Spring 5-29-2019

De novo Sequencing and Analysis of *Salvia hispanica* Transcriptome and Identification of Genes Involved in the Biosynthesis of Secondary Metabolites

James Wimberley
Chapman University, wimbe106@mail.chapman.edu

De novo sequencing and analysis of *Salvia hispanica* transcriptome and identification of genes involved in the biosynthesis of secondary metabolites

A Thesis by

James Wimberley

Chapman University

Orange, CA

Schmid College of Science and Technology

Submitted in partial fulfillment of the requirements for the degree of

Master of Science in Computational and Data Sciences

May 2019

Committee in charge:

Hagop Atamian, Ph.D., Chair

Cyril Rakovski, Ph.D.

Gennady Verkhivker, Ph.D.

CHAPMAN UNIVERSITY
SCHMID COLLEGE OF SCIENCE AND TECHNOLOGY
Computational and Data Sciences

The thesis of James Wimberley is approved.

Hagop Atamian, Ph.D., Chair

on his behalf; Hesham El-Askary, Program Director

Cyril Rakovski, Ph.D.

Gennady Verkhivker

Gennady Verkhivker (May 25, 2019)

Gennady Verkhivker, Ph.D.

May 2019

De novo sequencing and analysis of *Salvia hispanica* transcriptome and identification of genes involved in the biosynthesis of secondary metabolites

by James Wimberley

ABSTRACT

De novo sequencing and analysis of *Salvia hispanica* transcriptome and identification of genes involved in the biosynthesis of secondary metabolites

by James Wimberley

Salvia hispanica L. (commonly known as chia) is gaining popularity worldwide and specially in US as a healthy oil and food supplement for human and animal consumption due to its favorable oil composition, and high protein, fiber, and antioxidant contents. Despite these benefits and its growing public demand, very limited gene sequence information is currently available in public databases. In this project, we generated 90 million high quality 150 bp paired-end sequences from the chia leaf and root tissues. The sequences were de novo assembled into 103,367 contigs with average length of 1,445 bp. The resulted assembly represented 92.2% transcriptome completeness. Around 69% of the assembled contigs were annotated against the uniprot database and represented a diverse array of functional and biological categories. A total of 14,267 contigs showed significant expression difference between the leaf and root tissues, with 6,151 and 8,116 contigs up-regulated in the leaf and root, respectively. The sequence data generated in this project will provide valuable resources for future functional genomic research in chia. With the availability of transcriptome sequences, it would be possible to identify genes involved in the important metabolic pathways that give chia its unique nutritional and medicinal properties. Finally, the generated data will contribute to the genetic improvement efforts of chia to better serve the public demand.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

Abbreviation	Meaning
bp	Base pair
cDNA	Complimentary DNA
GC/MS	Gas chromatography/mass spectrometry
GO	Gene Ontology
NCBI	National Center for Biotechnology Information
qPCR	Quantitative polymerase chain reaction

1 Introduction

1.1 Introduction

Salvia hispanica L. (commonly known as chia) is an annual self-pollinated species that belongs to the mint family (*Lamiaceae*) and is native to central and southern Mexico and Guatemala [1]. *S. hispanica* has a long history of plant–human interaction. In pre-Columbian Mesoamerica, the plant was a major commodity similar to bean, corn, and squash, and Aztecs valued its seeds for food, medicine, and oil [2]. The codices of 16th century Mexico provide a wealth of ethnobotanical information and indicate large areas of agricultural land were devoted exclusively to chia cultivation [2]. However after Spanish contact and colonization, the level of cultivation plummeted and the plant was largely overlooked as a food crop until its re-emergence as an alternative crop and a health food in the beginning of the 20th century [1]. *S. hispanica* grows up to three feet long and develops lush green foliage rich in essential oils before producing long purple or white flowers. These flowers develop to produce thousands of small (2 mm in length) highly nutritious edible seeds.

Chia seed provides remarkably balanced and close to complete nutritional source with 34.4% total dietary fiber, 31% total lipids, 16% protein, 5.8% moisture, and high amounts (335–860 mg/100 g) of calcium, phosphorus, potassium, and magnesium [1] [3] [4]. The oil content of chia seed (31%) is higher than that of other oilseeds of commercial

importance, such as soybean (24%) and cotton-seed (24%) [4]. The fatty acids of chia seed oil are highly unsaturated, with their main components being linolenic (50–57%) and linoleic (17–26%) fatty acids. This represents the highest known percentage of linolenic fatty acid of any plant source [5]. However, the leaf fatty acid has 60% more palmitic acid content compared to the seed but only 25% the concentration of α-Linolenic acid [6]. Besides fatty acids, chia leaves contain essential oils that have the potential for commercial uses in food flavoring and fragrance industry. The leaf oils also have antimicrobial properties and could be used as biopesticides to protect plants from pathogen and insect attacks. GC/MS analysis of the leaf oil composition from plants grown in southern California, southeastern Texas, and northwestern Argentina identified large number of components of which the most abundant were sesquiterpenes β-caryophyllene, globulol, γ-muurolene, α-humulene, germacrene-B, and widdrol and the monoterpene β-pinene [7]. Similarly, thorough analysis of chia leaf oil constituents by Elshafie et al (2018) identified 60 different sesquiterpenes, accounting for 84.5% of the oil [8]. The chia leaf oil sesquiterpenes were mostly represented by sesquiterpene hydrocarbons (53.9%) and oxygenated sesquiterpenes (30.6%). Some abundant sesquiterpene hydrocarbons include (Z)-caryophyllene (11.5%), (E)-caryophyllene (10.6%), α-humulene (4.8%), δ-amorphene (3.1%), and γ-gurjunene (3.1%). Oxygenated sesquiterpenes showed more uniform distribution with α-eudesmol (3.8%), caryophyllene oxide (2.7%), and spathulenol (2.2%) as the main representatives. Monoterpenes constitute the 0.4% of the oil. Phenolic compounds constitute 1.5% and oxygenated compounds constitute 5%. The metabolic profile of chia leaves also includes several flavonoids and hydroxycinnamic acids such as

apigenin and luteolin glycosides, aglycones quercetin methyl ether and naringenin, and quercetin and kaempoferol-based flavonoids [9].

Despite its favorable nutritional qualities and the plethora of secondary metabolites that it synthesizes with potential uses in food and fragrance industries, only 62 *S. hispanica* expressed sequence tag (EST) sequences are publicly available in the NCBI nucleotide database. RNA sequencing (RNA-Seq) is a powerful tool that enables profiling the gene constituent of non-model species. The *de novo* sequencing and assembly of the transcriptome is the first step in gaining insights into the genes and molecular pathways underlying the different phenotypes in non-model plant species.

In this study, we sequenced and assembled *S. hispanica* leaf and root transcriptome into 103,367 contigs with an estimated 92.8% completeness. Functional and Gene Ontology (GO) analysis identified diverse gene categories represented in the assembled transcriptome. Differential gene expression analysis identified 6,151 and 8,116 contigs that had higher expression in *S. hispanica* leaf and root, respectively. Genes encoding some key enzymes involved in vitamin biosynthesis pathway were identified. The sequencing data generated in this study will provide valuable resource to better understand the molecular mechanisms underlying the desirable characteristics of *S. hispanica* and will contribute to future research aimed at further improvement of these characteristics.

2 Methods

2.1 Plant Materials

Seeds of *S. hispanica* variety pinta was kindly provided by Dr. Joseph Cahill; Ventura Botanical Gardens. The seeds were germinated in Sunshine® All-Purpose potting mix and maintained in Conviron® growth chamber at 22°C with a 16-h light and 8-h dark photoperiod and 200 mol m-2 s-1 light intensity for two weeks. At four-leaf developmental stage, a pair of newly emerged leaves were harvested and immediately frozen in liquid nitrogen. Roots were washed thoroughly with tap water before harvesting. Tissues from six seedlings were combined together as one biological replicate. A total of three biological replicates were collected.

2.2 RNA extraction, library construction and Illumina sequencing

RNA was extracted from leaf and root tissues using TRIzol® (Invitrogen) according to manufacturer's instructions. The RNA was further purified using Spectrum™ Plant Total RNA Kit (Sigma-Aldrich) and subjected to on-column DNase treatment. The RNA quality and quantity were assessed using Agilent 2100 Bioanalyzer (Agilent Technologies). 500 ng total RNA was used for RNA-seq library preparation according to the protocol described by Kumar et al. (2012) [10]. Briefly, mRNA was isolated using oligo(dT) coated magnetic beads (Invitrogen) and treated with DNase followed by first and second strand cDNA

synthesis. The cDNA was fragmented using divalent cations and enriched for fragments around 300 bp. Finally, custom barcoded adaptors were ligated to the fragments followed by 10 cycles of PCR enrichment of the library products. The barcoded libraries were pooled together and subjected to 150 bp paired-end sequencing on Illumina HiSeq4000 machine [11].

2.3 Bioinformatic analysis

From the raw sequences, the adaptors and low quality bases were trimmed using Trimmomatic version 0.36 with 100 bp minimum length cutoff [12]. The remaining high quality reads were *de novo* assembled using Trinity version 2.5.1 [13]. The assembled contigs were clustered using the CD-HIT-EST program with 90% identity threshold [14] and the longest representative sequence in each cluster was selected using a custom python script. The completeness of the assembly was evaluated by Benchmarking Universal Single-Copy Orthologs (BUSCO) [15] using the embryophyta_odb9 database containing 1440 BUSCO categories. The contigs were annotated using the uniprot database [16], in addition to Arabidopsis [17] and tomato [18] protein sequences using DIAMOND [19] version 0.9.22. Gene Ontology (GO) annotation was performed using AgBase version 2.0 [20] and GO enrichment analysis was conducted using the PANTHER version 11 with conservative Bonferroni correction for multiple testing [21]. The RNA-seq reads were mapped against the *de novo* transcriptome assembly using Salmon version 0.8.1 [22] and differential gene expression analysis was performed using the generalized linear model (glm) functionality of the edgeR package [23]. Contigs with at least two-fold expression

difference and False Discovery Rate (FDR) < 0.01 were considered differentially expressed.

2.4 Clustering

The differentially expressed contigs (DECs) were hierarchically clustered into 30 groups by expression similarity using the hclust function of the stats package version 3.6.0 [24]. The clustering was done using the complete method, which considers the largest value of dissimilarities between clusters. The package dendextend version 1.9.0 [25] was used to plot a dendrogram demonstrating members which are similar in a subgroup, and members which are dissimilar and in distinct clusters. The results were then put through log transformation and displayed with a heatmap, using the gplots package [26] version 3.0.1.

2.5 Phylogenetic analysis

The phylogenetic relationship among 37 plant species from seven families was assessed using the chloroplast Maturase K (matK) gene. The protein sequences of the MatK gene were downloaded from the Genebank non redundant protein database. The sequences were aligned using the ClustalW program [27] and phylogenetic tree was constructed using Phylogeny.fr [28] using the maximum likelihood method with 1000 Bootstrap replicates.

2.6 cDNA synthesis & qPCR analysis

Total RNA was extracted from frozen leaf and root samples and DNase treated as described above. cDNA was prepared from 100 ng total RNA using Superscript III first strand cDNA synthesis kit (Invitrogen USA). qPCR Primers were designed using the online Primer 3 software [29]. The housekeeping genes Serine/threonine-protein phosphatase 2A (PP2A)

and Cyclophilin (CYP) were used as internal controls to normalize the data [30]. Three biological replicates were used. qPCR was run on the Bio-Rad CFX96 machine using the following conditions: 95 °C for 5 min, followed by 40 cycles of 95 °C for 20 sec and 60 °C for 1 min. The fold change in gene expression levels was calculated using the $2^{-\Delta\Delta CT}$ method [31].

3 Result and Discussion

3.1 Sequencing and de novo assembly

To obtain an overview of the *S. hispanica* transcriptome, RNA-Seq libraries were prepared from leaf and root tissues of two week old seedlings. A total of 230 million high quality 150 bp paired-end reads were generated. The reads were *de novo* assembled into 333,889 contigs greater than 300 bp. The number of contigs assembled is considerably higher than the number of protein-coding genes in well studied plants with similar size genomes such as Arabidopsis (35,386), *Medicago truncatula* (62,319), *Ananas comosus* (27,024), and *Populus trichocarpa* (73,013) [32], suggesting transcript redundancy. Unlike genome-guided assemblers, the currently available *de novo* assembly programs are known to generate high level of redundancy. Among the contributors of this redundancy are the sequencing errors and single nucleotide polymorphisms (SNPs) which create mismatches. Accordingly, redundant sequences get generated as the assembly programs fail to consolidate highly similar sequences. This fact is exacerbated with increasing the number of reads used in the transcriptome assembly [33]. To assess the completeness of our

transcriptome and the level of redundancy, BUSCO analysis was performed which revealed a completeness score of 92.8% (Figure 1).

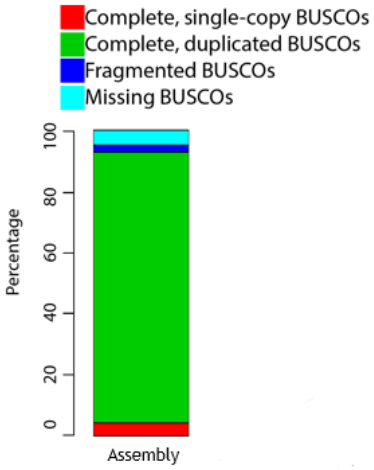

Figure 1: Results of the initial BUSCO analysis of the contigs

This indicates that most of the core gene set is present in our assembly, suggesting a high quality assembly. However, as anticipated, high level (88.8%) redundancy was evident. The redundant sequences in our initial assembly were consolidated using the CD-HIT-EST

program, which resulted in 103,367 contigs and reduced the redundancy to 42% while maintaining BUSCO completeness score of 92.2% (Figure 2).

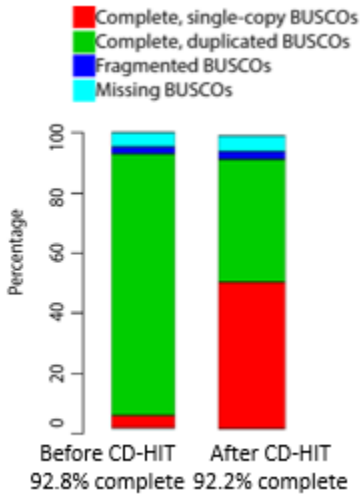

Figure 2: Comparison of the two BUSCO analyses; before and after CD-HIT

The remaining redundancy could be attributed to the heterogeneity of the *S. hispanica* genotype sequenced in this study, in addition to sequencing and assembly errors. Around 53% of the assembled contigs had length distribution between 300 and 1000 base pairs (bp)

(Figure 3), with N50 equal to 2330 bp and maximum transcript size of 26,500 bp (Table 1).

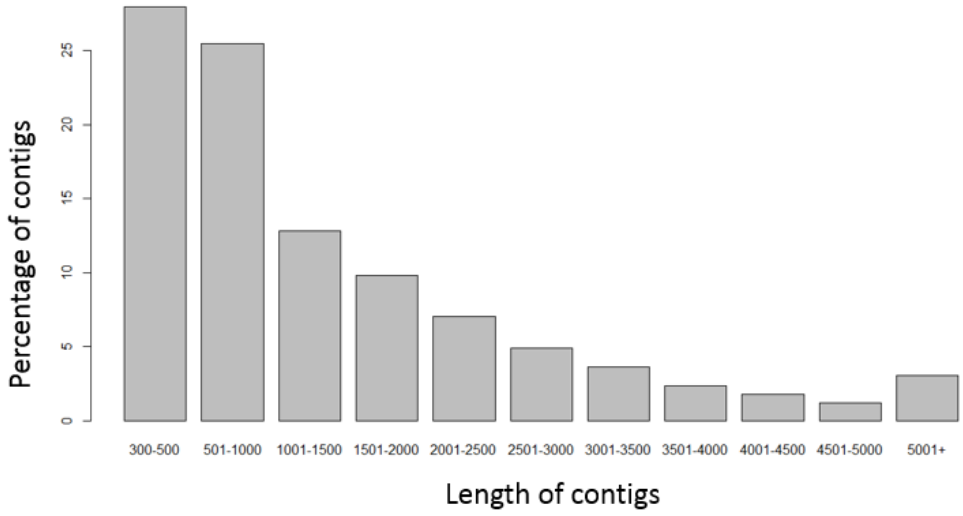

Figure 3: The length distribution of the contigs after CD-HIT consolidation

Table 1: Statistics of our assembly

Total Number of Contigs	103,367
Mean Length of Contigs	1,445
N50 Length of Contigs	2,330
Maximum Length of Contigs	26,781

3.2 Annotation and phylogenetic analysis

Based on Blastx analysis, 69% of the assembled contigs were annotated against the uniprot database with an E-value cut-off of 1e-3. Among these results, 211 contigs matched to non-plant species and were filtered out, leaving a total of 71,401 *S. hispanica* contigs matching to 30,628 unique sequences of plant origin in the Uniprot database. According to the

species distribution, a total of 102 plant genera showed homology to at least 10 *S. hispanica* sequences. The top 10 species belonged to orders Lamiales, Solanales, Gentianales, and Ericales (Figure 4).

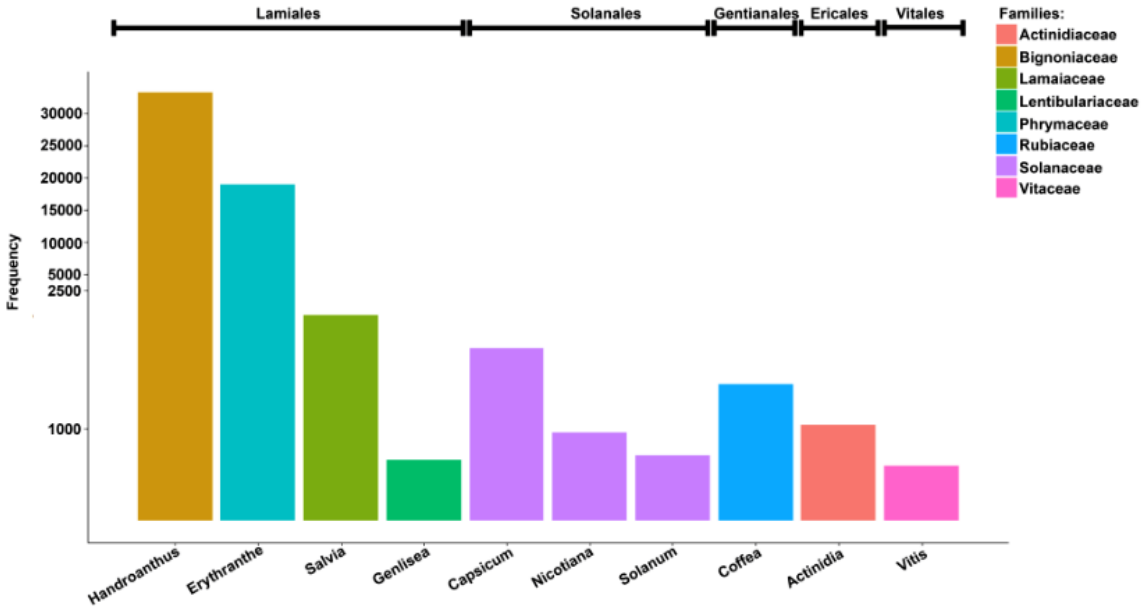

Figure 4: Species distribution of the sequences that our contigs matched to

Phylogenetic analysis was conducted using the maximum-likelihood method based on the chloroplast Maturase K (matK) gene, which has been widely used in plant analysis at family and genus level [34]. *S. hispanica* (family *Lamiaceae*) grouped with families *Solanaceae* and *Rubiaceae* (Figure 5), consistent with the top species showing homology to *S. hispanica* contigs.

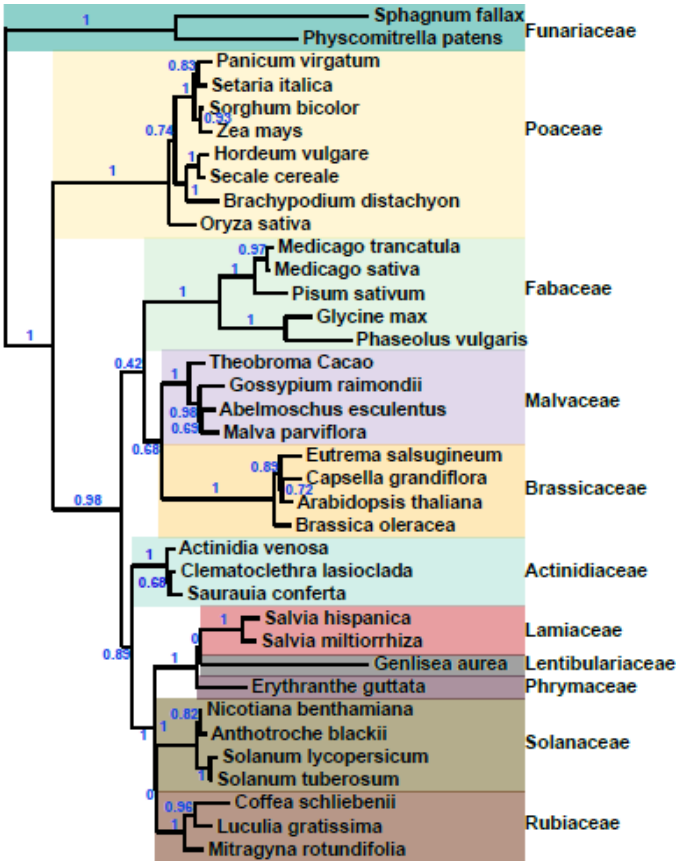

Figure 5: Phylogenetic tree showing the similarity of other species to the *S. hispanica*

The assembled transcripts were further annotated using Gene Ontology (GO) [35] and KEGG [36] databases. Diverse set of GO terms are represented in the assembled transcriptome (Figure 6). The biosynthetic, cellular protein modification, and cellular nitrogen compound metabolic processes are the top three representative terms within the Biological Process category. Ion binding is the top term in the Molecular Function category followed by Kinase and Oxidoreductase activities and DNA binding. The top three terms in the Cellular Component category are intracellular, nucleus, and cell.

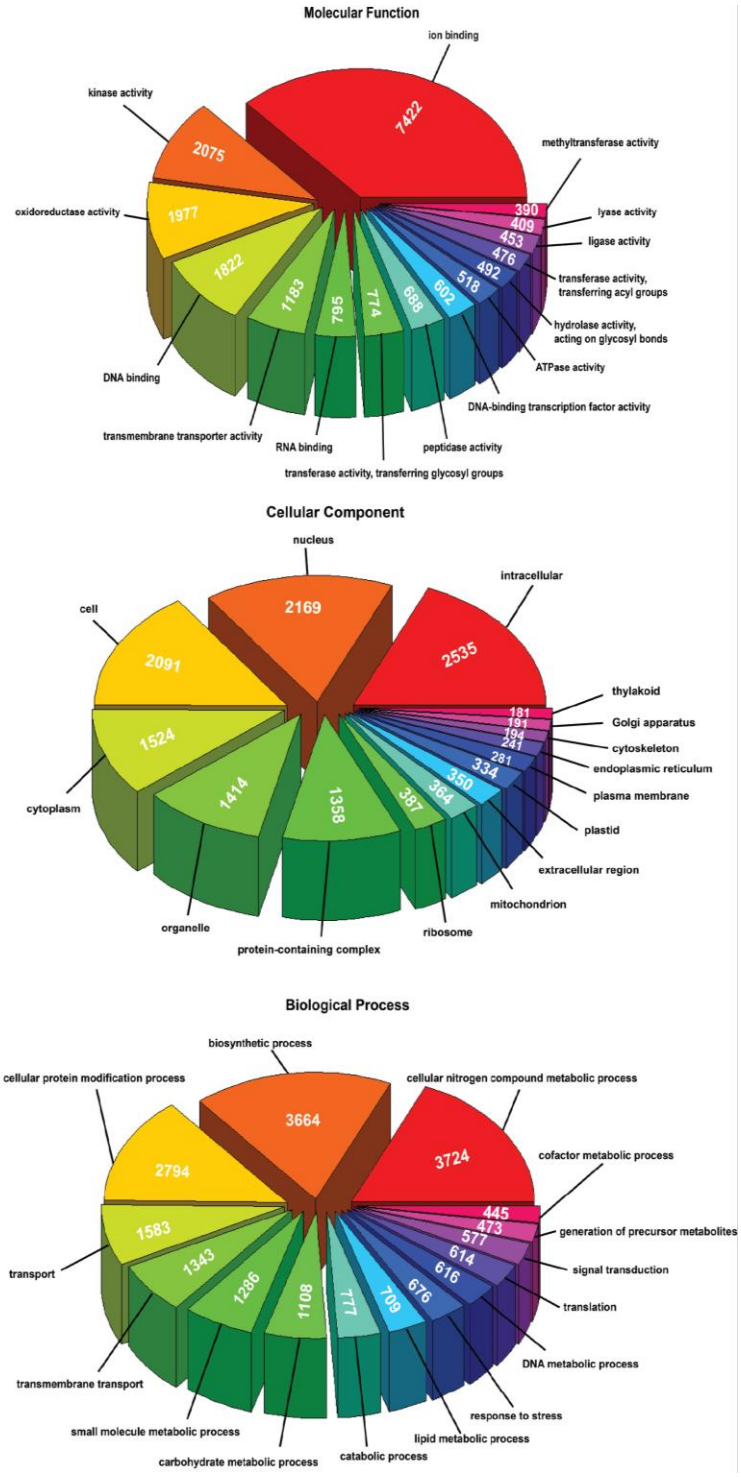

Figure 6: Diverse set of GO terms based on the annotations of our assembly

3.3 Differential gene expression and enrichment analysis

The leaf and root RNA-Seq reads were independently mapped against the assembled contigs and differential expression analysis was performed using the edgeR package [23] version 3.8. A total of 14,267 contigs showed significant difference (fold change >= 2; FDR < 0.01) in expression, among which 6,151 and 8,116 contigs were up-regulated in the leaf and root, respectively. Enrichment analysis of the differentially expressed contigs (DECs) and comparison between leaf and root tissues identified diverse and non-overlapping GO terms (Figure 7). Overall, fold enrichment of the GO terms was higher in the leaf compared to the root.

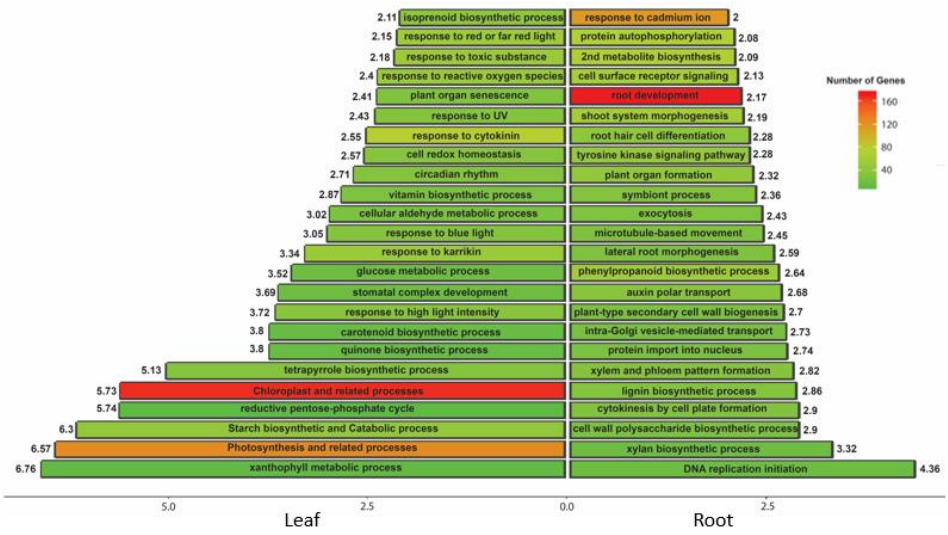

Figure 7: Gene Ontology enrichment analysis. Length of bars represent the fold enrichment and is also shown next to the bars.

For ease of visualization, a heatmap was generated based on hierarchical clustering of the DECs according to their expression levels (Figure 8). The leaf specific cluster 28 was enriched for lignin metabolic process, while the root specific cluster 16 was enriched for photosynthesis (Figure 9).

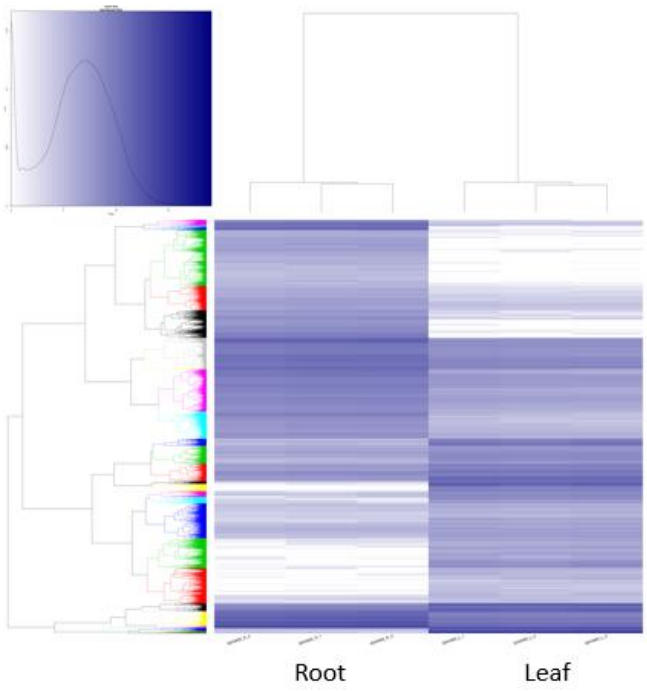

Figure 8: Dendrogram of contigs clustered by similarity, grouped into 30 clusters with heatmap demonstrating the expression

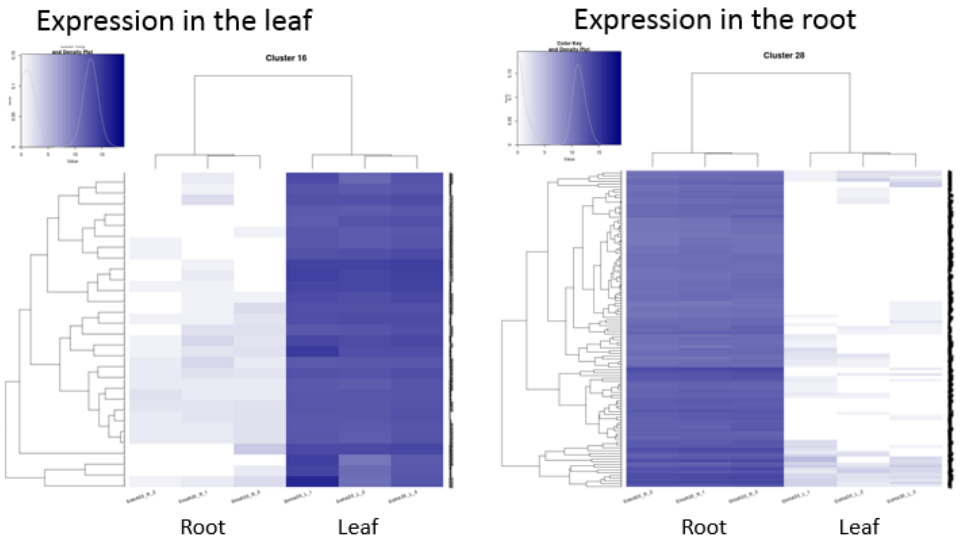

Figure 9: Clusters 16 and 28 from Figure 8

References

[1] R. Ayerza and W. Coates, *Chia: Rediscovering a Forgotten Crop of the Aztecs.*, University of Arizona, 2005.

[2] J. P. Cahill, "Ethnobotany of Chia, Salvia hispanica L. (Lamiaceae)," *Economic Botany,* vol. 57, no. 4, pp. 604-618, 2003.

[3] E. Reyes-Caudillo, A. Tecante and M. Valdivia-López, "Dietary fibre content and antioxidant activity of phenolic compounds present in Mexican chia (Salvia hispanica L.) seeds," *Food Chemistry,* vol. 107, no. 2, pp. 656-663, 2008.

[4] M. R. Sandoval-Oliveros and O. Paredes-López, "Isolation and Characterization of Proteins from Chia Seeds (Salvia hispanica L.)," *Journal of Agricultural and Food Chemistry,* vol. 61, no. 1, pp. 193-201, 2013.

[5] R. Ayerza and W. Coates, "Protein Content, Oil Content and Fatty Acid Profiles as Potential Criteria to Determine the Origin of Commercially Grown Chia (Salvia hispanica L.)," *Industrial Crops Products,* no. 34, pp. 1366-1371, 2011.

[6] G. Ouzounidou, V. Skiada, K. K. Papadopoulou, N. Stamatis, V. Kavvadias, E. Eleftheriadis and F. Gaitis, "Effects of soil pH and arbuscular mycorrhiza (AM) inoculation on growth and chemical composition of chia (Salvia hispanica L.) leaves," *Brazilian Journal of Botany,* vol. 38, no. 3, pp. 487-495, 2015.

[7] M. Ahmed, I. P. Ting and R. W. Scora, "Leaf Oil Composition of Salvia hispanica L. from Three Geographical Areas," *Journal of Essential Oil Research,* vol. 6, no. 3, pp. 223-228, 1994.

[8] H. S. Elshafie, L. Aliberti, M. Amato, V. De Feo and I. Camele, "Chemical composition and antimicrobial activity of chia (Salvia hispanica L.) essential oil," *European Food Research and Technology,* vol. 244, no. 9, pp. 1675-1682, 2018.

[9] M. Amato, M. C. Caruso, F. Guzzo, F. Galgano, M. Commisso, R. Bochicchio, R. Labella and F. Favati, "Nutritional quality of seeds and leaf metabolites of Chia (Salvia hispanica L.) from Southern Italy," *European Food Research and Technology,* vol. 241, no. 5, pp. 615-625, 2015.

[10] R. Kumar, Y. Ichihashi, S. Kimura, D. H. Chitwood, L. R. Headland, J. Peng, J. N. Maloof and N. R. Sinha, "A high-throughput method for Illumina RNA-Seq library preparation," *Frontiers in Plant Science,* vol. 3, p. 202, 2012.

[11] *This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant.*

[12] A. M. Bolger, M. Lohse and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," *Bioinformatics,* vol. 30, no. 15, pp. 2114-2120, 2014.

[13] M. Grabherr, B. Haas, M. Yassour, J. Levin, D. Thompson, I. Amit, X. Aidconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman and A. Regev, "Full-length transcriptome assembly from RNA-Seq data without a reference genome," *Nature Biotechnology,* vol. 29, pp. 644-652, 2011.

[14] L. Fu, B. Niu, Z. Zhu, S. Wu and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinformatics,* vol. 28, no. 23, pp. 3150-3152, 2012.

[15] F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva and E. M. Zdobnov, "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs," *Bioinformatics,* vol. 31, no. 19, pp. 3210-3212, 2015.

[16] The UniProt Consortium, "UniProt: a worldwide hub of protein knowledge," *Nucleic Acids Research,* vol. 47, no. D1, pp. D506-D515, 2019.

[17] "The Arabidopsis Information Resource (TAIR)," [Online]. Available: ftp://ftp.arabidopsis.org/Proteins/TAIR10_protein_lists/. [Accessed 8 May 2019].

[18] N. Fernandez-Pozo, N. Menda, J. Edwards, S. Saha, I. Tecle, S. Strickler, A. Bombarely, T. Fisher-York, A. Pujar, H. Foerster, A. Yan and L. Mueller, "The Sol Genomics Network (SGN) - from genotype to phenotype to breeding," *Nucleic Acids Research,* vol. 43, no. D1, pp. D1036-D1041, 2014.

[19] B. Buchfink, C. Xie and D. H. Huson, "Fast and sensitive protein alignment using DIAMOND," *Nature,* vol. 12, pp. 59-60, 2015.

[20] F. McCarthy, N. Wang, G. Magee, B. Nanduri, L. ML, E. Camon, D. Barrell, D. Hill, M. Dolan, W. Williams, D. Luthe, S. Bridges and S. Burgess, "AgBase: a functional genomics resource for agriculture," *BMC Genomics,* vol. 7, p. 229, 2006.

[21] H. Mi, X. Huang, A. Muruganujan, H. Tang, C. Mills, D. Kang and P. D. Thomas, "PANTHER version 11: expanded annotation data from Gene Ontology and

Reactome pathways, and data analysis tool enhancements," *Nucleic Acids Research,* vol. 45, no. D1, pp. D183-D189, 2017.

[22] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry and C. Kingsford, "Salmon provides fast and bias-aware quantification of transcript expression," *Nature Methods,* vol. 14, pp. 417-419, 2017.

[23] M. Robinson, D. McCarthy and G. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics,* vol. 26, no. 1, pp. 139-40, 2010.

[24] R Core Team, *R: A language and environment for statistical computing*, Vienna, 2018.

[25] T. Galili, "dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering," *Bioinformatics,* vol. 31, no. 22, pp. 3718-3720, 2015.

[26] Warnes, Bolker, Bonebakker, Gentleman, Huber, Liaw, Lumly, Maechler, Magnusson, Moeller, Schwartz and Venables, *gplots: Various R Programming Tools for Plotting Data,* R package version 3.0.1, 2016.

[27] [Online]. Available: https://www.ebi.ac.uk/Tools/msa/clustalw2/.

[28] A. Dereeper, V. Guignon, G. Blanc, S. Audic, S. Buffet, F. Chevent, J. Dufayard, S. Guindon, V. Lefort, M. Lescot, J. Claveries and O. Gascuel, "Phylogeny.fr: robust phylogenetic analysis for the non-specialist," *Nucleic Acids Research,* vol. 36, pp. W465-9, 2008.

[29] [Online]. Available: http://bioinfo.ut.ee/primer3-0.4.0/.

[30] R. Gopalam, S. Rupwate and A. Tumaney, "Selection and validation of appropriate reference genes for quantitative real-time PCR analysis in Salvia hispanica.," *PLoS One,* vol. 12, no. 11, 2017.

[31] K. J. Livak and T. D. Schmittgen, "Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method," *Methods,* vol. 25, no. 4, pp. 402-408, 2001.

[32] [Online]. Available: https://phytozome.jgi.doe.gov.

[33] X. Huang, K.-G. Chen and P. A. Armbruster, "Comparative performance of transcriptome assembly methods for non-model organisms," *BMC Genomics,* 2016.

[34] W. Dong, J. Liu, J. Yu, L. Wang and S. Zhou, "Highly Variable Chloroplast Markers for Evaluating Plant Phylogeny at Low Taxonomic Levels and for DNA Barcoding," *PLoS ONE,* vol. 7, no. 4, 2012.

[35] [Online]. Available: http://www.geneontology.org/.

[36] [Online]. Available: http://www.genome.jp/kegg/.