Gene Expression Analysis using RNA-Seq

Doriaun Calvo, Shawn Culpepper, Ruben Garcia, Nathan Pellegrin

A. Background

Arabidopsis thaliana is a small plant apart of the mustard family native to Europe, Asia, and Africa - commonly found alongside roads and disturbed lands. One of the important characteristics for plants to function are stomata (stoma), and they are tiny pores or openings on the underside of land plant leaves and on top for aquatic that allows carbon dioxide and oxygen exchange for photosynthesis, and helps maintain water loss by closing in extreme (hot/dry) conditions. During the day, a plant's intake of carbon dioxide, water, and light are essential to produce glucose as a food source, and excess water vapor and oxygen are released into their surrounding environment through the stomata opening and closed at night to conserve water.

The BIG question now is how is this related to varying levels of humidity? The opening and closing of stomata has a lot to do with diffusion as well as a response to its external environment, and humidity is a great example of how certain conditions can impact plant physiology because when humidity conditions are optimal (moderate temperatures and moisture), stomata are open and if humidity levels decrease due to increased temperatures (deserts) or windy conditions, more water vapor is needed to diffuse from the plant into the air and plants must close their stomata to prevent excess water loss.

Some of the guided research questions we used for understanding the results were 1) does changing humidity affect transcriptomics? 2) Are there differentially expressed genes as a response to changes in humidity? 3) Are there clusters of genes that share the same pattern of expression across levels of humidity? 4)
What else can we learn from databases about these highly differential expression patterns and coexpressed genes?

B. Methods

The following steps summarize the experiment described in GEO accession GSE236463. Three groups of Arabidopsis thaliana plants (four in each group for a total of 12 biological replicates) were exposed to low, medium and high humidity for one hour. RNA was extracted from the plant tissues and paired-end sequenced using an Illumina NovaSeq 6000. Alignment to the TAIR 10 reference genome was performed using HISAT2 software and raw expression counts were normalized according to the fragments per Transcripts Per Million reads (TPM) method. RSEM (RNA-Seq by Expectation Maximization) software was used to quantify gene abundances. Data file containing results for all genes was downloaded from GSE236463.

Data pre-processing and gene expression analysis was performed in R (4.3.2). The edgeR library was used to filter low expression genes (genes where sum of CPM values across samples was less than 7 were excluded) and to perform multidimensional scaling to visually confirm the presence of differentially expressed genes. The limma library was used to estimate the mean-variance relationship and compute appropriate observation-level weights, fit a linear model of log fold change across treatments for

each gene and obtain Bayesian-adjusted t-tests on the null hypothesis of no difference. Highly differentially expressed genes were identified as those with p-value <0.01.

Gene coexpression analysis was performed after using the DESeq2 library to apply a variance stabilizing transformation and selecting the top 25% of genes based on expression levels. The WGCNA library was used to build coexpression networks based on correlated expression patterns across samples. Gene modules were identified by hierarchical clustering The top four modules in terms of coverage over all genes, were selected for further analysis. Eigen-genes for each module was extracted and the MEk scores for each gene were calculated. R script for the above methods is available on github (https://github.com/nathanpellegrin/btech_610_project). The set of HDE genes, and the top 10 genes in each module, were submitted to Panther to obtain annotation data on biological function, molecular function and cell components.

C. Results and discussion

- High humidity did create a measurable difference in the gene expression of certain genes as measured by RNA abundance in Arabidopsis thaliana plants.

The expression counts file obtained from GEO contained 32,833 genes, 9,128 (27%) of which had no expression in any sample, and an additional 7,111 (22%) were found to have low expression. 105 genes were found to be highly differentially expressed (HDE) across treatment levels. Annotations for these HDE include, among other features, processes related to the growth of the cell wall and response to exogenous stimuli, including temperature.

- Co-expressed genes were identified in groups (modules)
    - Top four modules had a typically characteristic (eigen-gene) gene identified

Analysis of coexpression networks identified twelve gene modules. Brown, Turquoise, Blue and Yellow were the top four modules based on the number of genes associated.

- The Highly differentially expressed genes and the eigen-genes were submitted to Panther for purpose determination.
    - Majority of Molecular function results were for catalytic activity
    - Majority of Biological process results were for cellular response

Cross referencing with the tair/PANTHER databases grouped the HDE and coexpression genes based on molecular function and biological process.

- Successfully completed a RNA-Seq pipeline experiment pulling out interesting, biologically relevant gene expression patterns.

HDE and coexpression genes can be extrapolated, explored and cross referenced in databases proving the applicability of powerful bioinformatics tools.

References

1. Chang, J. (n.d.). *WGCNA Gene Correlation Network Analysis*. Bioinformatics Workbook. Retrieved December 15, 2023, from https://bioinformaticsworkbook.org/tutorials/wgcna.html
2. *GEO Accession viewer*. (n.d.). Retrieved December 11, 2023, from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM7548222
3. Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1), 559. https://doi.org/10.1186/1471-2105-9-559
4. Law, C. W., Alhamdoosh, M., Su, S., Dong, X., Tian, L., Smyth, G. K., & Ritchie, M. E. (2018). *RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR* (5:1408). F1000Research. https://f1000research.com/articles/5-1408
5. Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*(1), 323. https://doi.org/10.1186/1471-2105-12-323
6. Wang, J. (2023, November). *Lab 11* [Lab presentation]. BTECH 610 - Bioinformatics, St. Mary's College. https://github.com/wjidea/BTECH610_Bioinformatics/tree/main/lab11
7. Woodward, A. W., & Bartel, B. (2018). Biology in Bloom: A Primer on the *Arabidopsis thaliana* Model System. *Genetics*, *208*(4), 1337–1349. https://doi.org/10.1534/genetics.118.300755https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5887134/pdf/1337.pdf
8. https://www.arabidopsis.org/portals/education/aboutarabidopsis.jsp