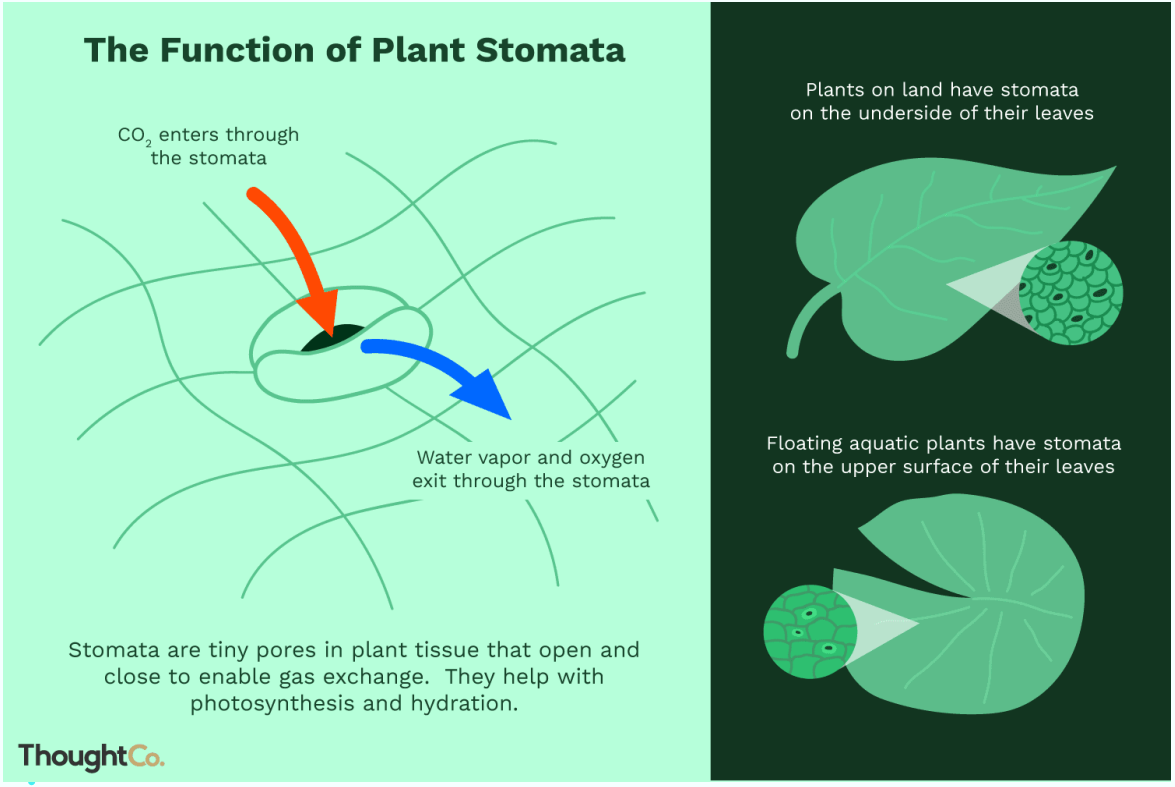
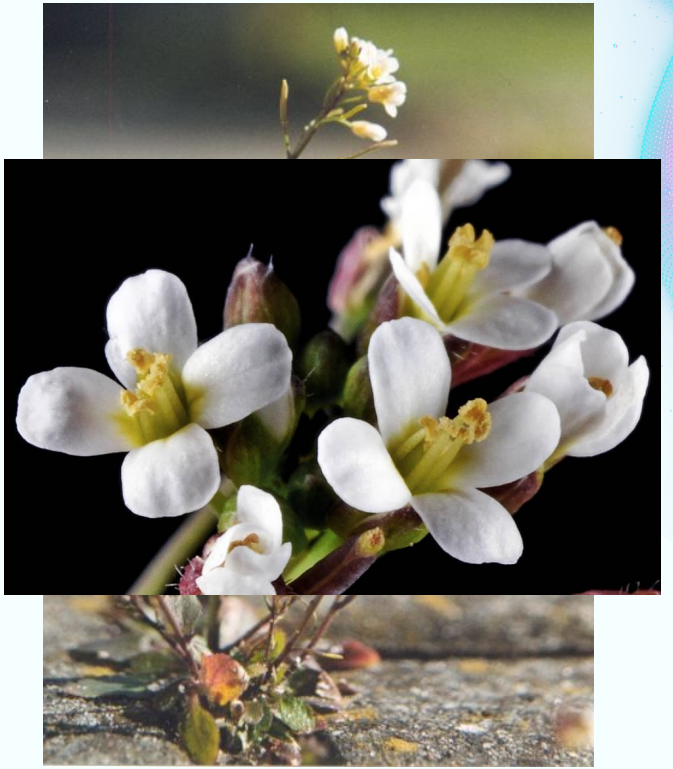
Gene Expression Analysis using RNA-Seq

Doriaun Calvo, Shawn Culpepper,
Ruben Garcia, Nathan Pellegrin

BACKGROUND


The Function of Plant Stomata
CO_2 enters through the stomata
Water vapor and oxygen exit through the stomata
Stomata are tiny pores in plant tissue that open and close to enable gas exchange. They help with photosynthesis and hydration.
ThoughtCo.
Plants on land have stomata on the underside of their leaves
Floating aquatic plants have stomata on the upper surface of their leaves



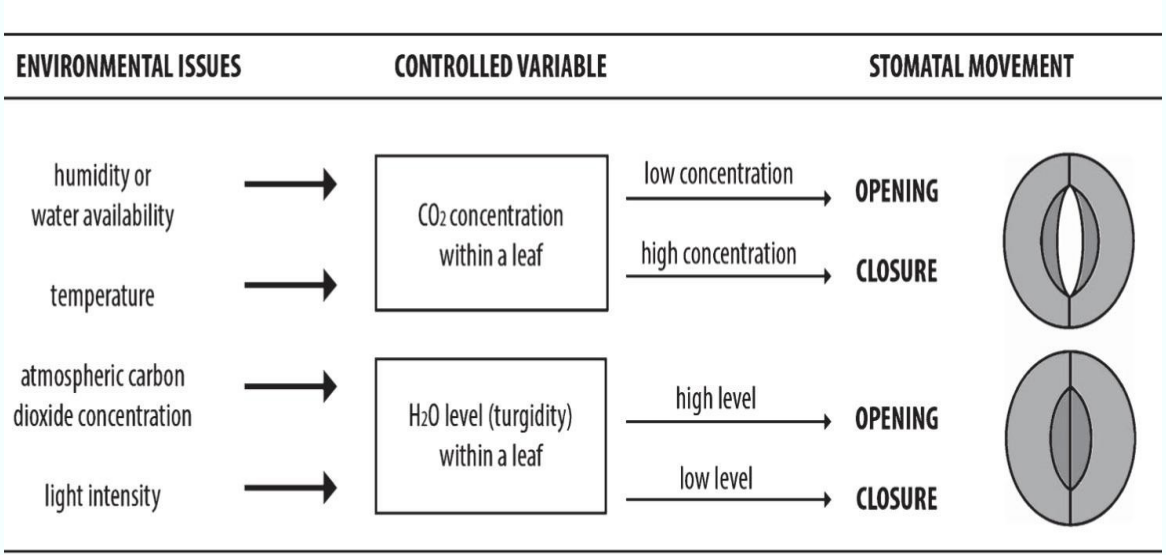


Fig. 3. Diagram showing the mechanism of stomatal movement.

Mechanism:

1. **Unfavorable conditions** = guard cells pump potassium ions out & close to prevent dehydration (excess water evaporation)
2. **Favorable conditions** = guard cells remain open to intake as much CO_2 and O_2 as possible to promote functional processes and

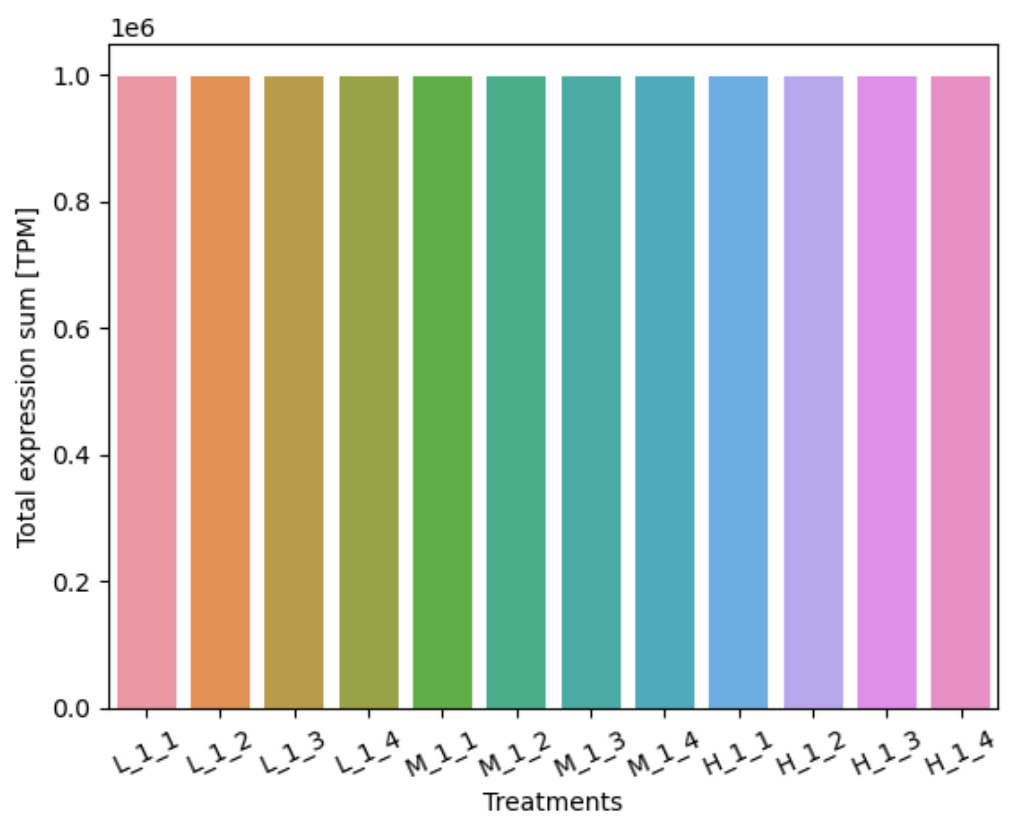
GUIDED RESEARCH QUESTIONS

1. Does change in humidity levels 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 HDE and coexpressed genes?

METHODS (1)

1. **Experimental Setup**:
    - Groups: Three groups of *Arabidopsis thaliana* plants.
    - Replicates: Four biological replicates per group, totaling 12.
    - Treatment: Exposure to low, medium, and high humidity for one hour.
2. **RNA Extraction**:
    - Lab process: Extraction of RNA from plant tissues.
3. **Sequencing**:
    - Method: Paired-end sequencing.
    - Equipment: Illumina NovaSeq 6000.
4. **Alignment**:
    - Reference Genome: TAIR 10.
    - Software: HISAT2 for alignment.
5. **Data Normalization**:
    - Method: Fragments per Transcripts Per Million reads (TPM).
6. **Quantification**:
    - Software: RSEM (RNA-Seq by Expectation Maximization) to quantify gene abundances.

Data Normalization (TPM)



- sns.barplot
- TPM (transcripts per million reads) - opposite order or normalization than RPKM
- RPK / total read counts = TPM data
- TPM facilitates more accurate reads across samples

Bar graph representation of total expression sum vs. treatment

RESULTS

DIFFERENTIAL GENE EXPRESSION

METHODS (2)

1. **Data Acquisition**:
   - Download gene expression data from GEO accession GSE236463.

2. **Data Pre-processing and Analysis in R (version 4.3.2)**

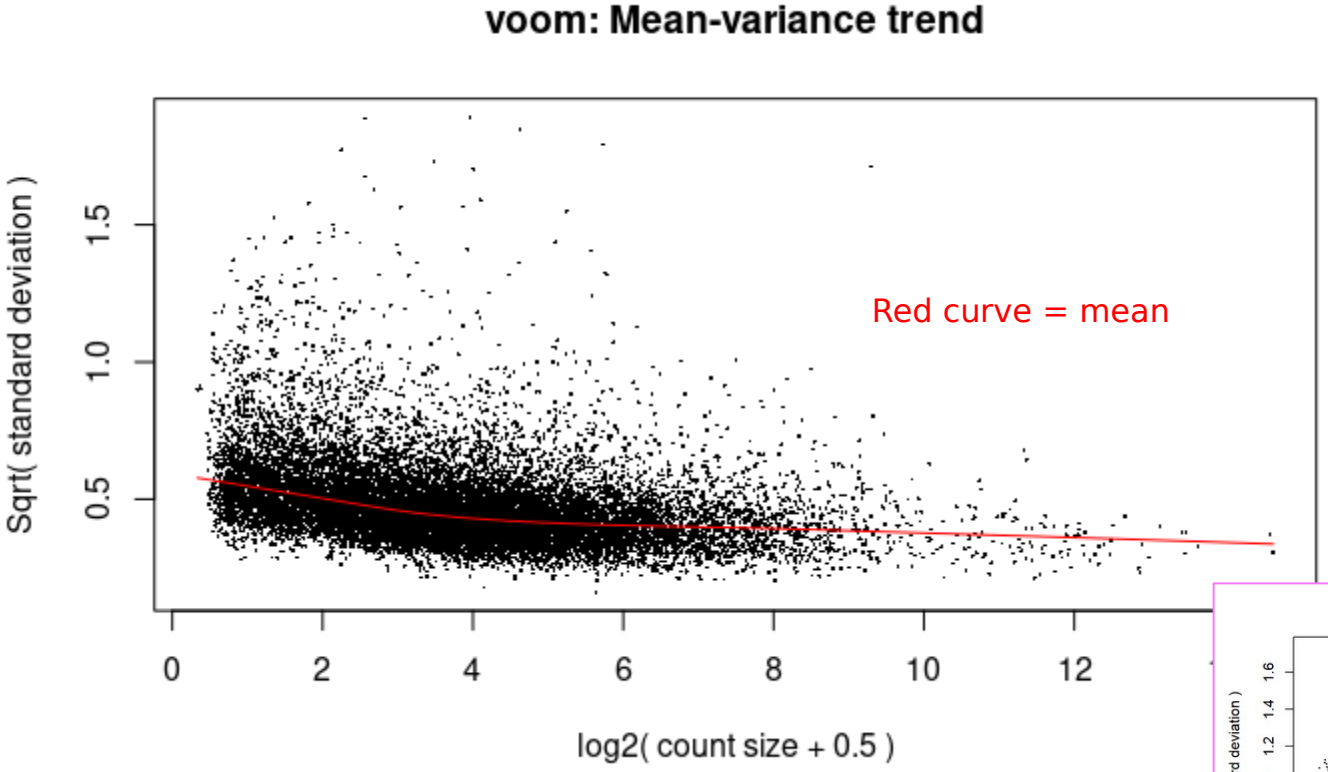
3. **Gene Expression Filtering (using edgeR library)**:
   - Criteria: Exclude genes with low expression (sum of CPM values less than 7 across samples).

4. **Multidimensional Scaling (MDS) Analysis (using edgeR)**:
   - Purpose: Visually confirm the presence of differentially expressed genes.

6. **Mean-Variance check and Linear Modeling (using limma library)**:
   - Estimation: Mean-variance relationship of genes.
   - Weights: Compute observation-level weights.
   - Model: Fit a linear model of log fold change across treatments for each gene.
   - **Statistical Testing**:
   - Method: Bayesian-adjusted t-tests.
   - Hypothesis: Test the null H of no difference in expression.

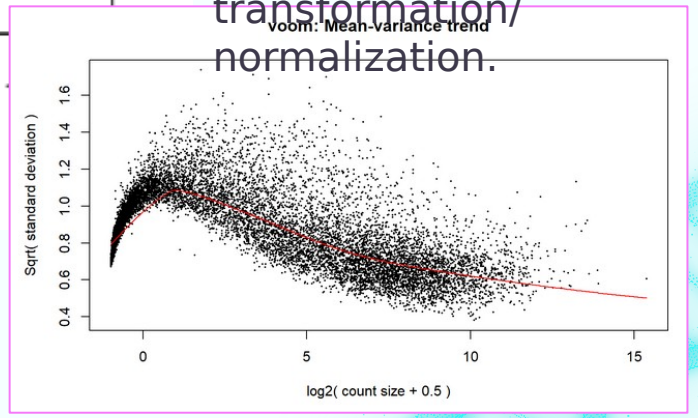
7. **Identification of Highly Differentially Expressed Genes**:
   - Criteria: Genes with p-value < 0.01 from t-tests

Residuals versus log (count)

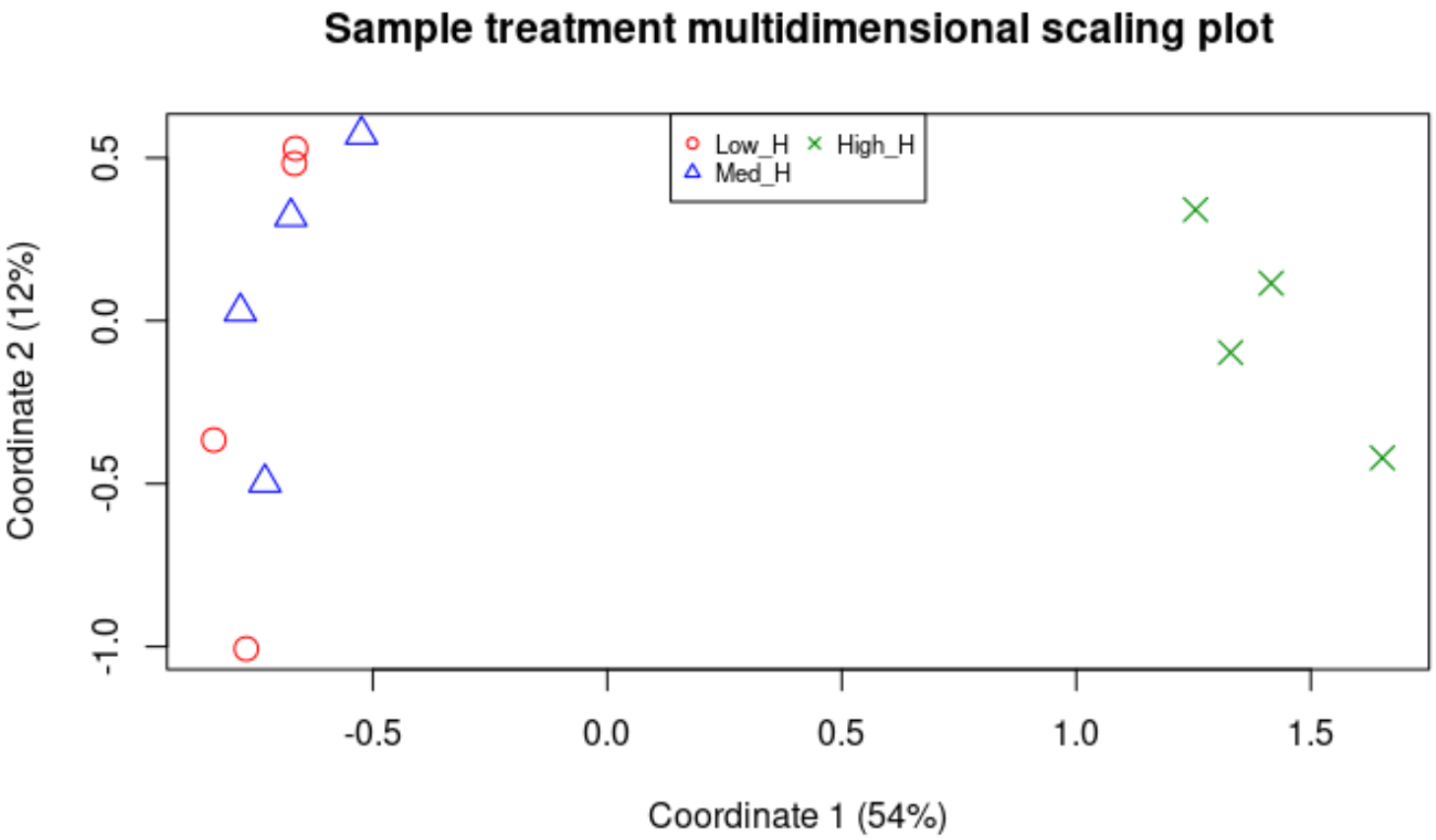


Residuals of linear models fitted to each gene (after applying log CPM) appear relatively uniform across expression levels.

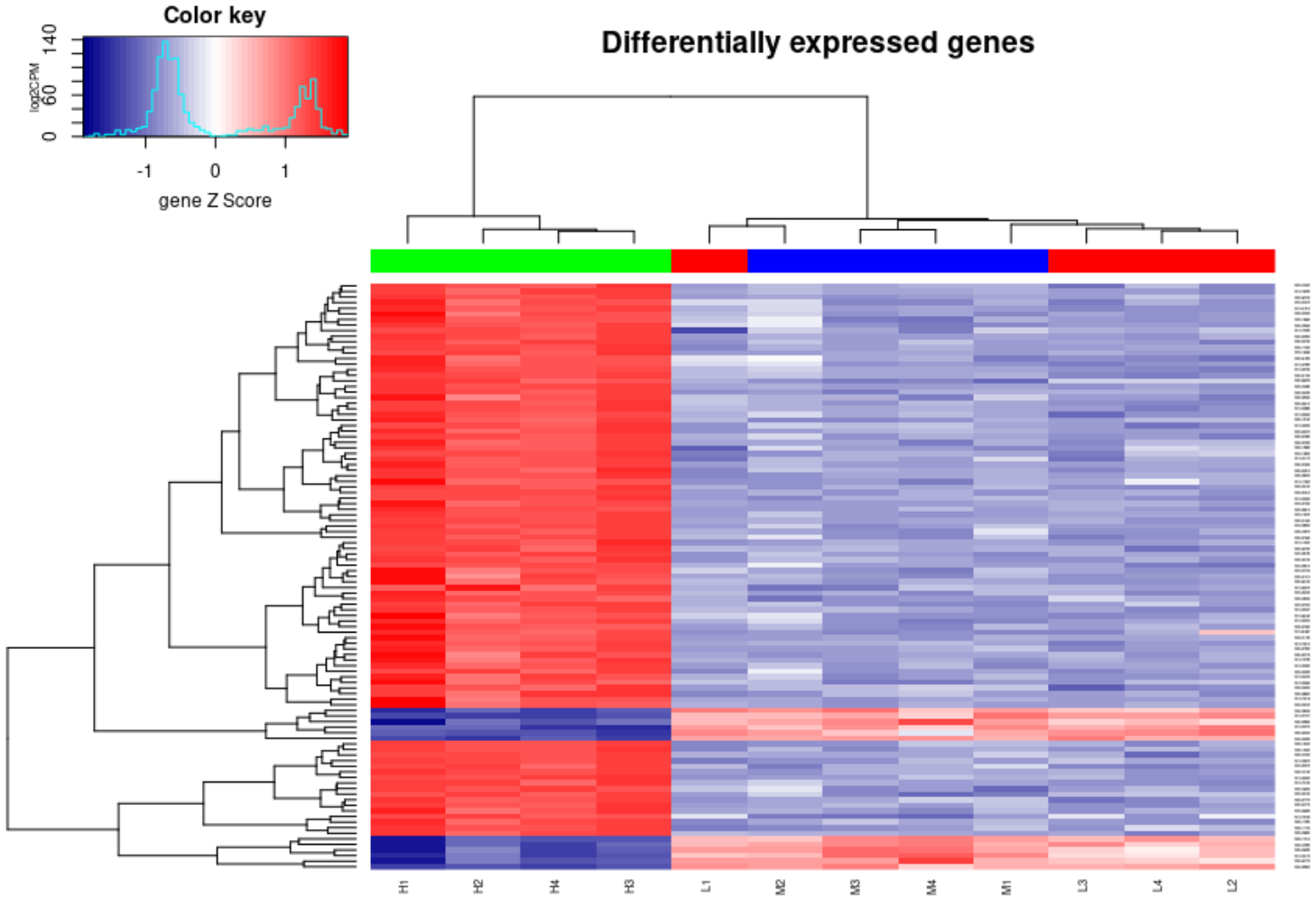
Example below (not from our data) shows ‘pathological’ case needing transformation/ normalization.



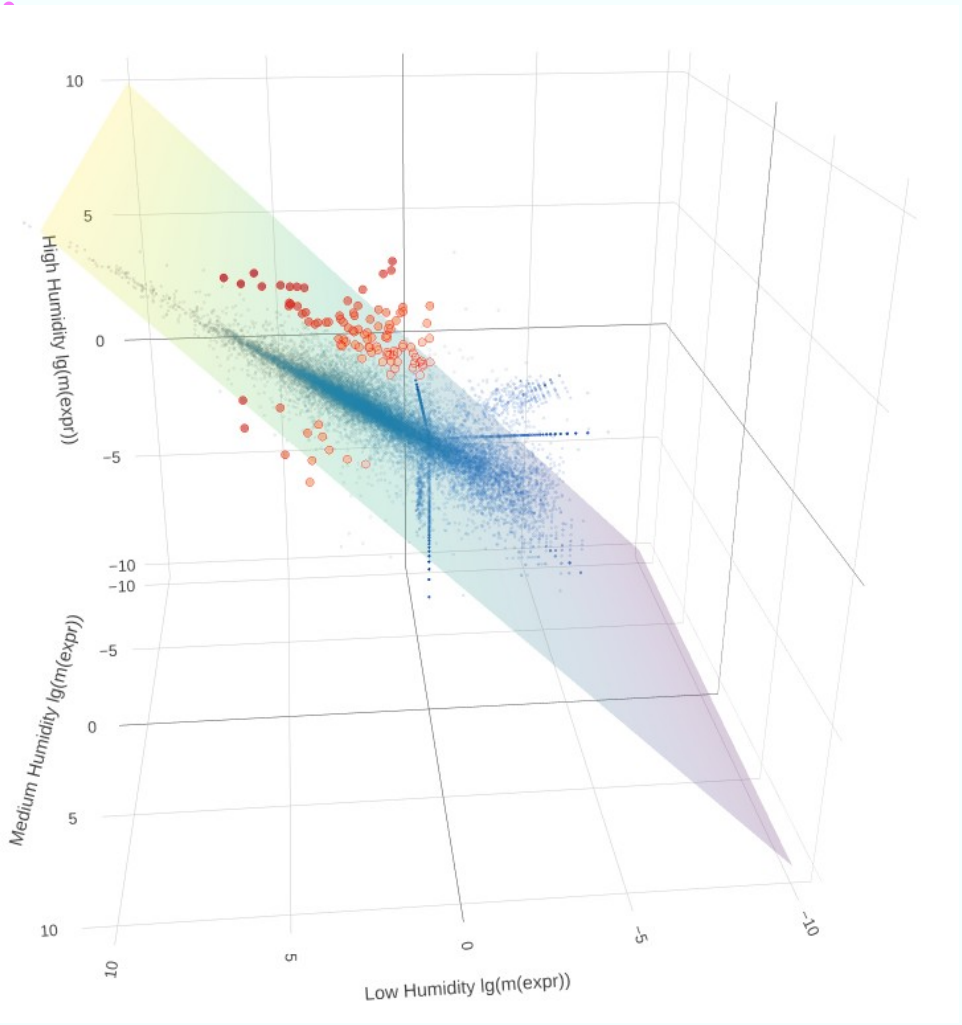
https://rpubs.com/jrgonzalezISGlobal/transcriptomic_analyses



MDS plot shows clear separation of samples based on treatment; high humidity samples are clustered and isolated, indicating differential expression of genes



- Heatmap shows high humidity samples are statistically differentiated from samples (left 4 cols) at both low and medium levels, with most genes expressing at greater rates (red cells) under high humidity (? %,N) except for two subgroups



3D Scatterplot provides visual confirmation of genes selected as highly differentially expressed (red circles). Relative to all other genes, high DE genes are in a sparse subspace at the edges of the distribution and located further from the plane of equality (perfectly correlated expression across treatments)

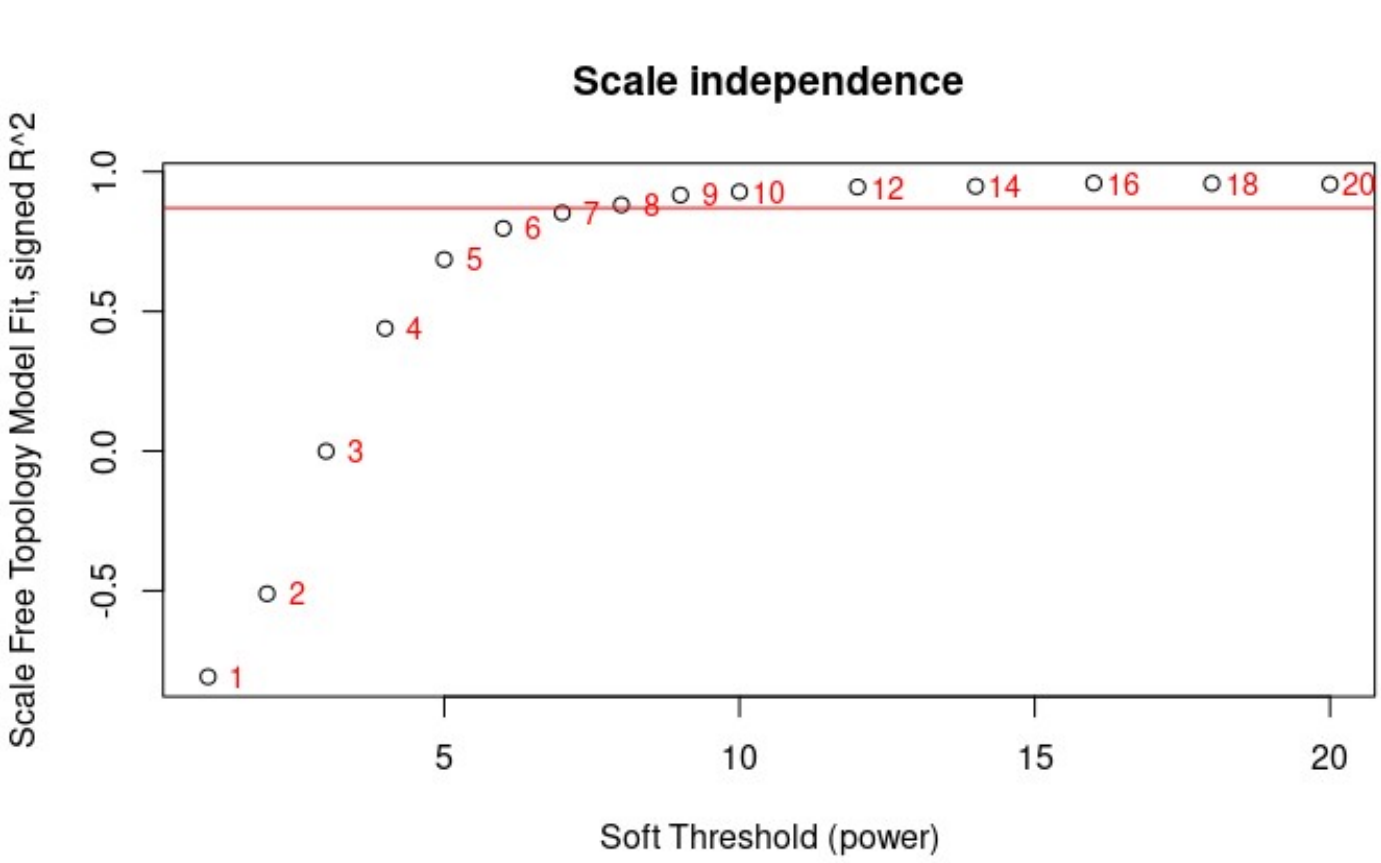
105 Highly DE Genes

```
  [1] "AT5G51550" "AT4G16563" "AT2G30930" "AT2G17230" "AT2G34510" "AT4G13340" "AT2G38310"
  [8] "AT2G42580" "AT3G62720" "AT4G08950" "AT1G22530" "AT5G44130" "AT1G76090" "AT3G23730"
 [15] "AT1G55330" "AT5G03120" "AT3G45970" "AT1G03870" "AT4G37450" "AT1G21910" "AT3G05490"
 [22] "AT2G44500" "AT5G57560" "AT1G14520" "AT2G39800" "AT3G28200" "AT1G03457" "AT2G33570"
 [29] "AT4G37240" "AT5G54380" "AT5G15350" "AT2G23100" "AT2G34770" "AT4G28190" "AT3G44990"
 [36] "AT2G06850" "AT5G24030" "AT1G23030" "AT5G66200" "AT1G22882" "AT4G30270" "AT2G36410"
 [43] "AT2G23130" "AT1G50040" "AT2G23290" "AT5G57550" "AT5G11740" "AT1G75310" "AT1G57990"
 [50] "AT5G05440" "AT1G72150" "AT5G62730" "AT1G75750" "AT1G72790" "AT5G20250" "AT1G72416"
 [57] "AT2G16660" "AT3G13520" "AT2G03090" "AT3G57930" "AT2G31730" "AT3G58850" "AT1G20070"
 [64] "AT5G44020" "AT1G17620" "AT1G22330" "AT3G05320" "AT4G24570" "AT5G49360" "AT1G53730"
 [71] "AT2G46330" "AT1G25560" "AT4G26690" "AT3G54810" "AT3G26290" "AT3G24550" "AT3G17390"
 [78] "AT3G30775" "AT1G22400" "AT3G62150" "AT1G35350" "AT2G01190" "AT1G07477" "AT4G15800"
 [85] "AT3G19680" "AT5G39860" "AT1G70090" "AT4G04745" "AT5G61590" "AT4G29310" "AT2G31010"
 [92] "AT1G61667" "AT3G55500" "AT1G24170" "AT4G33420" "AT3G59350" "AT5G37770" "AT5G64770"
 [99] "AT2G48030" "AT3G19150" "AT1G05170" "AT5G28030" "AT1G66160" "AT3G17510" "AT1G63570"
```

GENE COEXPRESSION NETWORKS

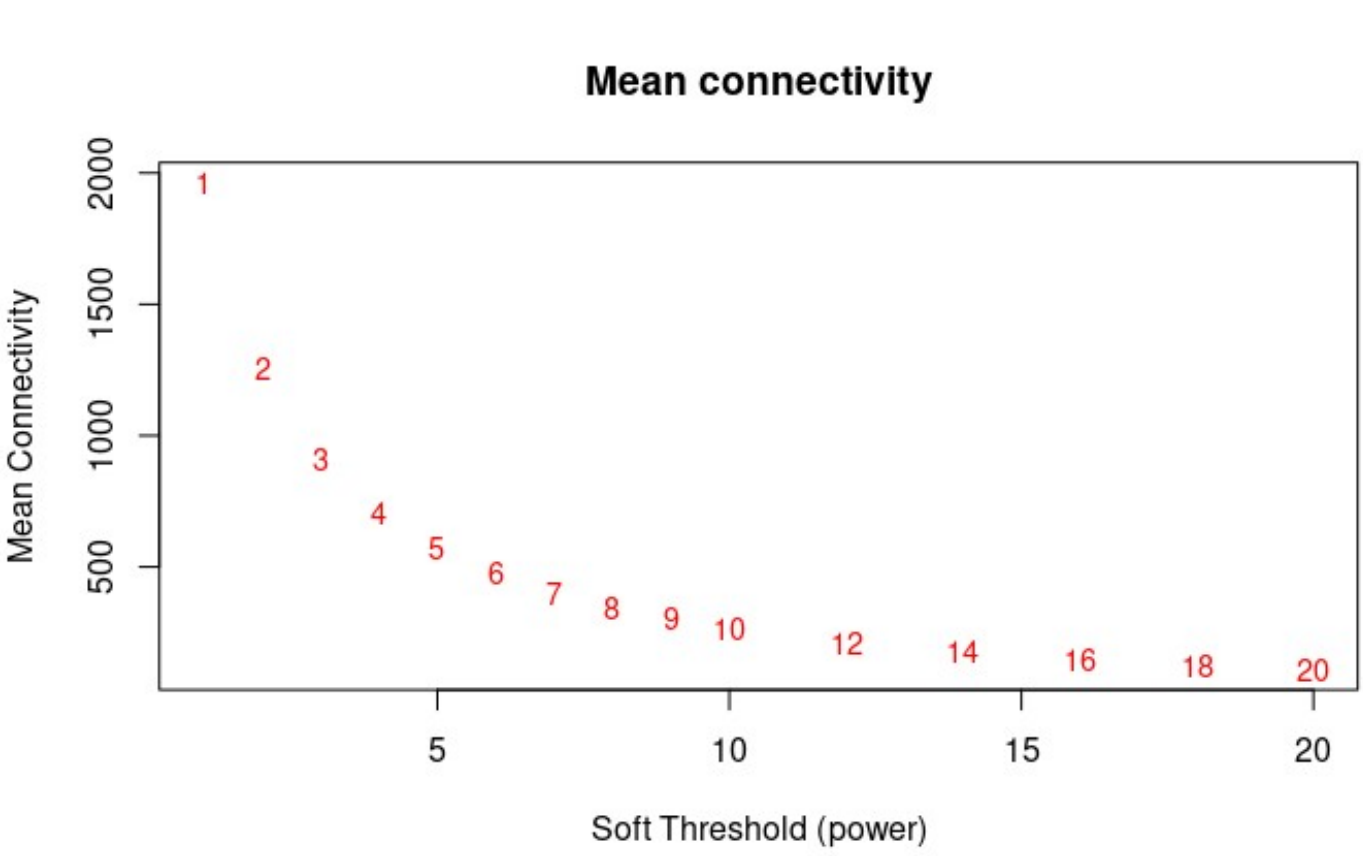
METHODS (3)

1. **Gene Coexpression Analysis (using DESeq2 library)**:
    - Transformation: Apply variance stabilizing transformation.
2. **Filter: Choose the top 25% of genes based on expression levels.**
3. **Coexpression Network Construction (using WGCNA library)**:
    - Method: Build networks based on correlated expression patterns across samples.
    - **Gene Module Identification**:
    - Technique: Hierarchical clustering to identify gene modules (dendrogram vis)
    - Eigen-gene and MEk
4. **Selection of Top 10 most similar genes in each module**

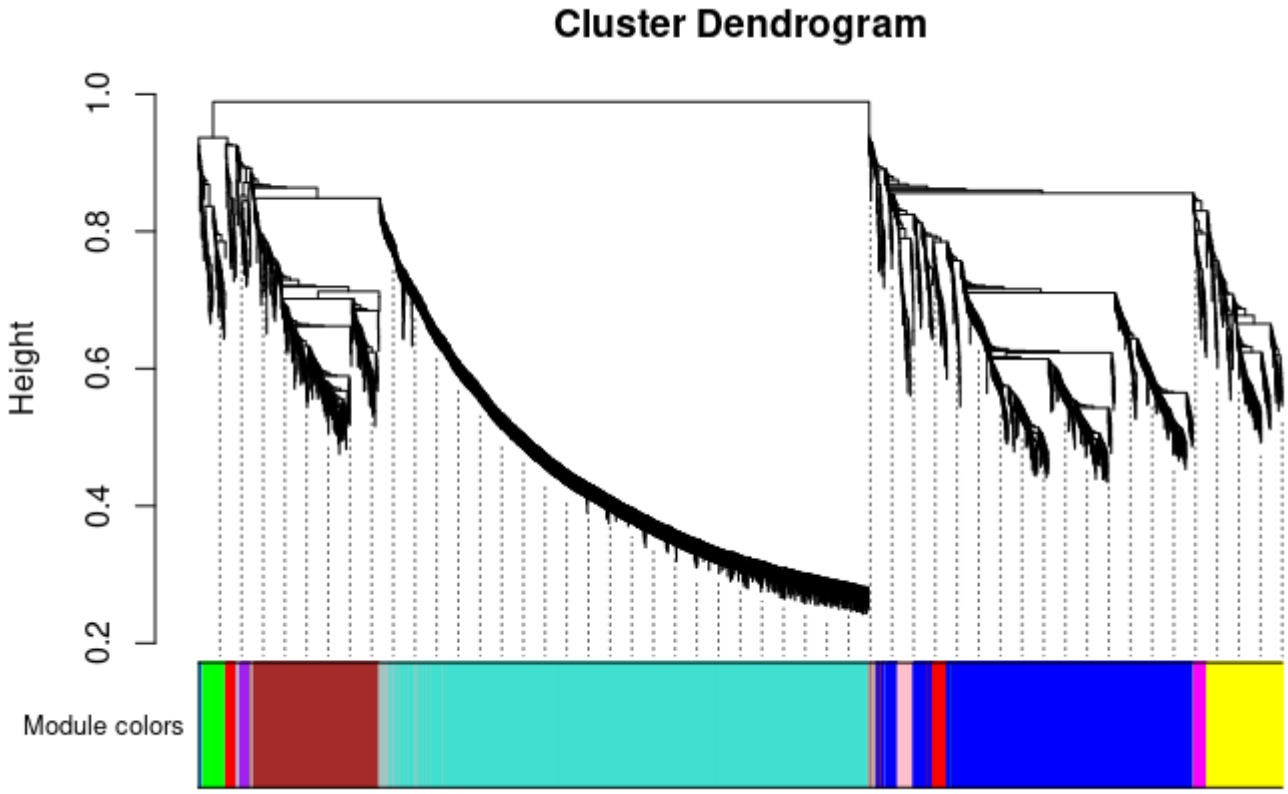


The WGCNA lib selects a “soft threshold” based on best R^2 fit to a scale free network topology.

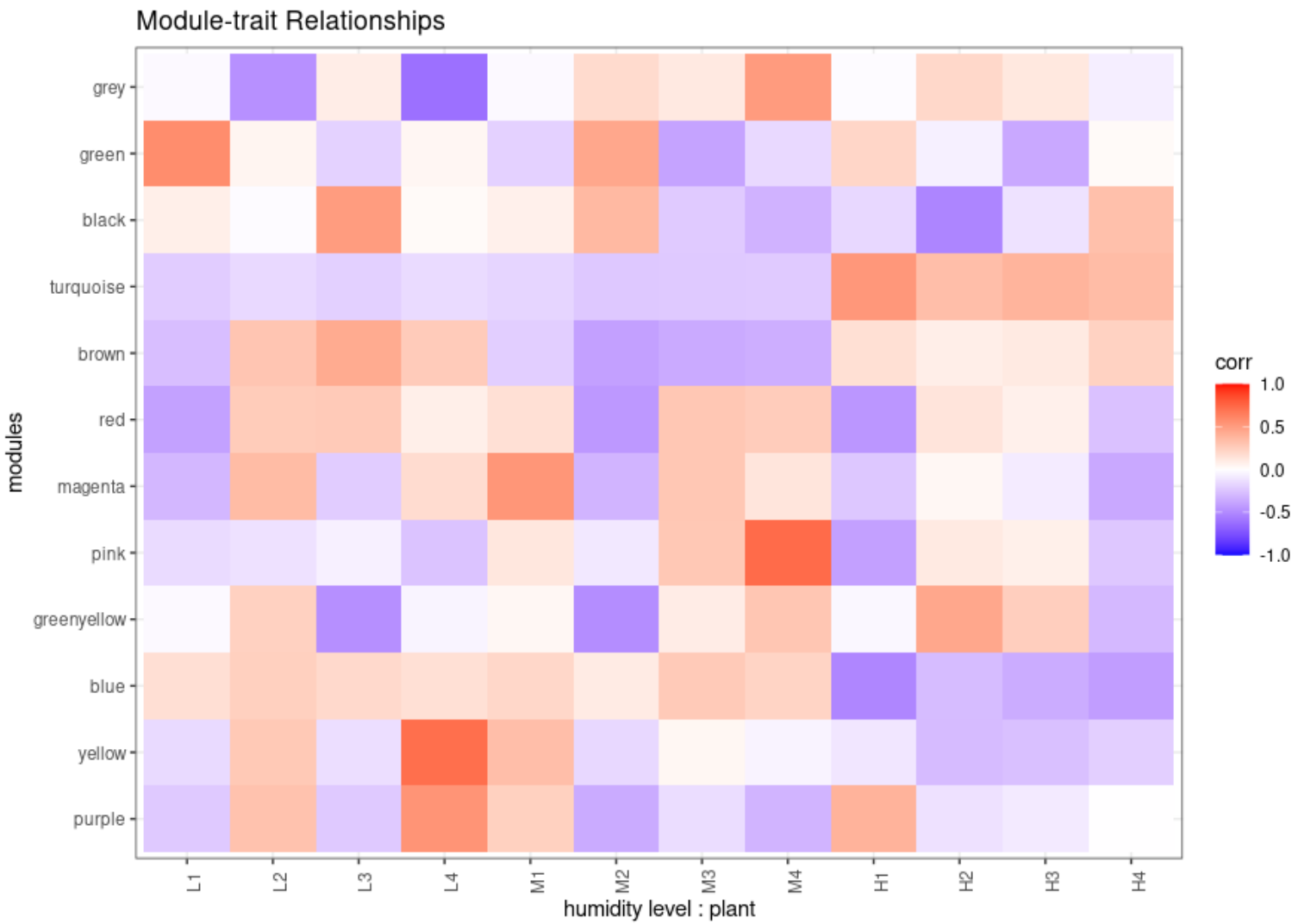
The power of 7 was determined to result in the best fit; applied to the gene-gene correlations used to construct coexpression network



The average connectivity of the resulting network is just under 500 links (edges) per gene (node).

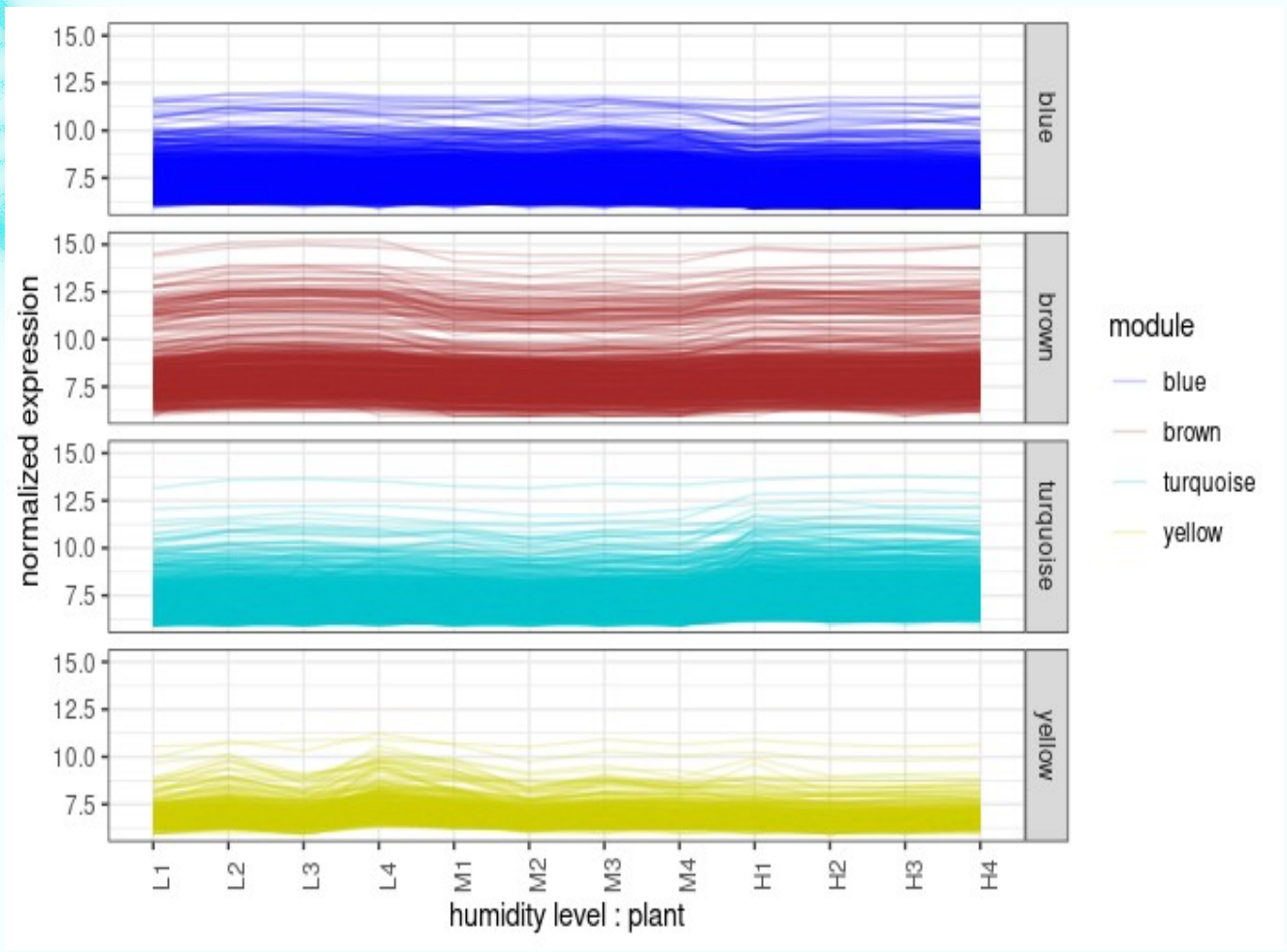


Hierarchical clustering of genes based on correlational network topology. Highly correlated and highly connected genes define 4 clear modules.



Characteristic expression patterns define each module.

Eigen-genes were extracted for scoring similarity of genes to each module and select examples for further research.



Line plot of expression levels of individual with each module of interest visualizes within-group and between-group differences

Top 10 genes most similar to the eigen-gene of each module

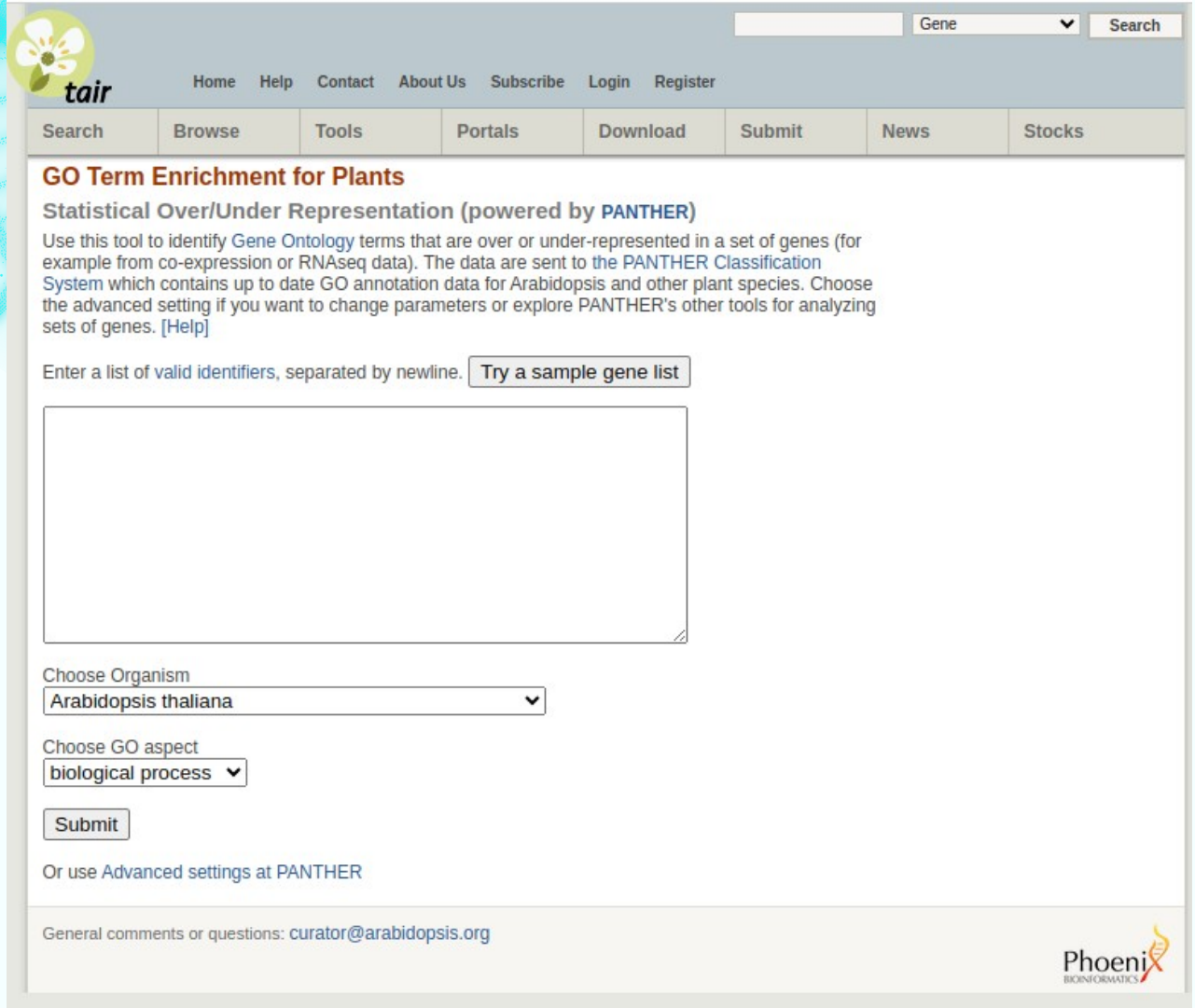
[ ] = Highly DE genes

	Gene_id	module
1	AT1G51140	blue
2	AT1G79520	blue
3	AT2G19860	blue
4	AT2G29970	blue
5	AT1G51440	blue
6	AT5G03190	blue
7	AT5G03240	blue
8	AT5G22290	blue
9	AT4G30960	blue
10	AT1G60590	blue
11	AT3G50820	brown
12	AT3G46780	brown
13	AT1G07320	brown
14	AT3G52960	brown
15	AT3G61470	brown
16	AT5G66570	brown
17	AT1G15140	brown
18	ATMG01390	brown
19	AT2G18020	brown
20	AT1G29910	brown

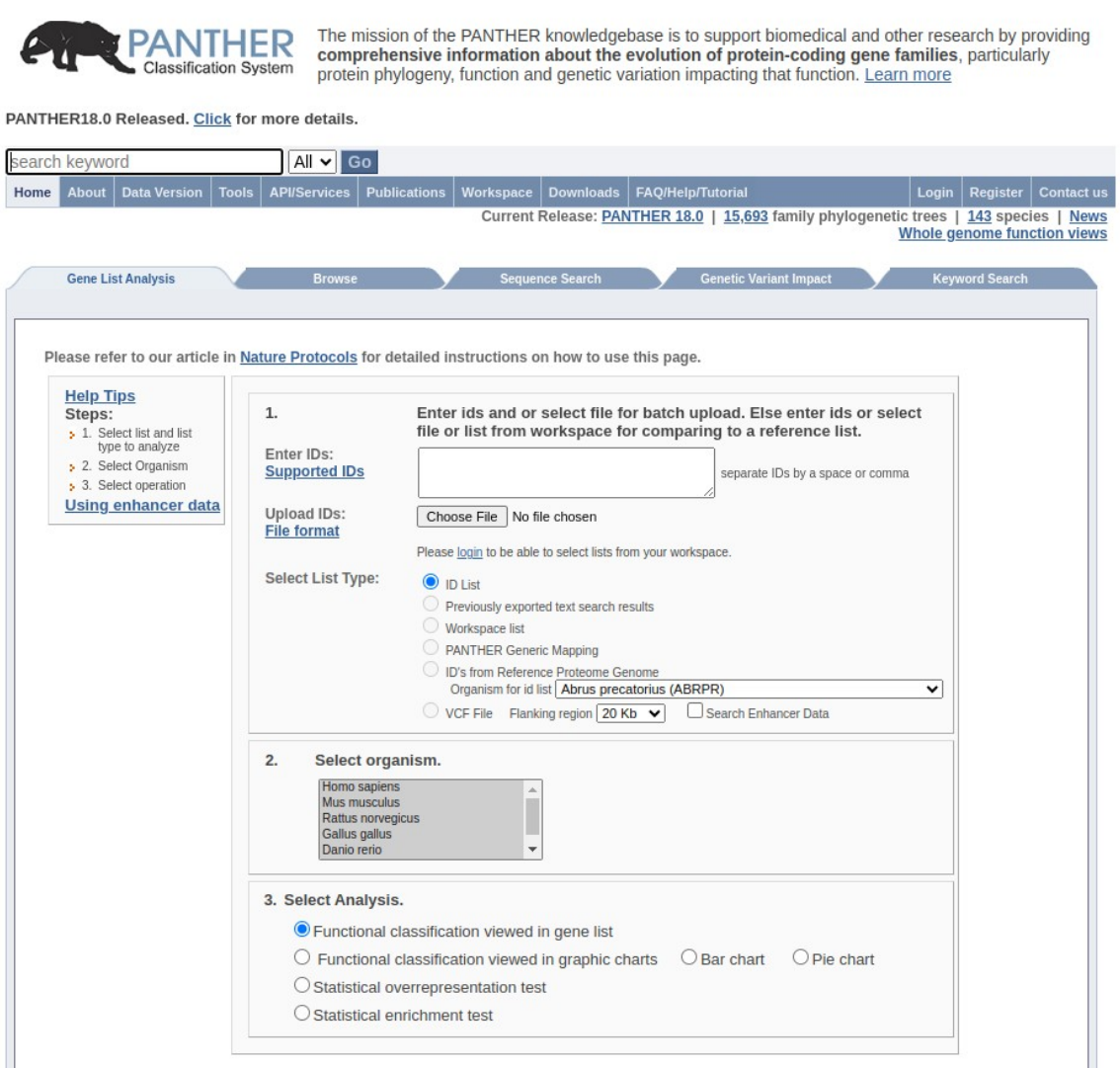
	Gene_id	module
21	AT3G24550	turquoise
22	AT3G51550	turquoise
23	AT1G74380	turquoise
24	AT5G24030	turquoise
25	AT2G01190	turquoise
26	AT1G29690	turquoise
27	AT3G57930	turquoise
28	AT3G19150	turquoise
29	AT4G26690	turquoise
30	AT4G37450	turquoise
31	AT1G19570	yellow
32	AT1G32640	yellow
33	AT3G45140	yellow
34	AT4G27560	yellow
35	AT1G33230	yellow
36	AT4G16760	yellow
37	AT5G05730	yellow
38	AT2G38750	yellow
39	AT4G31500	yellow
40	AT1G19670	yellow

Database Query

TAIR and PANTHER



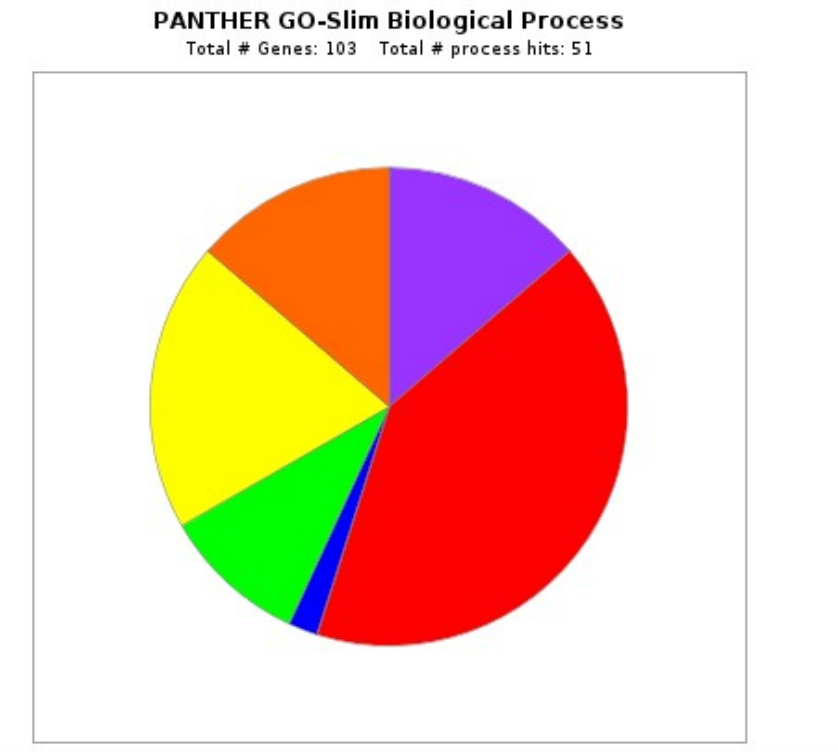
TAIR - The Arabidopsis Information Resource



PANTHER - Protein ANalysis THrough Evolutionary Relationships

PANTHER Database

Protein ANalysis THrough Evolutionary Relationships



Click to get gene list for a category:

- biological regulation (GO:0065007)
- cellular process (GO:0009987)
- developmental process (GO:0032502)
- localization (GO:0051179)
- metabolic process (GO:0008152)
- response to stimulus (GO:0050896)

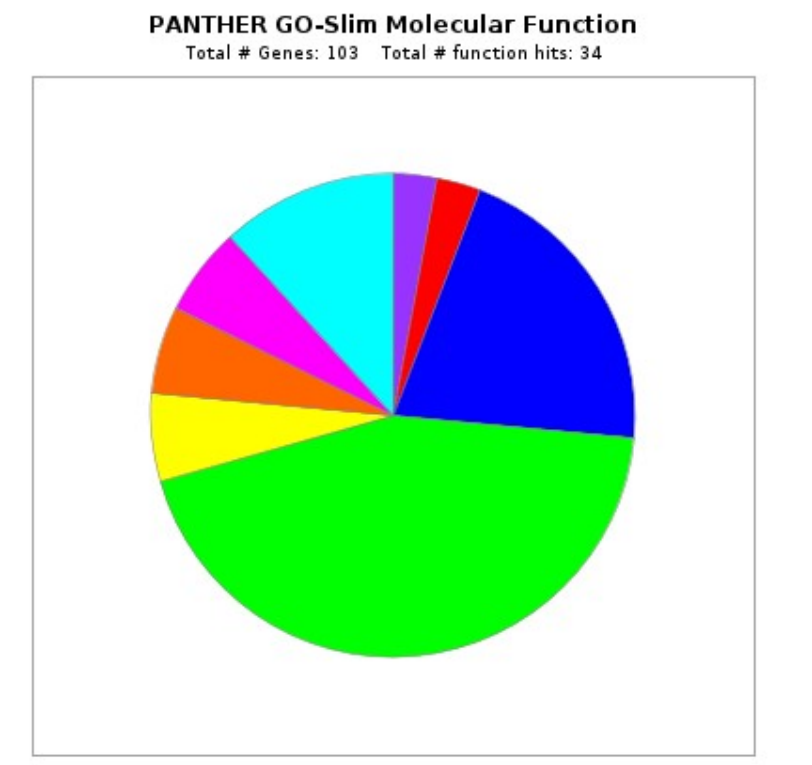
Color picker powered by Web Colors by VisiBone

Cellular Process = transcription factors, secondary carrier transport

Metabolic process = transferase, lyase, oxygenase

PANTHER Database

Protein ANalysis THrough Evolutionary Relationships



Click to get gene list for a category:

- ATP-dependent activity (GO:0140657)
- antioxidant activity (GO:0016209)
- binding (GO:0005488)
- catalytic activity (GO:0003824)
- molecular function regulator activity (GO:0098772)
- molecular transducer activity (GO:0060089)
- transcription regulator activity (GO:0140110)
- transporter activity (GO:0005215)

Color picker powered by Web Colors by VisiBone

Catalytic activity = oxidases, transferases (All eigen-genes from turquoise module are in this category)

Binding = RNA metabolism proteins, membrane traffic, secondary carrier

Transporter = secondary transporter, ATP binding cassette

Interpreting results in the context of plant physiology (Chat GPT 4)

Stomatal Regulation: Plants regulate their water balance through stomata. In high humidity, stomata may remain open for longer periods as the need to conserve water is reduced. This can affect the rate of photosynthesis and transpiration.

Cell Wall Modifications: High humidity can affect cell wall properties. Genes involved in cell wall biosynthesis and modification may be upregulated or downregulated depending on the plant's need to adapt to the humid environment.

Osmotic Adjustments: To maintain cellular turgor and water balance, plants adjust their internal solute concentrations. This may involve the upregulation of genes responsible for the biosynthesis of osmolytes or solutes that help in maintaining cell turgidity.

Defense Responses: High humidity can create conditions favorable for the growth of pathogens like bacteria and fungi. Therefore, genes associated with pathogen recognition and the production of antimicrobial compounds might be expressed more under these conditions.

Hormone Signaling: Hormones like abscisic acid and gibberellins play a crucial role in plant growth and stress responses. The expression of genes involved in the synthesis and signaling pathways of these hormones might be upregulated or downregulated in response to the humid environment.

Metabolic Changes: Elevated humidity can influence the plant's overall metabolic processes, including energy production and utilization. Genes involved in various metabolic pathways might show altered expression to optimize energy use and resource allocation under these conditions.

Molecular Chaperones and Heat Shock Proteins: High humidity can cause stress at a cellular level, leading to protein misfolding. Genes encoding molecular chaperones and heat shock proteins, which help in proper protein folding and stress recovery, might be upregulated.

Secondary Metabolites: Plants might produce more secondary metabolites under stress conditions, including high humidity. The biosynthesis of these compounds could be more active in response to humid conditions.

CONCLUSION

- 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.

Questions?

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. https://www.thoughtco.com/plant-stomata-function-4126012
8. Lopez, Marlen & Croxford, Ben & Rubio, Ramón & Martín, Santiago & Jackson, Richard. (2015). Active materials for adaptive architectural envelopes based on plant