Probing the mechanism of 1, 4-conjugate elimination reactions catalyzed by terpene synthases. *Journal of the American Chemical Society* **134**, 20844–20848.
- Gao Y, Honzatko RB and Peters RJ (2012) Terpenoid synthase structures: a so far incomplete view of complex catalysis. *Natural Product Reports* **29**, 1153–1175.
- Gasteiger E, Hoogland C, Gattiker A, Wilkins MR, Appel RD and Bairoch A (2005) Protein identification and analysis tools on the ExPASy server. In Walker JM (ed.), *The Proteomics Protocols Handbook*. Springer Protocols Handbooks. USA: Humana Press, pp. 571–607.
- Greener JG, Filippis I and Sternberg MJE (2017) Predicting protein dynamics and allostery using multi-protein atomic distance constraints. *Structure* **25**, 546–558.
- Guo J, Huang Z, Sun J, Cui X and Liu Y (2021) Research progress and future development trends in medicinal plant transcriptomics. *Frontiers in Plant Science* **12**, 691838.
- Iijima Y, Davidovich-Rikanati R, Fridman E, Gang DR, Bar E, Lewinsohn E and Pichersky E (2004) The biochemical and molecular basis for the divergent patterns in the biosynthesis of terpenes and phenylpropenes in the peltate glands of three cultivars of basil. *Plant Physiology* **136**, 3724–3736.
- Irmisch S, Muller AT, Schmidt L, Gunther J, Gershenzon J and Kullner TG (2015) One amino acid makes the difference: the formation of *ent*-kaurene and 16 $\alpha$ -hydroxy-*ent*-kaurene by diterpene synthases in poplar. *BMC Plant Biology* **15**, 1–13.
- Kaneria M, Rakholiya K, Jakasania R, Dave R and Chanda S (2017) Metabolite profiling and antioxidant potency of *Couroupita guianensis* Aubl. using LC-QTOF-MS based metabolomics. *Research Journal of Phytochemistry* **11**, 150–169.
- Khameneh B, Iranshahy M, Soheili V and Bazzaz BSF (2019) Review on plant antimicrobials: a mechanistic viewpoint. *Antimicrobial Resistance & Infection Control* **8**, 1–28.
- Khan MR, Kihara M and Omoloso AD (2003) Antibiotic activity of *Couroupita guianensis*. *Journal of Herbs, Spices & Medicinal Plants* **10**, 95–108.
- Khan AM, Shivashankara KS and Roy TK (2014) Determining composition of volatiles in *Couroupita guianensis* Aubl. through headspace-solid phase micro-extraction (HS-SPME). *Journal of Horticultural Sciences* **9**, 161–165.
- Kim YS, Han JY, Lim S and Choi YE (2009) Ginseng metabolic engineering: regulation of genes related to ginsenoside biosynthesis. *Journal of Medicinal Plants Research* **3**, 1270–1276.
- Knudsen JT, Eriksson R, Gershenzon J and Stahl B (2006) Diversity and distribution of floral scent. *The Botanical Review*, **72**, 1.
- Koksai M, Zimmer I, Schmitzler JP and Christianson DW (2010) Structure of isoprene synthase illuminates the chemical mechanism of teragram atmospheric carbon emission. *Journal of Molecular Biology* **402**, 363–373.
- Kumar A, Mulge DS, Thakar KJ, Pandreka A, Warhekar AD, Ramkumar S, Sharma P, Upadrasta S, Shanmugam D and Thulasiram H (2023) Functional characterization of five triterpene synthases through de-novo assembly and transcriptome analysis of *Euphorbia grantii* and *Euphorbia tirucalli*. *bioRxiv*, 2023-04.
- Kumar S, Korra T, Thakur R, Arutselvan R, Kashyap AS, Nehela Y, Chaplygin V, Minkina T and Keswani C (2023) Role of plant secondary metabolites in defence and transcriptional regulation in response to biotic stress. *Plant Stress* **8**, 100154.
- Laskowski RA, Chistyakov VV and Thornton JM (2005) PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Research* **33**, 266–268.
- Laskowski RA, Jabłońska J, Pravda L, Vařeková RS and Thornton JM (2018) PDBsum: structural summaries of PDB entries. *Protein Science* **27**, 129–134.
- Lemoine F, Correia D, Lefort V, Doppelt-Azeroual O, Mareuil F, Cohen-Boulakia S and Gascuel O (2019) NGPhylogeny. fr: new generation phylogenetic services for non-specialists. *Nucleic Acids Research* **47**, 260–265.

- Lesburg CA, Zhai G, Cane DE and Christianson DW (1997) Crystal structure of pentalenene synthase: mechanistic insights on terpenoid cyclization reactions in biology. *Science* **277**, 1820–1824.
- Lim TK (ed) (2012) *Couroupita guianensis*. In *Edible Medicinal And Non Medicinal Plants*, vol. 3. Dordrecht: Springer, pp. 133–137. [https://doi.org/10.1007/978-94-007-2534-8\\_14](https://doi.org/10.1007/978-94-007-2534-8_14)
- Liu W, Feng X, Zheng Y, Huang CH, Nakano C, Hoshino T, Bogue S, Ko TP, Chen CC and Cui Y (2015) Structure, function and inhibition of *ent*-kaurene synthase from *Bradyrhizobium japonicum*. *Scientific Reports* **4**, 1–9.
- Mann CM and Markham JL (1998) A new method for determining the minimum inhibitory concentration of essential oils. *Journal of Applied Microbiology* **84**, 538–544.
- Navale GR, Sharma P, Said MS, Ramkumar S, Dharne MS, Thulasiram HV and Shinde SS (2019) Enhancing epi-cedrol production in *Escherichia coli* by fusion expression of *farnesyl pyrophosphate synthase* and *epicedrol synthase*. *Engineering in Life Sciences* **19**, 606–616.
- Pauli A and Kubeczka KH (2010) Antimicrobial properties of volatile phenylpropanes. *Natural Product Communications* **5**, 1387–1394.
- Pazouki L and Niinemets U (2016) Multi-substrate terpene synthases: their occurrence and physiological significance. *Frontiers in Plant Science* **7**, 1019.
- Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC and Ferrin TE (2004) UCSF Chimera visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **25**, 1605–1612.
- Rafiqi UN, Gul I, Saifi M, Nasrullah N, Ahmad J, Dash P and Abdin MZ (2019) Cloning, identification and in silico analysis of terpene synthases involved in the competing pathways of artemisinin biosynthesis pathway in *Artemisia annua* L. *Pharmacognosy Magazine* **15**, 38.
- Raguso RA (2016) More lessons from linalool: insights gained from a ubiquitous floral volatile. *Current Opinion in Plant Biology* **32**, 31–36.
- Rai A, Saito K and Yamazaki M (2017) Integrated omics analysis of specialized metabolism in medicinal plants. *The Plant Journal* **4**, 764–787.
- Redestig H and Costa IG (2011) Detection and interpretation of metabolite transcript coresponses using combined profiling data. *Bioinformatics* **27**, 357–365.
- Rising KA, Crenshaw CM, Koo HJ, Subramanian T, Chehade KAH, Starks C, Allen KD, res DA, Spielmann HP and Noel JP (2020) Formation of a novel macrocyclic alkaloid from the unnatural farnesyl diphosphate analogue anilino-geranyl diphosphate by 5-epi-aristolochene synthase. *ACS Chemical Biology* **10**, 1729–1736.
- Sanz-Biset J, Campos-de-la-Cruz J, Epiquián-Rivera MA and Canigueral S (2009) A first survey on the medicinal plants of the Chazuta valley (Peruvian Amazon). *Journal of Ethnopharmacology* **122**, 333–362.
- Schwede T, Kopp J, Guex N and Peitsch MC (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Research* **31**, 3381–3385.
- Sharkey TD, Gray DW, Pell HK, Breneman SR and Topper L (2013) Isoprene synthase genes form a monophyletic clade of acyclic terpene synthases in the TPS-b terpene synthase family. *Evolution: International Journal of Organic Evolution* **67**, 1026–1040.
- Shekhawat MS and Manokari M (2016) *In vitro* propagation, micromorphological studies and ex vitro rooting of cannon ball tree (*Couroupita guianensis* Aubl.): a multi-purpose threatened species. *Physiology and Molecular Biology of Plants* **22**, 131–142.
- Srivastava PL, Daramwar PP, Krithika R, Pandreka A, Shankar SS and Thulasiram HV (2015) Functional characterization of novel sesquiterpene synthases from Indian sandalwood *Santalum album*. *Scientific Reports* **5**, 1–12.
- Tholl D (2006) Terpene synthases and the regulation, diversity and biological roles of terpene metabolism. *Current Opinion in Plant Biology* **9**, 297–304.
- Wendt KU, Poralla K and Schulz GE (1997) Structure and function of a squalene cyclase. *Science* **277**, 1811–1815.
- Wiegand I, Hilpert K and Hancock REW (2008) Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. *Nature Protocols* **3**, 163.
- Yang S, Wang N, Kimani S, Li Y, Bao T, Ning G, Li L, Liu B, Wang L and Gao X (2022) Characterization of terpene synthase variation in flowers of wild aquilegia species from Northeastern Asia. *Horticulture Research* **9**, uhab020.
- Zhou F and Pichersky E (2020) More is better: the diversity of terpene metabolism in plants. *Current Opinion in Plant Biology* **55**, 1–10.
- Zhou K, Xu M, Tiernan M, Xie Q, Toyomasu T, Sugawara C, Oku M, Usui M, Mitsuhashi W and Chono M (2012) Functional characterization of wheat *ent*-kaurene (-like) synthases indicates continuing evolution of labdane-related diterpenoid metabolism in the cereals. *Phytochemistry* **84**, 47–55.