CSB352 Lab Report 2, Brandon Kim

Introduction:

In this lab report, *Arabidopsis thaliana,* a eudicot species deriving from the Brassicaceae family that is a notorious model organism due to its small, completely sequenced genome will be investigated. Due to the previous research done on the past lab report, it was deduced that the 7 gene was a highly conserved gene across taxa, fulfilling the autophagy process role, and allowing for cellular health maintenance. The At5g66930 gene will be further examined to facilitate higher levels of understanding of the characterization of this autophagy function using various methods such as the analysis of the protein domains, observing the protein-protein interactions[5], looking at the gene expression patterns, and performing a coexpression analysis; comparing it against other expressions and deducing an educated answer from there.

Methods:

To begin, the protein domain must be retrieved and analyzed. The nucleotide sequence was retrieved from the NCBI BLAST[7] website in FASTA format, converted to a protein sequence using EMBOSS Transeq[6], and inputted into the InterProScan[4] with default search parameters. 6 separate results will be generated, one for each of the 6 partial sequences. The most notable one to observe is the fourth sequence where you can examine the cellular processes, albeit 2 of them. Default parameters were chosen to maximize the amount of results generated. Secondly, protein-protein interactions data was retrieved from the BioGRID[8] search engine, searching with the AGI ID of At5g66930, and under the organism search of *Arabidopsis thaliana (Columbia)*. The results generated show 2 interactors; RAPTOR1, and TGA1 under the GO component tab. Interactors of interactors were attempted to be found using the search features such as RAPTOR1 or At5g66930, however, no further results were found. Other tools such as

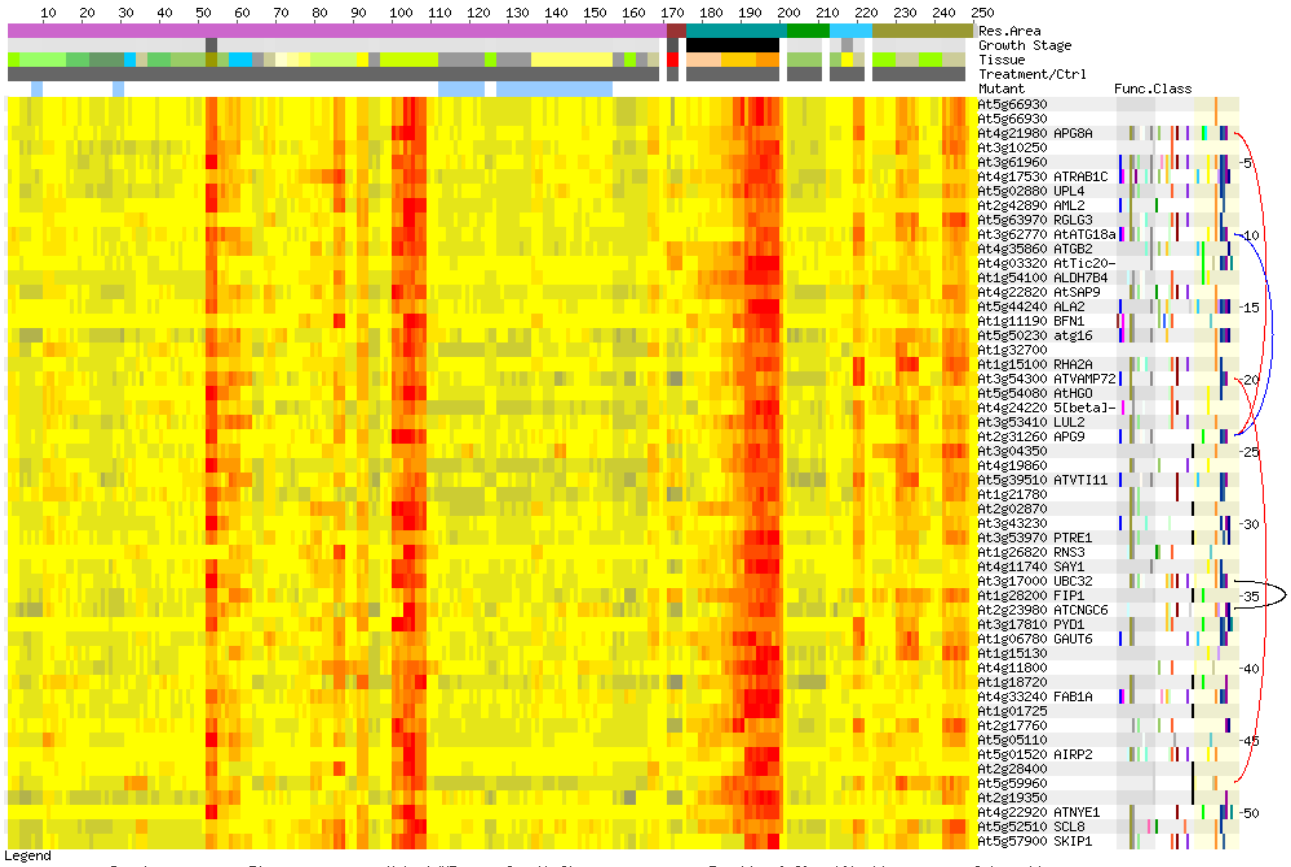
the Arabidopsis Interactions Viewer[1] were used to attempt to retrieve more protein-protein interaction results, however, the same two results showed. Due to this, these protein-protein interactors were unable to be analyzed using AgriGO[9] due to the lack of results. Despite this, protein function prediction can still be identified and characterized deeper using the function and processes of RAPTOR1 and TGA1.

After this, it is important to investigate coexpressions by performing a coexpression analysis against the AtGenExpress Tissue Compendium expression database and doing a GO enrichment analysis on this. The coexpression analysis can be done on the BAR Expression Angler tool[10] by inputting the AGI ID of At5g66930, returning the top 50 hits instead of 25, and searching against the AtGenExpress Tissue Compendium expression database. This expression database was performed against because the Tissue Compendium most closely relates to our gene at hand. After all, its function is for proteins and we search against the AtGenExpress database because it is a database with a reliably large sample set. With these 50 results, they can be viewed by formatted data set after median centering and normalization to observe the heatmap of the gene expression levels (Input a value of 0.25 for the MAX to increase the visibility of low expressing genes). Following this, we would like to further observe these 50 genes using a GO enrichment analysis using AgriGO[9]. Navigating to AgriGO[9], we can retrieve the analysis by using the Singular Enrichment Analysis tool, with the species of *Arabidopsis thaliana*, inputting the 50 genes in the query list, referencing against Affymetrix ATH1 Genome Array (blast), changing the statistical test method to Hypergeometric, keeping the significance level at default 0.05, and decreasing the minimum number of mapping entries to 3 to promote more, but still quality results. This Affymetrix reference was used over the others because it allows the comparison against genes that are only present on the ATH1 Genome Array. The hypergeometric

statistical test was the best fit here because it allows for observation of the probability of successes without replacement. Next, to solidify and return more potential predictive results, the gene ID At5g66930 was inputted into the Integrative online prediction program; GeneMANIA[11] to output other potential interactors along with the RAPTOR1 and TGA1 listed beforehand.

Results and Discussions:

Beginning with InterProScan[4] protein domain analysis, the lack of results (2) may indicate the lack of information documented in the database they search the protein sequences against. To address the 2 results that showed; the 2 biological processes in the fourth sequence of autophagy, and autophagosome assembly, it can be noted that these 2 processes are done in the phagophore assembly site, an important note to keep in mind that will be touched on further in the report. Moving on from the protein domain analysis, the protein-protein interactions were examined using the BioGRID[8] tool, resulting in two protein interactors. RAPTOR1 (AT3G08850) and TGA1 (AT5G65210). Both have high throughput, meaning the experiments involved interacting with over 100 in their respective methods of Affinity Capture-MS, and Two-hybrid; both indicative of a party hub protein-protein interaction network which can mean that the protein interactors are highly related in expression in their specific locations with one another. Now for coexpression analysis to help with identifying some notable gene expression patterns. After using the expression angler[10] to fulfill this, viewing the heatmap as stated in the methods section reveals a great deal of information on the function, processes, components, and their connection altogether.



Legend

+ −
0.00
0.03
0.05
0.08
0.10
0.12
0.15
0.17
0.20
0.22
>0.25

Res.Area
CAGE
MS
Dev
MedWT
PhCh
Poll
Sd+Sq

Tissue
1stNd
CaulLf
Cotly
FlwRos
Flw
Flw10/11
Flw12
Flw15
Flw9
Crpl.F12
Petl.F12
Sepl.F12
Stmn.F12
Crpl.F15
Ped.F15
Petl.F15
Sepl.F15
Stmn.F15
Hyptl
Leaf
Lf1-2
Lf7.Dist
Lf7.Petiol
Lf7.Prox
Mat.Poll
Root
RosLf10
RosLf12
RosLf2
RosLf4
RosLf6
RosLf8
Sdlng.grn
Seed10-Sil
Seed3+Sil
Seed4+Sil
Seed5+Sil
Seed6-Sil
Seed7-Sil
Seed8-Sil
Seed9-Sil
SenLf
ShoApx.Infl
ShoApx.Tran
ShoApx.Veg
ShoApx.Veg+Lf
2ndInterNd
VegRos

Mutant/WT
Mutant
WT

Trtmnt/Ctrl
Treatment
Control

Growth Stage
Div
1.0
1.02
1.03
1.04
1.09
1.10
1.11
1.12
3.70
6.10
6.50
9.70

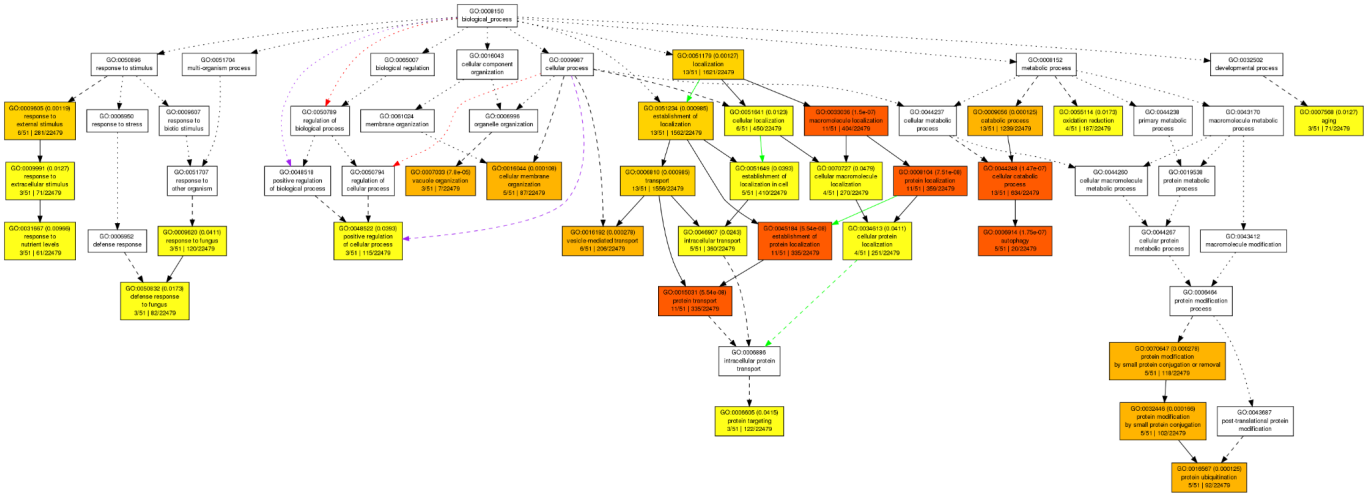
Functional Classification
Dna or rna metabolism
Cell organization and bio...
Developmental processes
Electron transport or ene...
Other biological processes
Other cellular processes
Other metabolic processes
Pollination
Protein metabolism
Response to abiotic or bi...
Response to stress
Signal transduction
Transcription,dna-dependent
Transport
Unknown biological processes
Dna or rna binding
Hydrolase activity
Kinase activity
Nucleic acid binding
Nucleotide binding
Other binding
Other enzyme activity
Other molecular functions
Protein binding
Receptor binding or activity
Structural molecule activity
Transcription factor acti...
Transferase activity
Transporter activity
Unknown molecular functions
Er
Golgi apparatus
Cell wall
Chloroplast
Cytoskeleton
Cytosol
Extracellular
Mitochondria
Nucleus
Other cellular components
Other cytoplasmic components
Other intracellular compo...
Other membranes
Plasma membrane
Plastid
Ribosome
Unknown cellular components

Interactions
Low Interolog CV
Med. Interolog CV
High Interolog CV
PubMed support

(Figure 1: Heatmap generated from the Expression Angler[10] that is formatted after median centering and normalization with a MAX value of 0.25 to increase visibility. The coloured bars at the top represent the different tissue types, which can be referenced to with the legend at the bottom of the figure for each tissue type and their corresponding colour. The gene IDs are listed on the right, and the heatmap can be read by knowing that the red colour is highly expressed, and the yellow colour is lowly expressed, with everything in the middle being a gradient).

The prominent vertical red strips indicate high expression levels of genes across that specific region. Note that there are 3 prominent vertical red strips; from left to right; one at the senescing leaf tissue column, one at the flowers stage 15 petals and stamens column, and one at the seeds stages 8-10 without siliques. These 3 tissue types show that there is a high expression here, and will help with characterizing this gene's function and process.

Next, a GO enrichment analysis was done on this coexpression analysis and returned a gene ontology.

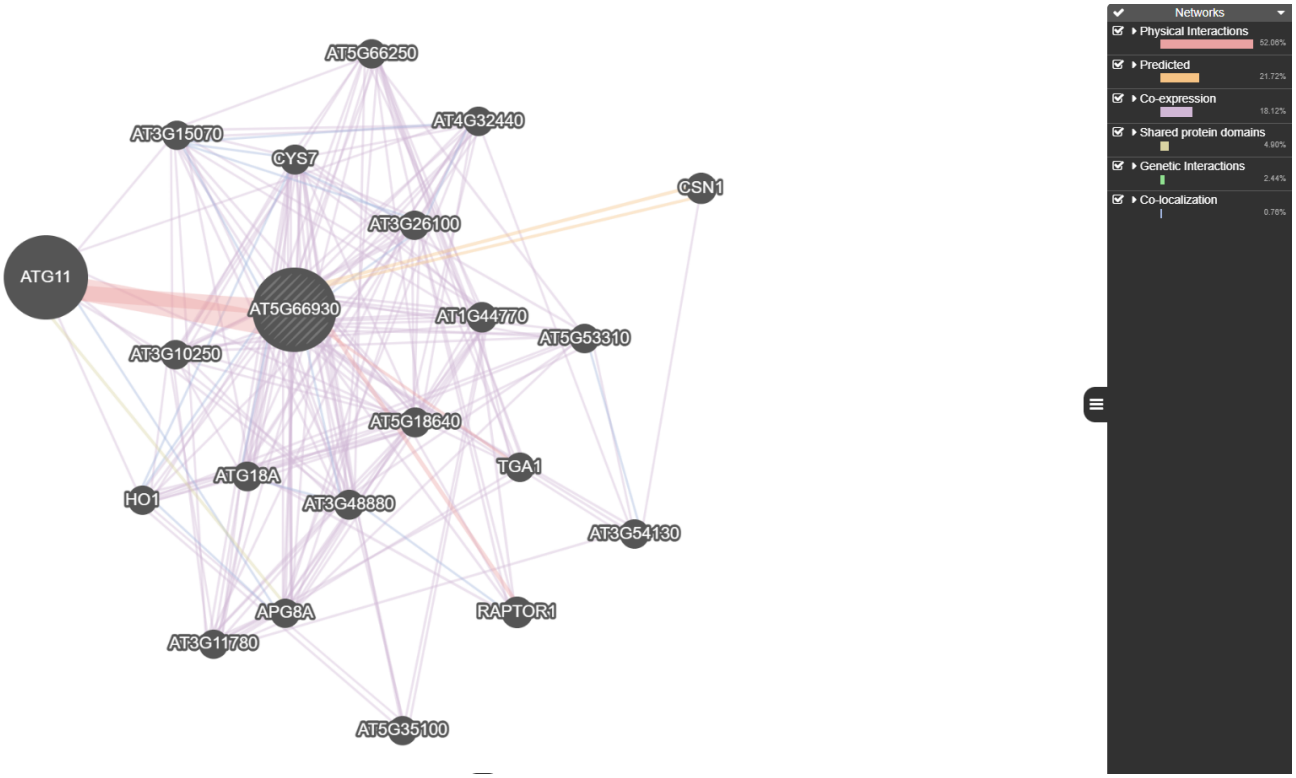


(Figure 2: A gene ontology enrichment analysis done on the top 50 resulting genes from the AtGenExpress database by AgriGO[9]. The red colour indicates highly enriched biological

processes and white/yellow indicates low enriched biological processes. Between red and yellow is a gradient respective to the magnitude).

In this enrichment analysis, there are 6 highly enriched biological processes according to Figure 2. From left to right in the figure, these are the highest enriched biological processes for At5g66930; **protein transport, the establishment of protein localization, macromolecule localization, protein localization, cellular catabolic processes, and autophagy.** Now, this ontology analysis reveals the processes that are most enriched, or rather, favoured. As stated in the introduction, we researched that autophagy was one of the main processes of the At5g66930 gene; however, more connections using coexpression analysis reveal more related processes, most related to localization or transport. It can also be noted that the ontology structure shows these highly enriched processes are under the parent process of localization; a process in which signal receptors recognize proteins for recruitment[3].This is interesting because, in the previous lab report, it was deduced that the function is autophagy, which it still is, however further characterizing it reveals the role of the gene in the biological processes; most notably the protein transport and protein localization. In addition, it is involved in the cellular catabolic process of chemically breaking down substances for energy use[2]. Therefore, it can be concluded that the At5g66930 gene is highly involved in the process of autophagy, protein transport, protein localization, and the cellular catabolic process; investigated through the coexpression analysis pathways.

Afterwards, GeneMANIA[11], a predictive integrative tool that shows the functions of a certain gene was used to display potential functions and interactors.



(Figure 3: GeneMANIA[11] output web of the various confirmed networks in addition to predicted networks based on other databases. Red lines indicate physical interactions, orange indicate predicted interactions, purple indicates coexpression, yellow indicates shared protein domains, green indicates genetic interactions, and blue indicates co-localization. The web shows the various connections and how they are related).

In this GeneMANIA[11] output, it can be seen that the physical interactions include the ones outputted from BioGRID[8], solidifying the RAPTOR1 interactor and the TGA1 interactor, but also included the ATG11 interactor, an autophagy-related protein, further proving the function of the autophagy process. The web also shows the coexpressed, as referenced before, as well as the interesting predicted CSN1, a CRISPR-associated gene. Seemingly unrelated, but could be

related because of the CSN1 nature of protein-coding and the function of AT5G66930 being protein localization and recognition.

These *in silico* experiments show very informative results however, it is important to note the importance of both *in silico* and *in vivo* experiments to arrive at a conclusion. A potential *in vivo* activity that can add further to the conclusion of the function of AT5G66930 could be a gene knockout of AT5G66930, investigating the lack of a function after the knockout; however, the problem with this is the importance of the gene's hypothesized functions. The species may die after the gene knockout, not concluding anything. Therefore, there are 2 proposed *in vivo* experiments. Firstly, gene silencing can be done to *reduce* the expression levels of the functions to investigate the change. Secondly, microarray analysis can be done to measure the levels of gene expression in different conditions. For the first experiment, it can be hypothesized that the function of autophagy, protein transport, protein localization, and the cellular catabolic process will all be suppressed and decreased in expression after the silencing, which indicates a correct *in silico* conclusion. In the second experiment, it can be hypothesized that the gene expression levels are high in the autophagy, protein transport, protein localization, and the cellular catabolic process function locations of the gene, and less for the others.

References

1. Dong, S. *et al.* Proteome-wide, structure-based prediction of protein-protein interactions/new molecular interactions viewer. *Plant Physiology* 179, 1893–1907 (2019).
2. Dröge, W. Redox regulation in anabolic and Catabolic Processes. *Current Opinion in Clinical Nutrition and Metabolic Care* 9, 190–195 (2006).
3. Hung, M.-C. & Link, W. Protein localization in disease and therapy. *Journal of Cell Science* 124, 3381–3392 (2011).
4. Jones, P. *et al.* InterProScan 5: Genome-scale protein function classification. *Bioinformatics* 30, 1236–1240 (2014).
5. Li, F., Chung, T. & Vierstra, R. D. Autophagy-Related11 plays a critical role in general autophagy- and senescence-induced mitophagy in *arabidopsis*. *The Plant Cell* 26, 788–807 (2014).
6. Li, W. *et al.* The EMBL-Ebi Bioinformatics Web and programmatic tools framework. *Nucleic Acids Research* 43, (2015).
7. McGinnis, S. & Madden, T. L. Blast: At the core of a powerful and diverse set of Sequence Analysis Tools. *Nucleic Acids Research* 32, (2004).
8. Oughtred, R. *et al.* The biogrid database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science* 30, 187–200 (2020).
9. Tian, T. *et al.* Agrigo v2.0: A go analysis toolkit for the Agricultural Community, 2017 update. *Nucleic Acids Research* 45, (2017).
10. Waese, J. *et al.* EPlant: Visualizing and exploring multiple levels of data for hypothesis generation in plant biology. *The Plant Cell* 29, 1806–1821 (2017).

11. Warde-Farley, D. *et al.* The genemania prediction server: Biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research* 38, (2010).