Plant Genetic Resources: Characterization and Utilization

cambridge.org/pgr

De novo assembly, annotation and molecular marker identification from the leaf transcriptome of *Ocimum gratissimum* L.

Tanuja[1], Nibir Ranjan Parasar[1], Ravichandiran Kumar[1], Purushothaman Natarajan[2] and Madasamy Parani[1]

[1]Genomics Laboratory, Department of Genetic Engineering, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu 603203, India and [2]Department of Biology, Gus R. Douglass Institute, West Virginia State University, Institute, West Virginia, USA

Research Article

Cite this article: Tanuja, Parasar NR, Kumar R, Natarajan P, Parani M (2021). *De novo* assembly, annotation and molecular marker identification from the leaf transcriptome of *Ocimum gratissimum* L. *Plant Genetic Resources: Characterization and Utilization* **19**, 469–476. https://doi.org/10.1017/S1479262121000563

Received: 28 November 2020
Revised: 22 September 2021
Accepted: 7 November 2021
First published online: 2 December 2021

Key words:
Basil; *de novo* assembly; *Ocimum gratissimum*; RNA sequencing; transcriptome

Author for correspondence:
Madasamy Parani,
E-mail: paranim@srmist.edu.in

Abstract

Ocimum gratissimum L. is a well-known medicinal plant with several therapeutic properties, but molecular studies on this species are lacking. Therefore, we have sequenced the whole transcriptome from the leaves of *O. gratissimum* and assembled 121,651 transcripts. The transcriptome of *O. gratissimum* was closely related to *Sesamum indicum* and *Erythranthe guttata* in congruence with the molecular phylogenetic relationships among these species. Further, 62,194 transcripts were annotated and classified according to the GO terms concerning the biological process, cellular component and metabolic function. In the KEGG pathway analysis, 34,876 transcripts were mapped to 149 pathways and 1410 of them were involved in the biosynthesis of secondary metabolites. In the phenylpropanoid pathway, 101 transcripts were associated with the biosynthesis of eugenol, the principal constituent of the essential oil of *O. gratissimum*. In the caffeine metabolism pathway, none of the transcripts was related to caffeine biosynthesis, supportive of the caffeine-free nature of *Ocimum*. Transcripts coding for the metallothionein were abundant in the leaves, supporting the observation that *O. gratissimum* is an accumulator of heavy metals. We also identified the 930 transcripts coding for 59 transcription factors families with myeloblastosis transcription factors being the most predominant. About 6500 simple sequence repeats were identified, which will be useful in DNA marker-based applications. This is the first report of the leaf transcriptome of *O. gratissimum*, which will serve as an essential resource for further molecular studies in this important medicinal species.

Introduction

Several species of *Ocimum* L. are called Tulsi or basil, and they have antioxidant, anti-inflammatory, antimicrobial, antidiabetic, anxiolytic, hepatoprotective, antitumor, gastroprotective, hyperlipidemic and antiplasmodic properties (Chattopadhyay, 1994; Lahon and Das, 2011; Manaharan *et al*., 2014; Saharkhiz *et al*., 2014; Parasuraman *et al*., 2015; Srivastava *et al*., 2016). There are more than 150 species in the genus *Ocimum*, and each species has a unique chemical composition and medicinal properties, which remain unexplored entirely. *Ocimum gratissimum* L., commonly known as clove basil, is also an important species in this genus with a high therapeutic value. It is a perennial shrub that grows 1–3 m of height, and it is distributed in tropical and subtropical areas, especially India and Africa. Due to its widespread use in India, it has got about 85 vernacular names in India with Ram Tulasi or its phonetic variants being common in Hindi, Kannada, Malayalam, Marathi, Sanskrit and Telugu (http://medicinalplants.in/).

O. gratissimum contains several phytochemicals such as eugenol, tannins, flavonoids, terpenoids, saponins, glycosides and reducing sugars (Ighodaro *et al*., 2010; Singh *et al*., 2016; Olamilosoye *et al*., 2018; Airaodion *et al*., 2019). Eugenol from *O. gratissimum* is an FDA-approved non-toxic inhibitor of advanced glycation end products (AGEs) used in the management of diabetes (Singh *et al*., 2016). The essential oils from the leaves of *O. gratissimum* are used as a larvicidal agent against *A. albopictus* mosquitoes (Sumitha and Thoppil, 2016). Phenolic compounds such as caffeic acid and its derivatives caftaric acid, chicoric acid and rosmarinic acid and the flavonoid compound, vicenin-2, from this species showed enhanced glucose-stimulated insulin secretion (GSIS) in the pancreatic islets isolated from mice (Casanova *et al*., 2017). The essential oil derived from *O. gratissimum* altered the permeability of the cell membrane and showed antimicrobial activity against gastroenteritis pathogens such as *Staphylococcus aureus*, *Escherichia coli, Salmonella typhimurium* and *Shigella flexneri* (Chimnoi *et al*., 2018).

The lack of sufficient genomic and transcriptomic data from *Ocimum* species is limiting the molecular studies on critical metabolic pathways in this group of medicinal plants. Partial genome sequences from short-read paired-end sequencing are available for *O. tenuiflorum* L. (syn. *O. sanctum* L.) and *O. tenuiflorum* subtype Krishna Tulsi (Rastogi *et al.*, 2014, 2015; Upadhyay *et al.*, 2015). Leaf transcriptomes are available for *O. tenuiflorum* subtypes Krishna and Rama Tulsi, *O. basilicum* L. and *O. americanum* L. (Rastogi *et al.*, 2014, 2015; Upadhyay *et al.*, 2015; Zhan *et al.*, 2016). Apart from these reports, differential gene expression profiling was carried out for the *O. basilicum* cv. 'Red Rubin' with purple leaves and *O. basilicum* cv. 'Tigullio' with green leaves (Torre *et al.*, 2016). However, there were not many molecular studies carried out in *O. gratissimum*. There were only 116 nucleotide sequences from *O. gratissimum* present in the GenBank database of NCBI as of July 2020. In this study, we report the whole transcriptome of the leaves of *O. gratissimum* based on RNA-seq data, and a detailed analysis of the genes, which are involved in phenylpropanoid and caffeine metabolism.

Material and methods

Sample collection and RNA isolation

Seedlings of *O. gratissimum* were collected from Nanmangalam, Medavakkam, Chennai, Tamil Nadu. The seedlings were transplanted in pots and grown in a greenhouse. The plants were taxonomically identified by a botanist (Prof. P. Jayaraman, Plant Anatomy Research Center, Chennai, Tamil Nadu). Young leaves from one healthy and mature plant at the flowering stage were used for RNA isolation using RNAiso Plus reagent (TaKaRa, Japan). DNA contamination was removed by DNase I treatment (Qiagen, GmbH, Germany) and RNA purification by using RNeasy Min Elute Clean-up Kit (Qiagen, GmbH, Germany). RNA quality and quantity were estimated by using agarose gel electrophoresis, spectrophotometer, Qubit3.0 fluorimeter (Invitrogen, California, USA) and Bioanalyzer 2100 (Agilent Technologies, California, USA).

RNA-Seq library preparation and sequencing

Library preparation was performed by using TrueSeq mRNA v2 Kit (Illumina, USA). Briefly, mRNA purification was done from total RNA by using oligo-dT attached magnetic beads, and the fragmentation of purified mRNA was done by mechanical shearing. After fragmentation, cDNA synthesis was done in the presence of reverse transcriptase and random primers. Double-stranded cDNAs were end-repaired and adenylated at the 3′end by using end-repair mix and A-Tailing mix, respectively (Illumina, USA). Then, adapters were ligated to the fragment and amplified by PCR. Finally, the library was validated by bioanalyzer 2100 (Agilent Technologies, Santa Clara, California, USA). The RNA-Seq library was quantified and subjected to paired-end sequencing in Nexseq500 (Ilumina, USA).

De novo transcriptome assembly and clustering

The paired-end sequencing image files in the bcl format were converted to FASTQ reads using bcl2fastq tool (Illumina, USA). The quality of raw reads was examined using the bioinformatics tool FastQCv0.11.8 (Andrews, 2010). Low quality reads with a Phred quality score <30, and the adapter sequences were removed by using Sickle (https://github.com/najoshi/sickle) and Cutadapt tool version 1.15 (Martin, 2011). We used Velvet-v1.2.10 (Zerbino and Birney, 2008), SOAPdenovo2 (Luo *et al.*, 2012) and Trinity-v2.6.6 (Grabherr *et al.*, 2013) for transcriptome assembly. Trinity gave the best assembly as reported before (Zhao *et al.*, 2011), which was used to construct unique transcripts and further analysis. Non-redundancy of the assembled data was achieved using CD-HIT version 4.7 (Fu *et al.*, 2012).

Assessment of gene completeness

Gene completeness analysis was done using the bioinformatics tool TRAPID (http://bioinformatics.psb.ugent.be/webtools/trapid) to obtain the total number of full length, quasi length and partial coding unigenes. This analysis was performed by comparing the assembled transcripts against PLAZA4 green plants clade database (Van *et al.*, 2018) with an *E*-value of < IE −5.

Functional annotation and classification

For functional annotation of unigenes, assembled sequences of *O. gratissimum* were searched against the non-redundant protein database at the National Center for Biotechnology Information (NCBI) (ftp://ftp.ncbi.nlm.nih.gov/blast/db/) using stand-alone BLAST^{+} package and BLASTX algorithm with threshold *E* value < IE− 5 for the homology search. Further BLASTX result was imported to the Blast2GO software for Gene Ontology (GO) analysis and enrichment of assembled unigenes. Functional classification and pathway prediction of the transcripts was carried out using KEGG Automated Annotation Server (KAAS) with default parameters.

Phylogenetic analysis

A few transcripts from the chloroplast genome were selected and assembled with the transcripts from 13 other species representing the Lamiaceae family. **Nicotiana tabacum** L. (Solanaceae) and **Arabidopsis thaliana** L. (Brassicaceae) were used as outgroups. Alignment was carried out using Clustal W (Thompson *et al.*, 2003) and a neighbour-joining tree with 1000 bootstrap replicates was constructed using MEGAX (Kumar *et al.*, 2018).

Transcript quantification

Quantification of the assembled transcripts and determination of isoform abundance were carried out by using RNA-Seq by expectation maximization (RSEM) tool (Li and Dewey 2011), and the transcripts per kilobase million (TPM) and fragments per kilobase million (FPKM) were calculated.

Prediction of transcription factors

The transcripts, which code for the transcription factors (TFs) belonging to different families, were predicted using the plant TF database (http://planttfdb.cbi.pku.edu.cn/prediction.php).

Identification of simple sequence repeats

Simple sequence repeats (SSRs) or the microsatellites among the assembled transcripts of *O. gratissimum* were identified using the MicroSAtellite (MISA) tool and misa.pl script (Beier *et al.*, 2017). The search criteria were adjusted for the identification of

https://doi.org/10.1017/S1479262121000563 Published online by Cambridge University Press

perfect di, tri, tetra, penta and hexa nucleotide motifs with minimum 6, 5, 5, 5 and 5 repeats, respectively. Primers were designed for 30 SSR loci using Primer3 software (https://primer3plus.com/cgi-bin/dev/primer3plus.cgi).

Results

RNA sequencing and de novo transcriptome assembly

The total RNA extracted from the leaves of *O. gratissimum* and purified after DNase treatment showed intact ribosomal RNA bands without any DNA contamination in agarose gel electrophoresis, A260/A280 ratio of 2.05 and RNA integrity number (RIN) of 7.30, which indicated its suitability for RNA-Seq library preparation. Sequencing of the RNA-Seq library generated 134.258 million raw reads (10.124 Gb). Since the reference genome sequence was not available for *O. gratissimum*, we performed *de novo* transcriptome assembly. After removing the adapter sequence and filtering out low-quality reads, 99.22 million reads (6.67 Gb bases) with a mean Phred score of 35.08 were obtained. These high-quality reads were assembled into 143,744 transcripts. Clustering removed 22,093 redundant transcripts and yielded 121,651 unique transcripts with an average length of 505 bases and an N50 length of 580 bases. Detail of the paired-end sequencing data generated and the assembly statistics are provided in Table 1. Assessment of gene completeness revealed that the unique transcripts included 4742 full-length, 25,644 quasi full-length and 42,415 partial coding unigenes. About 40% of the transcripts did not show significant similarity with any transcripts in the PLAZA 4.0 green plants clade database.

Table 1. Summary of paired-end sequencing and *de novo* assembly of the leaf transcriptome of *O. gratissimum*

Particulars	Number
Number of raw reads	134,258,184
Number of clean reads	99,220,462
Number of bases after processing	6,668,393,188
Mean Phred score	35.08
Percentage of assembled reads	65.57
Total number of assembled transcripts	121,651
Average length (bases)	505
Median contig length (bases)	367
Max length (bases)	26,362
Min length (bases)	201
Contig N50 (bases)	580
GC content (%)	46

Annotation of the transcripts

The transcripts that showed similarity with already reported genes were annotated for their putative biological functions based on a similarity search against the non-redundant database at NCBI. Out of the 121,651 unique assembled transcripts, 75,509 (62%) and 11,106 (9.1%) transcripts matched with existing gene models and uncharacterized proteins, respectively. The remaining 35,036 transcripts (28.8%) showed no significant similarity with any sequences in the database. These sequences may represent the sequences, which are unique to *O. gratissimum*. When the assembled transcripts were compared with the plant non-redundant protein databases from NCBI, as shown in Fig. 1, a large number of *O. gratissimum* transcripts matched with *Sesamum indicum* (40%) and *Erythranthe guttata* (14.9%).

Phylogenetic analysis

The neighbour-joining tree depicting the phylogenetic relationships among *O. gratissimum* and 13 other species from the Lamiaceae family is shown in online Supplementary Fig. S1.

Functional classification of the annotated transcripts

The assembled transcripts were functionally classified to be involved in the biological process, cellular component and metabolic function, and then assigned into several subcategories within each major component. The detailed functional classification of the transcripts under major categories and subcategories therein are shown in online Supplementary Fig. S2. The biological process category with 88,082 transcripts contained 2988 subcategories, and the oxidation-reduction process had the highest number of transcripts (3549 transcripts), followed by protein phosphorylation (2787 transcripts), regulation of transcription (2081 transcripts) and others. In the case of the cellular component category, 60,069 transcripts were grouped into 704 subcategories, with the highest number of transcripts being related to the integral component of the membrane (14,576 transcripts). Under the metabolic function category with 89,514 transcripts, 2154 subcategories were present in which the highest number of transcripts were related to ATP binding (8375 transcripts) followed by metal ion binding (3407 transcripts), RNA binding (2177 transcripts) and others. The expression levels of the *de novo* assembled transcripts from *O. gratissimum* leaves were estimated based on FPKM and TPM values. Among the top ten most abundant transcripts, three were coding for metallothionein proteins (online Supplementary Table S1).

Transcripts involved in biochemical pathways

We annotated 32,697 transcripts of *O. gratissimum* with enzyme commission (EC) numbers. These transcripts were mapped onto 149 pathways, which predominantly included the biochemical pathways of metabolism. Among the transcripts involved in metabolism, a large number of transcripts were mapped to purine and thiamine metabolism, which accounted for more than 50% of the transcripts (online Supplementary Fig. S3). Carbohydrate metabolism, nucleotide metabolism and metabolism of cofactors and vitamins accounted for more than 50% of the transcripts that were mapped under metabolism. It also included 1410 transcripts that were mapped onto 23 pathways for other secondary metabolites (online Supplementary Fig. S4). In *O. gratissimum*, among the transcripts which are involved in the biosynthesis of secondary metabolites, the highest number of transcripts was involved in phenylpropanoid biosynthesis, followed by caffeine metabolism (Fig. 2). The phenylpropanoid biosynthetic pathway begins with the formation of cinnamic acid from phenylalanine catalysed by phenylalanine ammonia-lyase (PAL). Cinnamic acid is then converted into cinnamoyl-CoA, p-coumaryl-CoA, p-coumaryl quinic acid, caffeoylquinic acid, caffeoyl-CoA, feruloyl-CoA and sinapoyl-CoA, which lead to the synthesis

https://doi.org/10.1017/S1479262121000563 Published online by Cambridge University Press

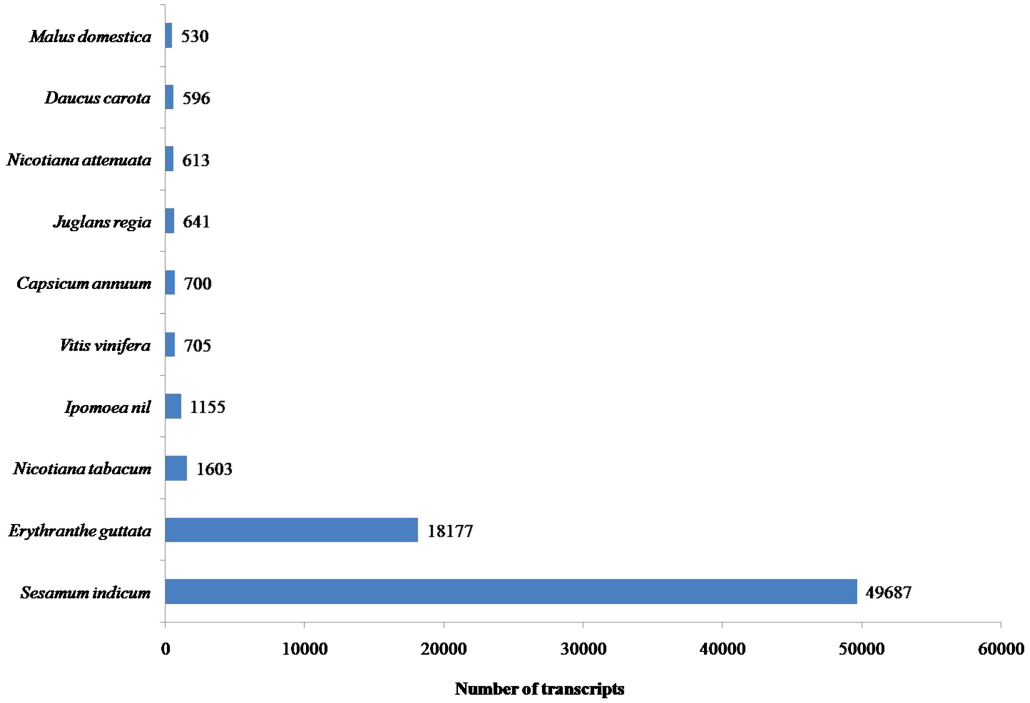

Fig. 1. Top hit species distribution of the transcripts of the *O. gratissimum* leaf transcriptome based on BLASTX search against plant non-redundant database.

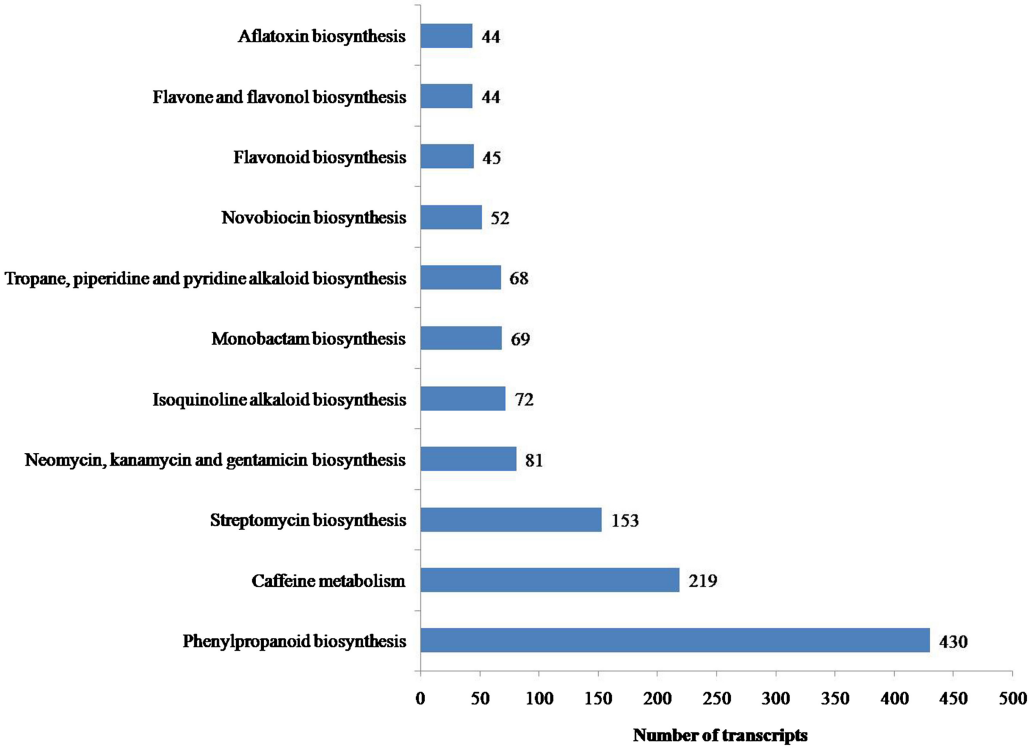

Fig. 2. The number of transcripts of the *O. gratissimum* leaf transcriptome, which code for the enzymes of different pathways of secondary metabolites.

of flavonoids, lignins, eugenol and other phenolic compounds. In the present study, KEGG analysis of leaf transcriptome of *O. gratissimum* revealed the presence of 430 transcripts coding for 14 enzymes involved in the biosynthesis of different compounds of the phenylpropanoid pathway (Fig. 3). Caffeine biosynthesis begins with the conversion of xanthosine to 7-methylxanthosine by *xanthosine* methyltransferase. Subsequently, 7-methylxanthosine is converted to 7-methylxanthine, paraxanthine and caffeine by the sequential actions of N-methyl nucleosidase and caffeine synthase. In the present study, the leaf transcriptome of

https://doi.org/10.1017/S1479262121000563 Published online by Cambridge University Press

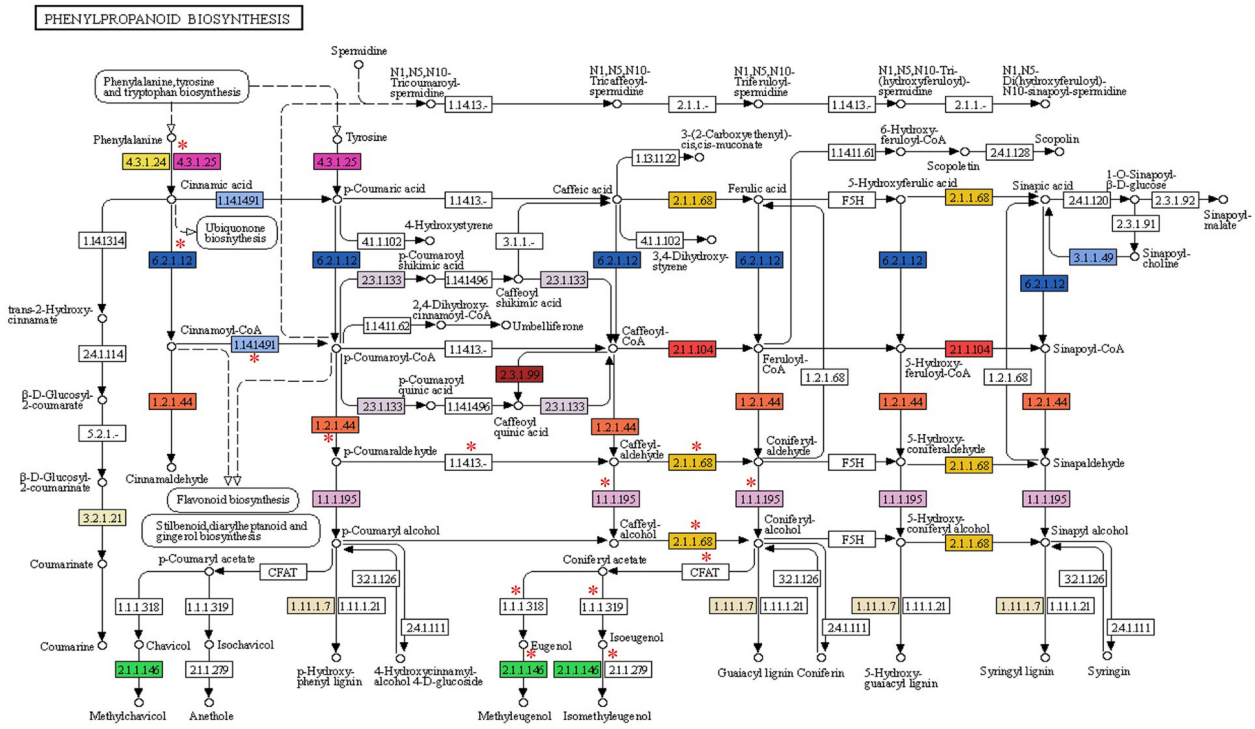

Fig. 3. Phenylpropanoid biosynthesis pathway showing different enzymes for which transcripts were identified from the *O. gratissimum* leaf transcriptome (each colour represents one Enzyme Code). The KEGG pathway map was adapted from http://www.kegg.jp/kegg/kegg1.html. Asterisk (*) indicate enzymes involved in the biosynthesis of eugenol.

Table 2. Number of different kinds of simple sequence repeats (SSRs) identified from the leaf transcriptome of *O. gratissimum*

	Number of repeats								
Repeat/Motif length	5	6	7	8	9	10	>10	Total	%
Di-repeat	0	1435	871	665	451	332	615	4369	67.13
Tri-repeat	1179	475	190	95	15	17	19	1990	30.58
Tetra-repeat	61	19	6	1	0	0	0	87	1.34
Penta-repeat	30	3	0	0	1	0	0	34	0.52
Hexa-repeat	12	14	2	0	0	0	0	28	0.43
Total	1282	1946	1069	761	467	349	634	6508	–
%	19.70	29.90	16.43	11.69	7.18	5.36	9.74	–	100

O. gratissimum contained 219 transcripts coding for four enzymes, which are involved in caffeine metabolism (online Supplementary Fig. S5).

Transcripts coding for TFs and SSRs

TFs are classified into different families according to the features of DNA binding domains. In this study, 930 transcripts coding for TFs belonging to 59 families were identified. The most abundant TFs belonged to the myeloblastosis (MYB) MYB TF family. The details of the transcripts coding for different TFs are presented in online Supplementary Fig. S6. We identified 6508 SSRs in which about 98% of them were di-nucleotide (67.1%) and tri-nucleotide repeats (30.6%). The details of the repeat number under different repeat/motif length and their frequency are given in Table 2. Among the di-nucleotide repeats, the AG/CT repeats were found to be the most abundant (67.4%) followed by AC/GT (17.8%), AT/AT (14.4%) and CG/CG (0.2%). In the case of the tri-nucleotide repeats, CGG/CCG was found to be most abundant. Primers designed for the amplification of 30 SSR loci with tri-nucleotide repeats are given in online Supplementary Table S2. The details of different types of repeats length and their frequency are given in Fig. 4.

Discussion

Genome sequences are not available for many of the medicinal plants bestowed with therapeutic compounds that are useful in traditional as well as modern medicine. Although the cost of DNA sequencing has come down drastically, *de novo* assembly

https://doi.org/10.1017/S1479262121000563 Published online by Cambridge University Press

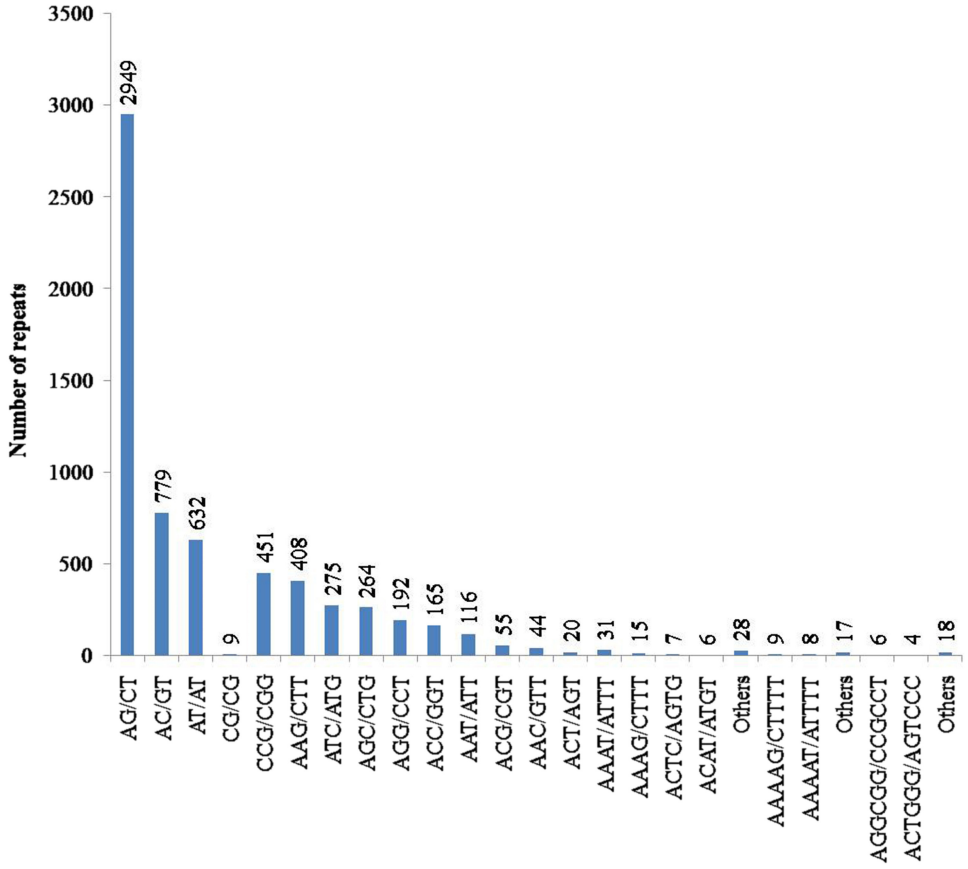

Fig. 4. Number of different types of simple sequence repeats (SSRs) identified from the *O. gratissimum* leaf transcriptome.

of chromosome-level genomes in the absence of physical maps is a challenge. Moreover, the accumulation of therapeutic compounds in medicinal plants depends on expression than the mere presence of unique genes. Therefore, the characterization of expressed genes from whole transcriptome analysis will be highly useful to understand the functional genes involved in the biosynthesis of therapeutic compounds. Clove basil (*O. gratissimum*) or Ram Tulasi is a commonly used medicinal plant, and its leaves are used as a single drug and as a component in herbal formulations. In this study, we report the leaf transcriptome of *O. gratissimum* assembled from Illumina's paired-end RNA-Seq data. Most of the assembled transcripts from *O. gratissimum* showed the highest nucleotide identity with *S. indicum* and *E. guttata*. This result is in concordance with the phylogenetic relationships among these species as per the latest Angiosperm Phylogeny Group (APG) IV classification (Chase *et al.*, 2016). While *O. gratissimum* belongs to Lamiaceae, *S. indicum* and *E. guttata* belong to Pedaliaceae and Phrymaceae, respectively. However, all of them are taxonomically related as they belong to the same order, Lamiales.

The number of genes involved in the synthesis of secondary metabolites is variable among different species. Even at the variety level, a significant difference in the number of such transcripts was observed between two varieties of *O. basilicum* (Torre *et al.*, 2016). In this study, we identified 1410 transcripts from the leaf transcriptome of *O. gratissimum* that are functional in 23 pathways of secondary metabolite biosynthesis. In a similar study, although relatively a comparable amount of RNA-Seq data were analysed, the number of transcripts related to secondary metabolite biosynthesis identified was only 501 and 952 from *O. tenuiflorum* (syn. *O. sanctum*) and *O. basilicum*, respectively (Rastogi *et al.*, 2014). Among the *O. gratissimum* transcripts mapped to the secondary metabolite biosynthesis in this study, those coding for the enzymes of phenylpropanoid biosynthesis and caffeine metabolism were abundant than the others. We also found a significant number of the MYB TFs, which are the key regulator of the phenylpropanoid metabolism in plants (Ma and Constabel, 2019). MYB TFs are also involved in the regulation of tolerance to biotic and abiotic stress tolerance (Lippold *et al.*, 2009; Du *et al.*, 2018).

Phenylpropanoids are a diverse group of compounds derived from phenylalanine, and are involved in plant defence, structural support and tolerance against biotic and abiotic stresses (Vogt, 2010). Eugenol is a phenylpropanoid compound with several therapeutic properties (Fujisawa and Murakami, 2016; Barboza *et al.*, 2018), and it is the primary ingredient in the essential oil of *Ocimum* species. We identified 101 transcripts, which code for seven enzymes that are involved in eugenol biosynthesis in *O. gratissimum*. Among the 101 transcripts, 64 were unique transcripts and 37 were truncated forms of the unique transcripts. The number of unique transcripts is reasonable considering that six out of the seven enzymes of the eugenol biosynthesis pathway are encoded by multigene families. Only eugenol O-methyltransferase is encoded by a single gene family. Interestingly, this gene is represented by a single transcript in this study also. Regarding caffeine metabolism, the leaf transcriptome of *O. gratissimum* contained 219 transcripts coding for four enzymes involved in caffeine metabolism. These enzymes are

https://doi.org/10.1017/S1479262121000563 Published online by Cambridge University Press

involved in either caffeine catabolism or diverting the precursors of caffeine to the synthesis of other metabolites. For example, urate hydroxylase is involved in the catabolism of caffeine to allantoin. Arylamine N-acetyltransferase converts paraxanthine, a precursor of caffeine, to 5-acetylamino-6-formylamino-3-methyl uracil. None of the transcripts related to caffeine metabolism identified from *O. gratissimum* is involved in caffeine synthesis, which supports the caffeine-free nature of the Ocimum species (Pattanayak *et al.*, 2010).

Some heavy metals are essential micronutrients (Co, Fe, Mn, Mo, Ni, Zn, Cu), but others are nonessential and toxic (Pb, Cd, As, Cr, Hg). *Ocimum* species were reported to accumulate some of the toxic heavy metals. While *O. basilicum* was reported to be a hyperaccumulator of Cd, Cr and Pb (Chand *et al.*, 2015; Dinu *et al.*, 2020), *O. gratissimum* was found to accumulate Cd (Chaiyarat and Suebsima, 2011). Though the accumulation of toxic heavy metals is helpful in bioremediation, it is a matter of concern when the accumulating plants are used for medicinal purposes. Metallothioneins (MTs) are cysteine-rich proteins, which bind with heavy metals and help in their accumulation in plants (Hamer, 1986). Our quantitative analysis showed that the transcripts coding for metallothionein proteins are abundant in the leaf transcriptome of *O. gratissimum*. This is the first report on the abundant expression of MT genes in *Ocimum*, which shall be useful for a detailed study on heavy metal accumulation in this important group of medicinal plants.

We also identified transcripts belonging to 59 TF families and 6508 SSR markers from the leaf transcriptome of *O. gratissimum*. The TFs are diverse and actively regulate all the vital functions of the plants, including germination, growth and reproduction. They are also involved in the synthesis of bioactive compounds, especially in the regulation of secondary metabolism. Therefore, identification of the genes coding for the TF is useful for understanding the regulatory mechanism of secondary metabolites. Though SSR markers are available for *O. tenuiflorum* and *O. basilicum* (Rastogi *et al.*, 2014, 2015), this is the first time a large number of SSR markers were identified for *O. gratissimum*. These SSRs shall be useful as DNA markers for comparative genomics, molecular breeding, genetic diversity assessment and gene mapping (Pyne *et al.*, 2018). The leaf transcriptome reported here will serve as a foundation for further molecular studies in *O. gratissimum* and related species. We also designed primers for 30 SSR loci to be useful for this purpose.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S1479262121000563

Acknowledgements. This study was financially supported by SRM-DBT Partnership Platform for Contemporary Research Services and Skill Development in Advanced Life Sciences Technologies (Order No. BT/PR12987/INF/22/205/2015).

Author contributions. MP and PN contributed to the study conception and design. Material preparation, data collection and analysis were performed by Tanuja, NRP and RK. The manuscript was written by MP and Tanuja. All authors read and approved the final manuscript.

Availability of data. The data were submitted to NCBI with Biosample Accession Number SAMN15582955.

References

Airaodion AI, Ibrahim AH, Ogbuagu U, Ogbuagu EO, Awosanya OO, Akinmolayan JD, Njoku OC, Obajimi OO, Adeniji AR and Adekale OA (2019) Evaluation of phytochemical content and antioxidant potential of *Ocimum gratissimum* and *Telfairia occidentalis* leaves. *Asian Journal of Research in Medical and Pharmaceutical Sciences* **7**, 1–11.

Andrews S (2010) FASTQC. A quality control tool for high throughput sequence data'.

Barboza JN, da Silva Maia Bezerra Filho C, Silva RO, Medeiros JV and de Sousa DP (2018) An overview on the anti-inflammatory potential and antioxidant profile of eugenol. *Oxidative Medicine and Cellular Longevity* **2018**, 3957262.

Beier S, Thiel T, Münch T, Scholz U and Mascher M (2017) MISA-web: a web server for microsatellite prediction. *Bioinformatics (Oxford, England)* **33**, 2583–2585.

Casanova LM, Gu W, Costa SS and Jeppesen PB (2017) Phenolic substances from Ocimum species enhance glucose-stimulated insulin secretion and modulate the expression of key insulin regulatory genes in mice pancreatic islets. *Journal of Natural Products* **80**, 3267–3275.

Chaiyarat R, Suebsima R, Putwattana N, Kruatrachue M and Pokethitiyook P (2011) Effects of soil amendments on growth and metal uptake by *Ocimum gratissimum* grown in Cd / Zn-contaminated soil. *Water, Air, & Soil Pollution* **214.1**, 383–392.

Chand S, Singh S, Singh VK and Patra DD (2015) Utilization of heavy metal-rich tannery sludge for sweet basil (*Ocimum basilicum* L.) cultivation. *Environmental Science and Pollution Research* **22.10**, 7470–7475.

Chase MW, Christenhusz MJM, Fay MF, Byng JW, Judd WS, Soltis DE, Mabberley DJ, Sennikov AN, Soltis PS, Stevens PF, Briggs Barbara, Brockington Samuel, Chautems Alain, Clark John C, Conran John, Haston Elspeth, Möller Michael, Moore Michael, Olmstead Richard, Perret Mathieu, Skog Laurence, Smith James, Tank David, Vorontsova Maria and Weber Anton (2016) An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: aPG IV. *Botanical Journal of the Linnean Society* **181**, 1–20.

Chattopadhyay RR (1994) Anxiolytic activity of *Ocimum sanctum* leaf extract. *Ancient Science of Life* **14**, 108–111.

Chimnoi N, Reuk-ngam N, Chuysinuan P, Khlaychan P, Khunnawutmanotham N, Chokchaichamnankit D, Thamniyom W, Klayraung S, Mahidol C and Techasakul S (2018) Characterization of essential oil from *Ocimum gratissimum* leaves: antibacterial and mode of action against selected gastroenteritis pathogens. *Microbial Pathogenesis* **118**, 290–300.

Dinu C, Vasile GG, Buleandra M, Popa DE, Gheorghe S and Ungureanu EM (2020) Translocation and accumulation of heavy metals in *Ocimum basilicum* L. Plants grown in a mining-contaminated soil. *Journal of Soils and Sediments* **20.4**, 2141–2154.

Du YT, Zhao MJ, Wang CT, Gao Y, Wang YX, Liu YW, Chen M, Chen J, Zhou YB, Xu ZS and Ma YZ (2018) Identification and characterization of GmMYB118 responses to drought and salt stress. *BMC Plant Biology* **18**, 1–18.

Fujisawa S and Murakami Y (2016) Eugenol and its role in chronic diseases. *Advances in Experimental Medicine and Biology* **929**, 45–66.

Fu L, Niu B, Zhu Z, Wu S and Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)* **28**, 3150–3152.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Bruce W, Friedman N and R A (2013) Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology* **29**, 644–652.

Hamer DH (1986) Metallothionein. *Annual Review of Biochemistry* **55**, 913–951.

Ighodaro OM, Agunbiade SO and Akintobi O (2010) Phytotoxic and anti-microbial activities of flavonoids in *Ocimum gratissimum*. *Life Science Journal* **7**, 45–48.

Kumar S, Stecher G, Li M, Knyaz C and Tamura K (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular Biology Evolution* **35**, 1547–1549.

Lahon K and Das S (2011) Hepatoprotective activity of *Ocimum sanctum* alcoholic leaf extract against paracetamol-induced liver damage in albino rats. *Pharmacognosy Research* **3**, 13–18.

Li B and Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323.

Lippold F, Sanchez DH, Musialak M, Schlereth A, Scheible WR, Hincha DK and Udvardi MK (2009) AtMyb41 regulates transcriptional and metabolic responses to osmotic stress in Arabidopsis. *Plant Physiology* **149.4**, 1761–1762.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y and Tang J (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 2047–217X.

Ma D and Constabel CP (2019) MYB repressors as regulators of phenylpropanoid metabolism in plants. *Trends in Plant Science* **24**, 275–289.

Manaharan T, Thirugnanasampandan R, Jayakumar R, Ramya G, Ramnath G and Kanthimathi MS (2014) Antimetastatic and anti-inflammatory potentials of essential oil from edible *Ocimum sanctum* leaves. *Scientific World Journal* **239508**, 5 pages.

Martin M (2011) CUTADAPT removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal* **17.1**, 10–12.

Olamilosoye KP, Akomolafe RO, Akinsomisoye OS, Adefisayo MA and Alabi QK (2018) The aqueous extract of *Ocimum gratissimum* leaves ameliorates acetic acid-induced colitis via improving antioxidant status and hematological parameters in male Wistar rats. *Egyptian Journal of Basic and Applied Sciences* **5**, 220–227.

Parasuraman S, Balamurugan S, Christapher PV, Petchi RR, Yeng WY, Sujithra J and Vijaya C (2015) Evaluation of antidiabetic and antihyperlipidemic effects of hydroalcoholic extract of leaves of *Ocimum tenuiflorum* (Lamiaceae) and prediction of biological activity of its phytoconstituents. *Pharmacognosy Research* **7**, 156–165.

Pattanayak P, Behera P, Das D and Panda S (2010) *Ocimum sanctum* Linn. A reservoir plant for therapeutic applications: an overview. *Pharmacognosy Reviews* **4**, 95–105.

Pyne RM, Honig JA, Vaiciunas J, Wyenandt CA and Simon JE (2018) Population structure, genetic diversity and downy mildew resistance among Ocimum species germplasm. *BMC Plant Biology* **18**, 1–15.

Rastogi S, Meena S, Bhattacharya A, Ghosh S, Shukla RK, Sangwan NS, Lal RK, Gupta MM, Lavania UC, Gupta V and Nagegowda DA (2014) De novo sequencing and comparative analysis of holy and sweet basil transcriptomes. *BMC Genomics* **15**, 1–18.

Rastogi S, Kalra A, Gupta V, Khan F, Lal RK, Tripathi AK, Parameswaran S, Gopalakrishnan C, Ramaswamy G and Shasany AK (2015) Unravelling the genome of holy basil: an "incomparable" "elixir of life" of traditional Indian medicine. *BMC Genomics* **16**, 413.

Saharkhiz MJ, Kamyab AA, Kazerani NK, Zomorodian K, Pakshir K and Rahimi MJ (2014) Chemical compositions and antimicrobial activities of *Ocimum sanctum* L. Essential oils at different harvest stages. *Jundishapur Journal of Microbiology* **8**, 1–7.

Singh P, Jayaramaiah RH, Agawane SB, Vannuruswamy G, Korwar AM, Anand A, Dhaygude VS, Shaikh ML, Joshi RS, Boppana R and Kulkarni MJ (2016) Potential dual role of eugenol in inhibiting advanced glycation end products in diabetes: proteomic and mechanistic insights. *Scientific Reports* **6**, 1–13.

Srivastava S, Adholeya A, Conlan XA and Cahill DM (2016) Acidic potassium permanganate chemiluminescence for the determination of antioxidant potential in three cultivars of *Ocimum basilicum*. *Plant Foods for Human Nutrition* **71**, 72–80.

Sumitha K V and Thoppil JE (2016) Larvicidal efficacy and chemical constituents of O. gratissimum L. (Lamiaceae) essential oil against aedes albopictus skuse (Diptera: Culicidae). *Parasitology Research* **115**, 673–680.

Thompson JD, Gibson TJ and Higgins DG (2003) Multiple sequence alignment using ClustalW and ClustalX. Current protocols in bioinformatics. Chapter 2: Unit 2.3.

Torre S, Tattini M, Brunetti C, Guidi L, Gori A, Marzano C, Landi M and Sebastiani F (2016) De novo assembly and comparative transcriptome analyses of red and green morphs of sweet basil grown in full sunlight. *PLoS ONE* **11**, 1–19.

Upadhyay AK, Chacko AR, Gandhimathi A, Ghosh P, Harini K, Joseph AP, Joshi AG, Karpe SD, Kaushik S, Kuravadi N and Lingu CS (2015) Genome sequencing of herb tulsi (*Ocimum tenuiflorum*) unravels key genes behind its strong medicinal properties. *BMC Plant Biology* **15**, 1–20.

Van Bel M, Diels T, Vancaester E, Kreft L, Botzki A, Van de Peer Y, Coppens F and Vandepoele K (2018) PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Research* **46.D1**, 1190–1196.

Vogt T (2010) Phenylpropanoid biosynthesis. *Molecular Plant* **3**, 2–20.

Zerbino DR and Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18**, 821–829.

Zhan X, Yang L, Wang D, Zhu JK and Lang Z (2016) De novo assembly and analysis of the transcriptome of *Ocimum americanum* var. *pilosum* under cold stress. *BMC Genomics* **17**, 1–12.

Zhao QY, Wang Y, Kong YM, Luo D, Li X and Hao P (2011) Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics* **12**, 1–12.