**Supplementary Materials**

Tissue-specific transcriptome analysis reveals candidate genes for terpenoid and phenylpropanoid metabolism in the medicinal plant *Ferula assafoetida*

Hajar Amini,\*,† Mohammad Reza Naghavi,†,[[1]](#footnote-1) Tong Shen,‡ Yanhong Wang,§ Jaber Nasiri,† Ikhlas A. Khan,§ Oliver Fiehn,‡,\*\* Philipp Zerbe,\* and Julin N. Maloof\*,1

\*Department of Plant Biology, University of California, Davis, CA, USA, 95616

†Department of Agronomy and Plant Breeding, Agricultural and Natural Resources College, University of Tehran, Iran, 77871-31587

‡West Coast Metabolomics Center, University of California, Davis, CA, USA, 95616

§National Center for Natural Products Research, Research Institute of Pharmaceutical Sciences, University of Mississippi, Oxford, MS, USA, 38677

\*\*Department of Biochemistry, Faculty of Sciences, King Abdulaziz University, Jeddah, Saudi-Arabia

MATERIAL AND METHODS

**Metabolite analysis**

GC-MS analyses were performed on an Agilent GC model 5977A equipped with a MD800 selective mass detector. Analyses were performed with electron ionization (EI) at 70 eV, an ion-source temperature of 200°C and an interface temperature of 280°C using split-splitless injection (split ratio 1:10) with a 280°C injector temperature and a1 μl injection volume. Separation was achieved using a fused silica column 5% phenyl-poly-dimethyl-siloxane **(**Chrompack CP-Sil 8 CB 50 m x 250 µm x 0.12 µm**)** and the following GC parameters: 50°C for 2 min, 4°C/min to 120°C, 2°C/min to 200 °C, 25°C/min to 280°C with hold for 8 min. Data acquisition was performed with Mass Lab software for the mass ranges 30-600 u with a scan speed of 1 scan/sec**.** Identification of compounds was based on Kovats retention indices as calculated using n-alkanes C11-C28 as reference compounds using three replicated measurements, as well as comparison to reference mass spectral libraries, including Adams (Robert Adams, 2012), US National Institute of Standards and Technology (NIST, USA), WILEY 1996 Ed. mass spectra library, and an in-house library of 600 mass spectra.

For ultra-high-performance liquid chromatography-quadrupole time-of-flight mass spectrometry (UHPLC-QToF-MS) studies, dry plant material was ground to a fine powder. Plant samples (0.5 g) were weighed in a centrifuge tube and 3.5 mL of methanol added before sonication in a water bath for 30 min and centrifugation at 959 *×g* for 15 min. The supernatant was transferred to a 10 mL flask. The extraction procedure was repeated twice and supernatants combined. The final volume was adjusted to 10 mL with methanol and mixed thoroughly. Prior to injection, an adequate sample volume was passed through a 0.45 µm PTFE filter and collected in a LC sample vial. All samples were analyzed on a Waters Acquity UPLC™ system including binary solvent manager, sampler manager, column compartment and a photodiode array (PDA) detector (Waters Acquity model code UPD). Data were acquired and interpreted using MassLynxTM NT 4.1 software. The separation was carried out on a Waters Acquity UPLC™ BEH Shield RP18 column (100 mm × 2.1mm i.d., 1.7 µm) which was equipped with a guard column (Vanguard 2.1 × 5 mm). The sample temperature and column temperature were maintained at 15°C and 35°C, respectively. The mobile phase consisted of water containing 0.05% formic acid (*v*/*v*) (A) and acetonitrile with 0.05% formic acid (B). The analysis was performed using the following gradient elution at a flow rate of 0.30 mL/min: 0-11 min, 30% B to 70% B and 11-15 min, 70% B to 100% B. The column was washed with 100% B for 3 min and re-equilibrated for 3.5 min between runs using the initial conditions. The strong needle wash (90/10; acetonitrile/water, *v*/*v*) and weak needle wash solution (10/90; acetonitrile/water, *v*/*v*) were used. The total run time for analysis was 15 min. Mass spectrometric analyses were performed using electrospray ionization (ESI) on a Waters ACQUITY™ Xevo QToF G2-S mass spectrometer (Waters Corporation, Manchester, UK). The ESI source was operated under the positive ionization mode in the following conditions: capillary of 0.65 kV, cone of 25 V, source temperature of 80°C, desolvation temperature 500°C, desolvation gas flow of 900 L/h, cone gas flow of 50 L/h, and collision energy of 6 eV. Leucine-enkephalin was used for the lock mass at a concentration of 5 ng/mL and flow rate of 5 µL/min. Ions [M+H]+ (*m/z* 556.2771 Da) and fragment ion (*m/z* 278.1141 Da) of leucine-enkephalin were employed to ensure mass accuracy during the MS analysis. The lock spray interval was set at 30 s, and the data were averaged over three scans. The mass spectrometer was programmed to step between low (6 eV) and elevated (10-30 eV) collision energies on the gas cell, using a scan time of 0.5 s per function over a mass range of 50–1200 *m/z*. When data were acquired with MSE, two interleaved scan functions were used. The first scan function acquired a wide mass range using low collision energy. This scan function collected precursor ion information for the sample. The second scan function acquired data over the same mass range; however, the collision energy was ramped from low to high. This scan function allowed for the collection of a full-scan accurate mass of fragments along with precursor ion information. An individual stock solution of standard compounds (umbelliferone, luteolin, and umbelliprenin) was prepared at a concentration of 1.0 mg/mL in methanol. The calibration curves were prepared at seven different concentration levels and listed in Supplementary Table S1.

***De novo* transcriptome assembly and evaluation**

We utilized four different *de novo* transcriptome assembly pipelines to ensure a high quality reference transcriptome was assembled.

First pipeline: Reads from all organs were combined and assembledwith Trinity v2.4.0 with kmer = 25 (Grabherr *et al.* 2011). We refer to this pipeline as “Trinity” pipeline.

Second pipeline: The Khmer v2.0 tools (Crusoe *et al.* 2015) was applied to variable kmer coverage abundance trimming to the reads prior to Trinity assembly. This reduces the computational cost of assembly without negatively affecting the quality of the assembly. We refer to this pipeline as the “Khmer\_Trinity” pipeline.

Third pipeline: The Drap v1.91 was applied as a post processing step after using Trinity to compact and correct the assembled transcriptome (Cabau *et al*. 2017). We refer to this pipeline as the “Trinity\_DRAP” pipeline.

Fourth pipeline: The Drap v1.91 was applied as a post processing step after using Oases v0.2.06 with kmer 25, 31,37, 43, 49 (Zerbino and Birney 2008; Schulz *et al.* 2012). We refer to this pipeline as the “Oases\_DRAP” pipeline.

### Tissue-specific differential expression analysis

The differentially expressed genes of the assembled transcriptome were investigated using the edgeR package in the R statistical environment (FDR <0.05) (Robinson *et al.* 2010b; R Core Team 2016) as follows: The differential expression analysis was calculated from 12 samples from the UC Davis facility (3 samples from 4 different organs) to avoid the possibility of a batch effect altering the expression analysis.We filtered to remove genes with low expression, only retaining genes with more than 10 reads in more than three samples. After this filtering, sample libraries were normalized by calculating the effective library size and normalization factor using the TMM method on counts data (Robinson and Oshlack, 2010a). The model was calculated based on “model.matrix (~organs, data=data)” formula. Organs refer to four organs including roots, flowers, leaves and stems for designing this model. We then identified genes differentially expressed among plant organs using a generalized linear model (glm) in edgeR and multiple-testing correction via the Benjamini and Hochberg (BH) procedure (Benjamini and Hochberg, 1995).

RESULTS AND DISCUSSION

***De novo* transcriptome assembly and evaluation**

By using the Illumina HiSeq platform, the libraries were sequenced, and then after removing low quality reads and adaptor sequences, approximately 218 million reads from roots, 190 million reads from flowers, 123 million reads from stems, and 91 million reads from leaves were generated (Supplemental Table S2). To obtain high quality assemblies, we compared results from four different *de novo* assembly pipelines including Trinity (Grabherr *et al.* 2011), Khmer\_Trinity (Crusoe *et al.* 2015), Trinity\_DRAP, and Oases\_DRAP (Cabau *et al.* 2017) (Supplemental Table S3). The Oases assemblies after DRAP correction produced the longest contigs (N50) in comparison to other pipelines (Supplemental Table S3).

To investigate the quality of the transcriptome assembly we determined the percentage of reads mapping and the completeness of universal single copy orthologs. Using the STAR aligner, we found that over 65% of the reads mapped uniquely to the transcripts produced by Oases\_DRAP (Figure S2). High quality assemblies should have a high representation of universally conserved single copy orthologs; we assessed this metric using BUSCO v3 (Simão *et al.* 2015). The universal single-copy orthologs used were from the OrthoDB database (Zdobnov *et al.* 2016). The transcriptome assembly produced by Oases and further corrected by DRAP contained complete assemblies of over 90% of these genes and more than 80% were present as single copy (Table S4), confirming that this assembled transcriptome was of high quality.

REFERENCES

Adams, R. P., 2012 *Identification of Essential Oils by Ion trap Mass Spectroscopy.* Academic Press Available at: <https://www.elsevier.com/books/identification-of-essential-oils-by-ion-trap-mass-spectroscopy/adams/978-0-12-044230-0>.

Benjamini, Y., and Y. Hochberg, 1995 Controlling the false discovery rate - a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B. Stat. Methodol. 57: 289–300. Available at: ://A1995QE45300017.

Cabau, C., F. Escudié, A. Djari, Y. Guiguen, J. Bobe *et al.,* 2017 Compacting and correcting Trinity and Oases RNA-Seq de novo assemblies. PeerJ 5: doi:10.7717/peerj.2988.

Crusoe, M. R., H. F. Alameldin, S. Awad, E. Boucher, A. Caldwell *et al*., 2015 The khmer software package: enabling efficient nucleotide sequence analysis. F1000Res4. doi:10.12688/f1000research.6924.1.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson *et al.*, 2011 Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotech*.* 29: 644–652. doi:10.1038/nbt.1883.

R Core Team, 2016 R: A language and environment for statistical computing. R Foundation for Statistical Computing. Available at: <https://www.R-project.org/>.

Robinson, M. D., and A. Oshlack, 2010a A scaling normalization method for differential expression analysis of RNA-seq data. Genome biol. 11: R25. doi:10.1186/gb-2010-11-3-r25.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth, 2010b edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139–140. doi:10.1093/bioinformatics/btp616.

Schulz, M. H., D. R. Zerbino, M. Vingron, and E. Birney, 2012 Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28: 1086–1092. doi:10.1093/bioinformatics/bts094.

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, 2015 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210–3212. doi:10.1093/bioinformatics/btv351.

Zdobnov, E. M., F. Tegenfeldt, D. Kuznetsov, R. M. Waterhouse, F. A. Simão *et al.,* 2016 OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. Nucleic Acids Res. 45: D744–D749. doi:10.1093/nar/gkw1119.

Zerbino, D. R., and E. Birney, 2008 Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18: 821–829. doi:10.1101/gr.074492.107.

1. Corresponding authors: Email; [jnmaloof@ucdavis.edu](mailto:jnmaloof@ucdavis.edu); [mnaghavi@ut.ac.ir](mailto:mnaghavi@ut.ac.ir) [↑](#footnote-ref-1)