Biology-4415: Deeper Understanding of Omics: Proteomics

Gavin Ieong

2023-04-06

Abstract

A previously published set of tobacco lines showed that a gene called Daucus carota (carrot) DcLCYB1 increased carotenoid content and photosynthetic efficiency. Carotenoids have many functions relating to plant physiology such as photosynthesis and photoprotection, pollination, and others.

Many approaches have been used to increase Β-carotene content by expressing the lycopene β-cyclase (LCYB) gene which codes for the enzyme (LCYB). Previous experiments showed that increased expression of LCYB provided beneficial traits such as better tolerance for abiotic stress (salt and drought), high light, and UV irradiation.[7]

However, as carotenoid content increases, it was accompanied by further induction of other genes, suggesting that there are many other factors controlling carotenoid production.

Due to these benefits and importance of these functions, a combined multi-omics approach involving transcriptomics, proteomics, and metabolomics, was used to study the effects of LCYB expression on tobacco genes and expression.

In these experiments, their own T4 generation of DcLCYB1 expressing tobacco as transgenic lines (Includes L14, L15, L16) was used and alongside it, Tabacco [Nicotiana tabacum cultivar Xanthi] was used as the wild type.

For each one, the technical Replicates is 3 (grown three times in three different petri dishes). Furthermore, there are three biological replicates per line.

On the transcriptomics level, they discovered all three transgenic lines (14,15,16) showed higher height than wild type in its phenotype, many of the Carotenoid related

genes were upregulated, and most Differentially Expressed Genes (DEG) were found in nucleus, cytosol, and chloroplast.

On the proteomics level, the PCA plot showed transgenic replicates grouped together, and wild type genotype grouped together (suggested considerable change in proteome), and similarly, the highest number of proteins showed up at chloroplast, cytoplasm, and mitochondria. Finally, they discovered 260 proteins had increase abundance while 149 Proteins had decrease abundance and further revealed abiotic stress related proteins were mainly upregulated in the transgenic lines.

Lastly, on the metabolomics level, they identified many changes in secondary metabolites of transgenic lines, and many other metabolites were increased or decreased in those lines and while the lipid composition varied between lines, they still showed significant difference to wild type line.

To further study these findings and the importance of the proteomic level of analysis, three additional analyzes were added to the original paper. This included visualizing existing plots and data in different approaches to highlight data, showcasing information that the paper was lacking or omitted, and integrating the findings of proteins and genes to further enhance connection between significance of proteomic level with the transcriptomic level.

Introduction

The science and study of "omics" can be seen as the studies of processes in totality. The emergence of high-throughput technologies allows us to study many "omics", and this encompasses major branches such as transcriptomics, proteomics and metabolomics. For instance, transcriptomics is the study of all transcripts created within an organism, and similarly proteomics is the study of the totality of the proteome under a given time and condition, and lastly metabolomics covers the study of all metabolites from within an organism.[1] Although the branches may differ, they are still intertwined together in many ways as we move from the basis of genes to the production of proteins, and to the metabolites that follow.[2] As trends noticed on the transcriptomics level might not correlate on the proteome level, a deeper look at the proteomic data is needed. To do this, existing data about the proteome will be visualized, and a focus on the intensity of the proteins will be further analyzed.

Methodology

Program Used:

- R version 4.2.2 (2023-04-02)
- StringDB (https://string-db.org/) (2023-04-02)

Libraries Used:

- Pheatmap[3]
- ggplot2[4]
- corrr[5]
- ggcorrplot[6]
- FactoMineR[7]
- factoextra[8]
- ggfortify[9]
- ggpubr[10]

Supplemental Tables Used:

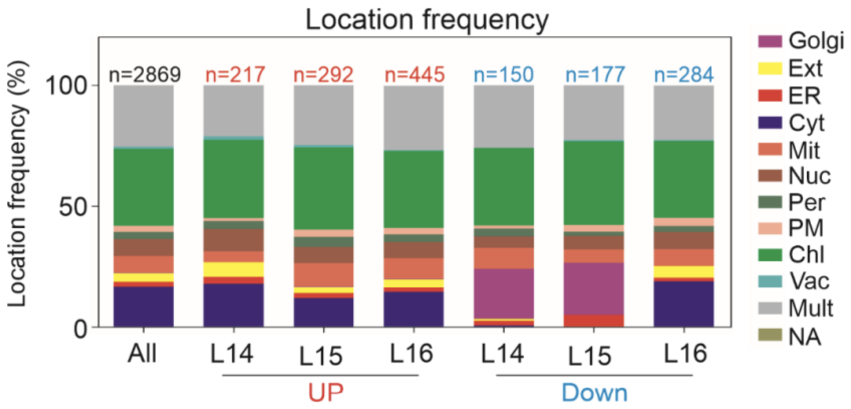
- Supplemental Table 10
- Supplemental Table 9

Results

Stacked Bar Graph & Bubble Plots:

Bubble plot using Supp. Table 10, with the number of significantly down regulated and up regulated proteins in each compartment.

(Total n = 2869, L14 Up = 217, L15 Up = 292, L16 UP = 445, L14 Down = 150, L15 Down = 117, L16 do



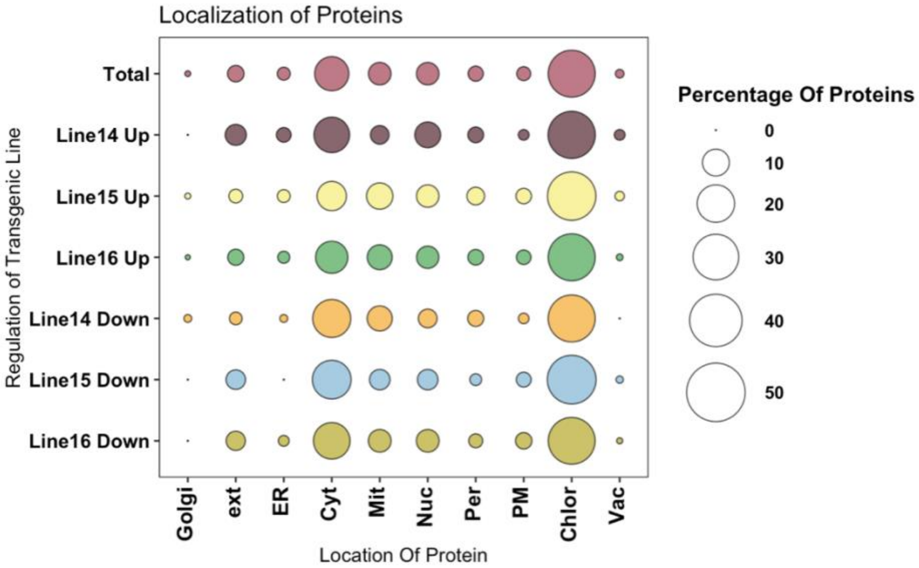


Figure 1. Original Stacked Bar Graph (top) and New Bubble Plot of Protein Localization (bottom).

PCA (Principal Component Analysis):

A PCA was done using the raw intensities of each of the lines (3 replicates each), and n =2869, and plotted using R.

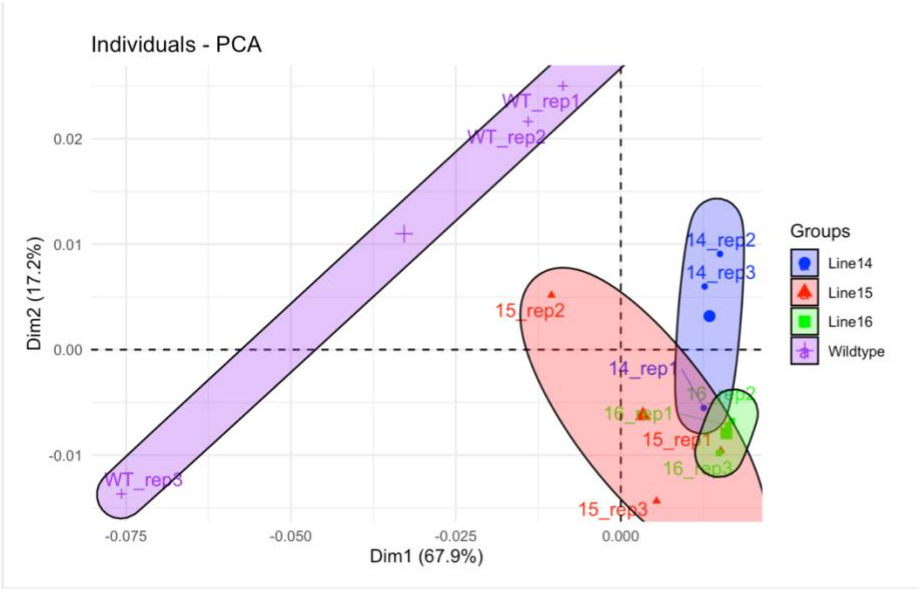


Figure 2. PCA Plot of Proteins with wild type, L14, L15. L16, each with 3 replicates n = 2869.

Scatterplot & Histogram:

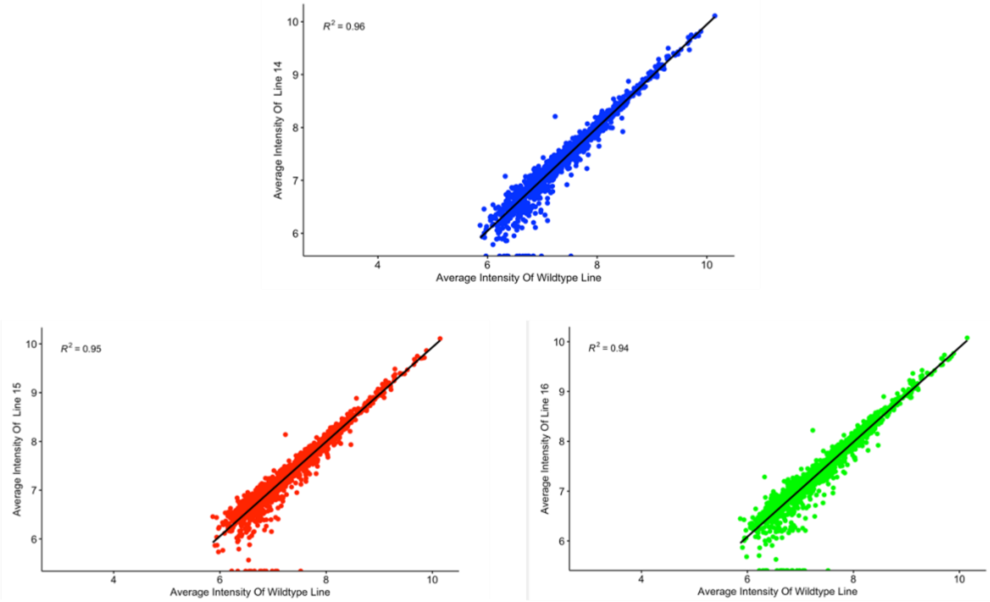


Figure 3. Scatterplot of the average of the 3 raw intensity data from each transgenic line n = 2869.

A distribution was plotted using the log 2-fold change for each transgenic line compared to the wildtype. The raw distribution shows changes compared to wildtype, and comparing the distribution shapes between each line showcases the similarities and differences between them. The distribution reveals L15 and L16 are more similar than L14 based on shape.

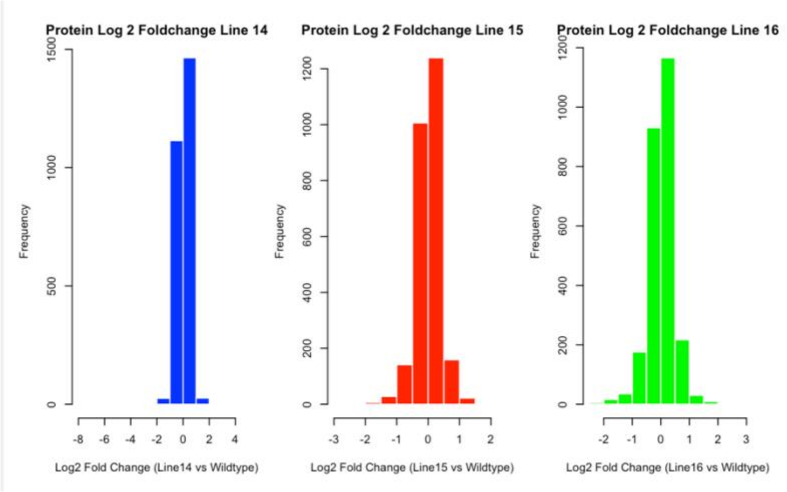


Figure 4. Histogram of intensity distribution after log2 foldchange for each transgenic line n =2869.

Heat Map:

Using the log 2-fold change for each transgenic line but using a different visualization technique in the form of a heat map. Shows that L15 and L16 are more similar compared to L14 because the clusters in middle and right show similar colour scheme, but the cluster on left has opposing colours, indicating differences. This is further supported by the dendrogram on the top of the plot, as L14 is outgroup while L15 and L16 are grouped together.

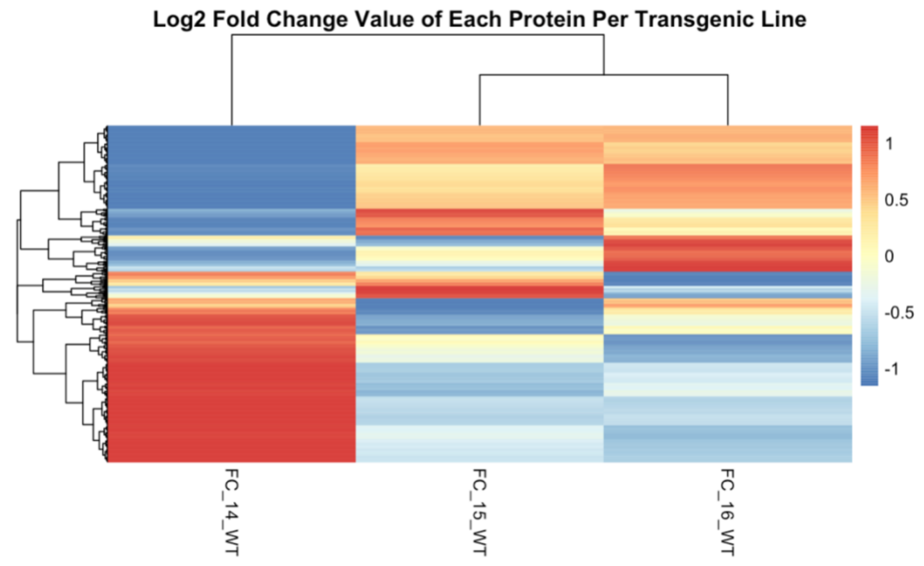


Figure 5. Heatmap of intensity distribution after log2 foldchange for each transgenic line n = 2869.

Network Analysis:

Overlapping genes and genes encoded for proteins as output, with the compartments used in the bubble plot (Chloroplast, Cytosol, ER, Golgi, Mitochodria, Nucleus, Peroxisome, Plasma Membrane, Vacuole) as proteins of interest. Highest Confidence level (0.9) was used.

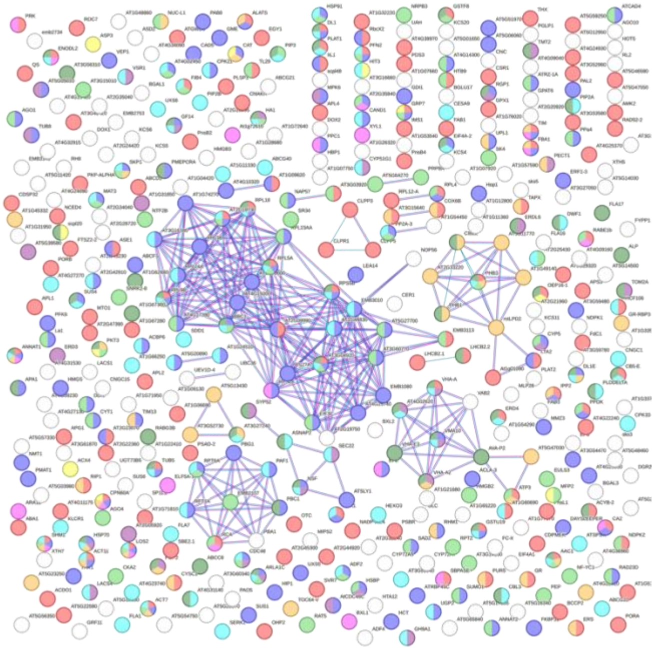


Figure 6. Cluster network of the overlap of gene names on transcriptomics level, and proteins on proteomics level. (464 Overlaps)

Same Protein Cluster Network as above, but with single nodes removed.

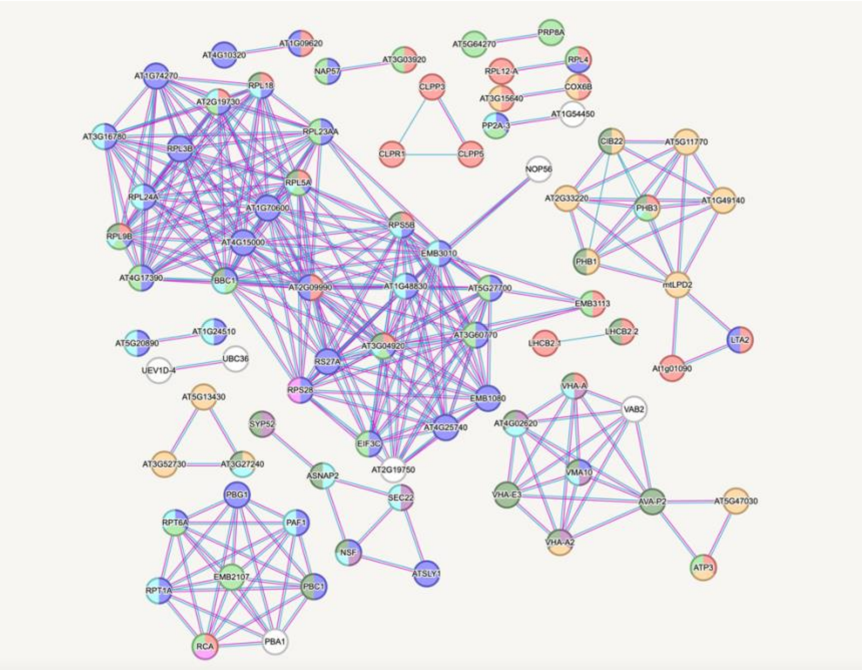


Figure 7. Cluster network of the overlap of gene names on transcriptomics level, and proteins on proteomics level, with single nodes removed.

Discussions

The main goal was to re-visualize the existing data and plots to make the information more readily understandable. The bubble plot accomplishes this goal by overcoming the flaws in the original stacked bar graph. The stacked bar graph (Figure 1) had problems with the colour scheme (making it hard to separate compartments) and it had dis-leveled bars which make it hard to compare across samples. With the new bubble plot, all the values are in alignment, and each with a different colour scheme. Furthermore, even the smaller values can be represented as a bubble and seen, unlike the stacked bar graph.

To study the intensity data more closely, the information was checked and verified with the scatterplot and PCA plot. The scatterplot showed no obvious outliers or problems with the data, while the PCA results matched the given ones from the paper. Furthermore, both dimensional axes of the generated PCA showed a higher percentage, indicating more of the original information was retained. The combination of the histogram and heatmap helped showcase and highlight in different ways the differences between L14 and L15 and L16. Finally, by using the overlapping genes of Line 14 found in table 4, and table 9, and stringDB, a protein cluster network was generated. From there, the compartments (chloroplast, cytosol, mitochondria, nucleus, extracellular, peroxisome, vacuole, endoplasmic reticulum, plasma membrane, and

Golgi body) mentioned in Figure 1 were selected to showcase which proteins were part of the most localized areas. This also showcased 4 distinct clusters once single nodes were removed.

Future Prospects

In the future, more improvements could be made to analyze the output cluster network. Although there are four distinct clusters with the protein names, neither function nor what greater system it is part of is shown.

Similarly, better techniques which make use of p-value and fold change could have been used to display the changes in upregulated and down regulated proteins.

Alternatively, future studies can delve into the metabolomics level, to see if these findings for L14 (differently expressed compared to L15 and L16) hold true.

Conclusions

With the addition of new visualization techniques, existing data can be made to be much easier to understand and analyze. This can play a key part in spreading this information to a broader audience and was accomplished by the new visualization techniques used such as the bubble plot. Additionally, with the use of 6 different plots, a significant difference was observed in the transgenic line 14. Based on the results of my own PCA plot, there is a significant difference in the proteome between the wildtype and the three transgenic lines (L14, L15, L16). The wild type is furthest apart from the transgenic lines, while the three transgenic lines are grouped relatively closer together. Within the transgenic lines, the PCA shows L15 and L16 being closer together than L14. This is further supported by the distribution of the log 2-fold change distribution of protein intensities within each transgenic line, as the distribution of L15 and L16 are much more similar compared to the distribution of L14. After using the same log 2-fold change values and inputting it into a heatmap, this result was confirmed via the dendrogram and clustering of the heat map.

Citations

1. Bharagava, R. N., Purchase, D., Saxena, G. & Mulla, S. I. Applications of metagenomics in microbial bioremediation of pollutants. *Microbial Diversity in the Genomic Era* 459–477 (2019). doi:10.1016/b978-0-12-814849-5.00026-5
2. Moreno, J.C., Martinez-Jaime, S., Kosmacz, M., Sokolowska, E.M., Schulz, P., Fischer, A., Luzarowska, U., Havaux, M., and Skirycz, A. (2021). *A Multi-OMICs Approach Sheds Light on the Higher Yield Phenotype and Enhanced Abiotic Stress Tolerance in Tobacco Lines Expressing the Carrot lycopene β-cyclase1 Gene*. Front. Plant Sci. 12, 624365. 10.3389/fpls.2021.624365.
3. Kolde R (2019). _pheatmap: Pretty Heatmaps_. R package version 1.0.12, <https://CRAN.R-project.org/package=pheatmap>.H. Wickham.
4. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
5. Kuhn M, Jackson S, Cimentada J (2022). _corrr: Correlations in R_. R package version 0.4.4, <https://CRAN.R-project.org/package=corrr>.
6. Kassambara A (2022). _ggcorrplot: Visualization of a Correlation Matrix using 'ggplot2'_. R package version 0.1.4, <https://CRAN.R-project.org/package=ggcorrplot>.
7. Sebastien Le, Julie Josse, Francois Husson (2008). FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software, 25(1), 1-18. 10.18637/jss.v025.i01
8. Kassambara A, Mundt F (2020). _factoextra: Extract and Visualize the Results of Multivariate Data Analyses_. R package version 1.0.7.999, <http://www.sthda.com/english/rpkgs/factoextra>.

9. Yuan Tang, Masaaki Horikoshi, and Wenxuan Li. "ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages." The R Journal 8.2 (2016): 478-489.

10. Kassambara A (2023). _ggpubr: 'ggplot2' Based Publication Ready Plots_. R package version 0.6.0, <https://CRAN.R-project.org/package=ggpubr>.