DataVis Process Book

Sarah Weintraub, Aidan Pecorale

May 2, 2022

1 Overview and Motivation

Network visualizations are notoriously difficult to parse and extract relevant biological information from multi-omics data sets. By creating a multivariate network will allow researchers to sort not only by annotation, but by clustered trends as well. Currently, this type of search is done with two excel sheets and manually search through clusters of interest and then switching sheets to that cluster and filtering through by annotation of interest. This visualization will provide ease of access to search the entire transcriptome for annotation as well as creating easier access per cluster.

2 Related Work

- Analysis of Network Clustering Algorithms and Cluster Quality Metrics at Scale [2].
- GraphOmics: an interactive platform to explore and integrate multi-omics data [17].
- The State of the Art in Visualizing Multivariate Networks [10].
- The State of the Art in Multilayer Network Visualization [8].
- Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Network [14]
- Tasks, techniques, and tools for genomic data visualization [11]
- NodeTrix: a hybrid visualization of social networks [4]
- Towards Rigorously Designed Preference Visualizations for Group Decision Making [5]
- Matrixwave: Visual comparison of event sequence data [18]

- Kepler's tally of planets [1]
- Caleydo [6]
- How the recession reshaped the economy, in 255 charts [12]
- Treevis.net [13]
- A Survey of Multi-faceted Graph Visualization. [3]
- Attribute signatures: Dynamic visual summaries for analyzing multivariate geographical data [16]
- Ordered tree map layouts [15]
- A nested model for visualization design and validation [9]
- WGCNA: an R package for weighted correlation network analysis [7]

3 Questions

We were interested in creating a tool for researchers to use for large data sets of genomics data. Specifically, we wanted this tool to allow for genomics comparison of expression trends.

4 Data

One major consideration we noted was data formatting and reproducibility. We wanted everything to be as simple as possible for the user to upload a limited number of files in the format that is in a standardized structure. This required us to implement a few cleanup and reformatting within the tool.

We performed the following formatting adjustments:

1. .csv to .json
2. data structure: wide to long

Initially we were working on the entire transcriptome, containing almost 6000 samples, and were clustered into about 12 groups, but after some literature review, we decided it would be best to filter based on fold change, keeping only the most relevant genes which ended up giving us about 3000 samples and only three clusters plus one, "grey" module, that contained genes without strong correlations to any cluster. We added a filter to remove the "grey" module from the heatmap visualization as it only had one gene, however it is still in our data set as we did not want to remove it for all future data sets.

5 Exploratory Data Analysis

We used PCAs, correlation matrices, and dendrograms to initially look at the data. This is what prompted the need for a better way to visualize this data as the correlation matrix and dendrograms were too large to gather any real insight from.

Because of this, we decided we needed a way to easily compare different modules to gain genetic insight. We knew we wanted to create a way to look at overall trends in each cluster so we plotted some heatmaps to see general trends.

We also wanted to add a way to see relationship between the clusters, so for this we used a network to show that relationship. Because our clusters were large (over 1000 genes), we related our modules using eigengenes, which is one of a set of right singular vectors of genes x samples matrix that tabulates gene expression of the genes across the samples. Essentially, this eigengene acted as our cluster representative which allowed us to plot the interconnectedness of our network.

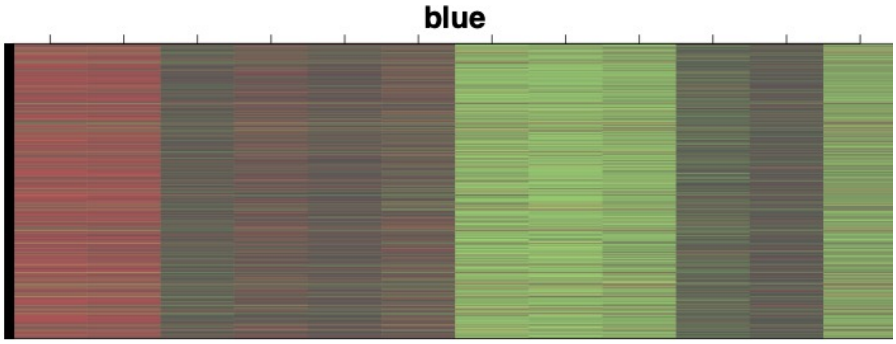
6 Design Evolution

As shown in our initial sketch in Figure 2, we had always planned to create a network with the eigengene bar chart plotted on the node. However, we were unsure how to plot the "zoomed-in" figure. We had sketched out a few tables and networks but none really worked for trend analysis and filtering. We ended up using small-multiple heatmaps for this visualization because that allowed us the flexibility to see multiple features at once for comparison.

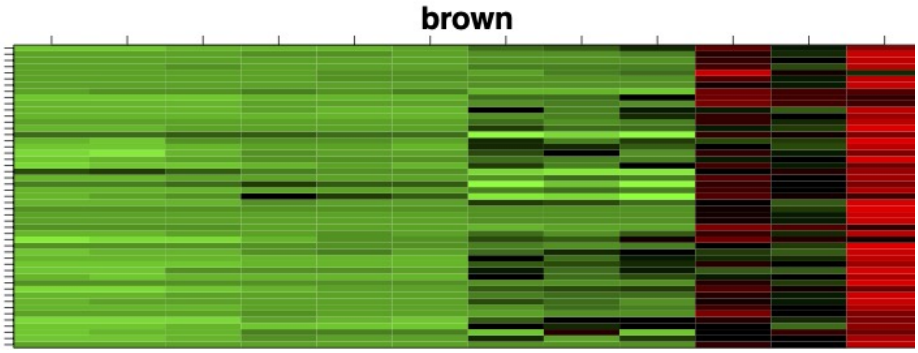
We also considered a few different structures for our network visualization, as shown in Figure 3, in the second note from the right, on the second row, displays one consideration of creating four distinct "areas" for each module and allow us to perform some additional visualizations within our network. However, when we considered how we want this tool to be used in the future, we decided against this method as it requires four clusters, and while it just so happens that our data has four clusters, it is unlikely that all future data sets also contain only four clusters. Thus, we created the network based on weights as show in Figure 3, in the fourth note from the right, on the first row, with a bar chart of the cluster eigengene as the node head.

7 Implementation

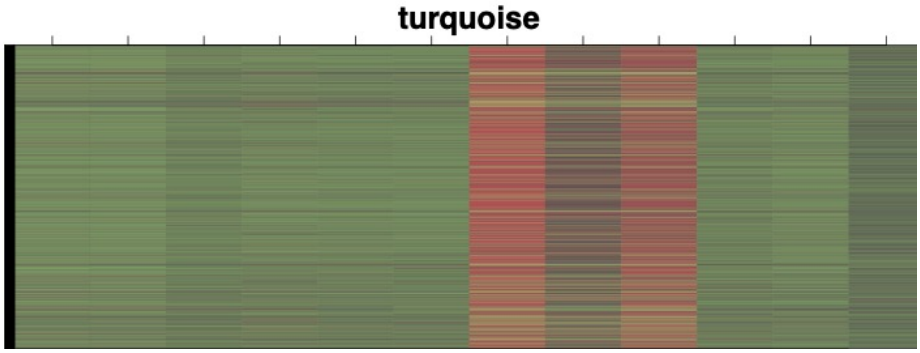
One difficult aspect of interactivity for this data is the various naming conventions for gene annotations. For example, when we are interested in finding all genes related to Zinc's metabolic pathway, we can search for "Zinc" and get lots of results, but we miss any genes labelled "Zn". This is unfortunately a



(a) Cluster "blue" heatmap



(b) Cluster "brown" heatmap



(c) Cluster "turquoise" heatmap

Figure 1: Initial heatmaps of clusters showing overall trends in data

common problem in gene nomenclature, especially across annotation platforms. We decided it would be best to include a filtering menu so that researchers can quickly scan through all unique identifiers to identify such issues.

We also connected our heatmaps to our filtering features so that the user

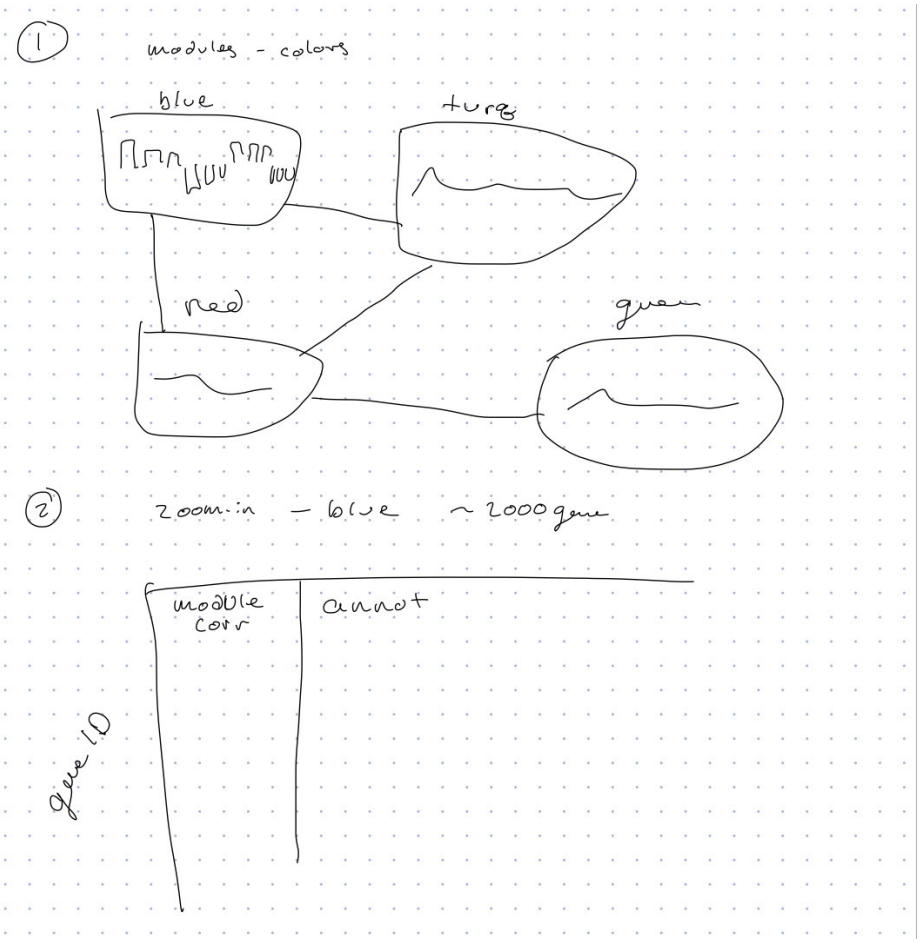


Figure 2: Initial sketch for network multivariate visualization

can select genes of interest and create the corresponding heatmap for them. Figure 5 shows the heatmaps made in d3 without any filters applied.

Once the filters are applied, only genes in those pathways show up in the heatmap and table as shown in Figure 6. Using this filtering method, users can scan through the table to identify other annotations belonging to those genes for further study.

We also added a filter to remove all "NA" values so users can remove all extra information and focus on only the annotated subsection.

8 Evaluation

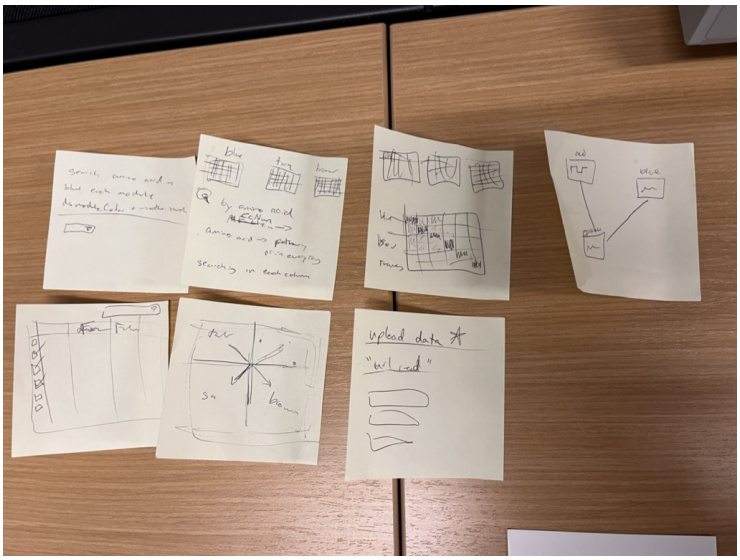


Figure 3: Initial sketches for network multivariate visualization

Figure 4: Initial sketches for network multivariate visualization

An interesting finding that came out of this visualization was just working with the data to create our visualizations forced us to consider methods to cut down our data and that's when we found the method of using fold change values to simplify the data set. Once we did this, as shown in Figure 1 we see clear patterns in each cluster. Before seeing this, we had assumed we needed clusters

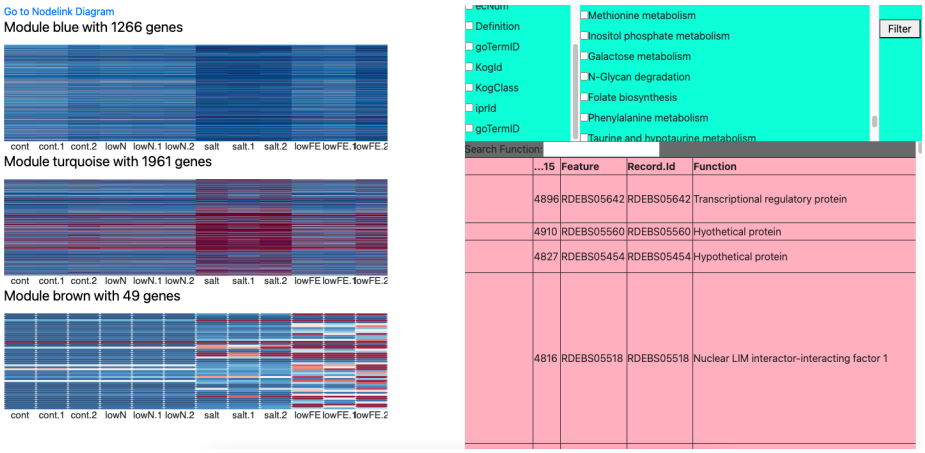


Figure 5: unfiltered visualization

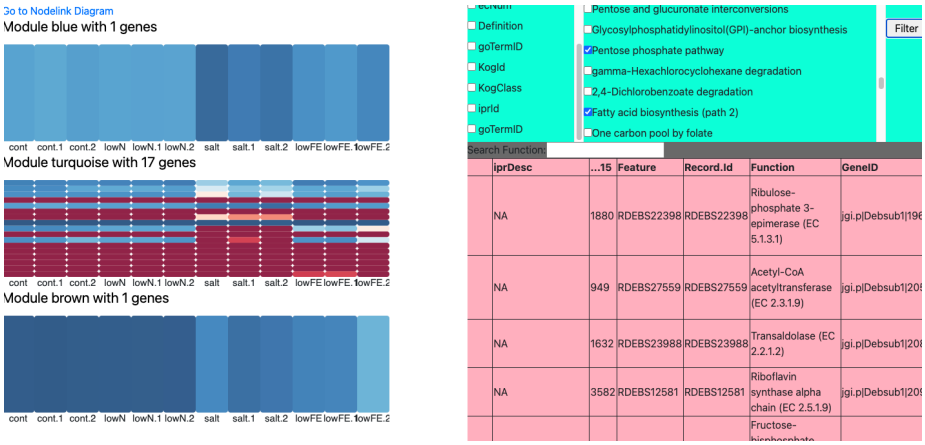


Figure 6: filtered visualization

of approximately less than 1000 to see clear trends.

In the future, we hope to add a data upload feature to our site so it is completely usable for users without JavaScript knowledge. This will make the tool completely user-friendly and allow researchers from any background to use our visualization tool.

With this in mind, we plan on letting researchers use this tool and provide feedback for potential updates.

We also would like to make the tool prettier with some CSS customization, but ran out of time.

References

[1] Jonathan Corum. Kepler's tally of planets, Apr 2013.

[2] Scott Emmons, Stephen Kobourov, Mike Gallant, and Katy Börner. Analysis of network clustering algorithms and cluster quality metrics at scale. *PloS one*, 11(7):e0159161, 2016.

[3] Steffen Hadlak, Heidrun Schumann, and Hans-Jörg Schulz. A survey of multi-faceted graph visualization. In *EuroVis (STARs)*, pages 1–20, 2015.

[4] Nathalie Henry, Jean-Daniel Fekete, and Michael J McGuffin. Nodetrix: a hybrid visualization of social networks. *IEEE transactions on visualization and computer graphics*, 13(6):1302–1309, 2007.

[5] Emily Hindalong, Jordon Johnson, Giuseppe Carenini, and Tamara Munzner. Towards rigorously designed preference visualizations for group decision making. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*, pages 181–190. IEEE, 2020.

[6] JKU Visual Data Science Lab. Caleydo.

[7] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):1–13, 2008.

[8] Fintan Mcgee, Mohammad Ghoniem, Guy Melançon, Benoit Otjacques, and Bruno Pinaud. The state of the art in multilayer network visualization. In *Computer Graphics Forum*, volume 38, pages 125–149. Wiley Online Library, 2019.

[9] Tamara Munzner. A nested model for visualization design and validation. *IEEE transactions on visualization and computer graphics*, 15(6):921–928, 2009.

[10] Carolina Nobre, Miriah Meyer, Marc Streit, and Alexander Lex. The state of the art in visualizing multivariate networks. In *Computer Graphics Forum*, volume 38, pages 807–832. Wiley Online Library, 2019.

[11] Sabrina Nusrat, Theresa Harbig, and Nils Gehlenborg. Tasks, techniques, and tools for genomic data visualization. In *Computer Graphics Forum*, volume 38, pages 781–805. Wiley Online Library, 2019.

[12] Alicia Parlapiano and Jeremy Ashkenas. How the recession reshaped the economy, in 255 charts, Jun 2014.

[13] Hans-Jorg Schulz. Treevis.net.

[14] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[15] Ben Shneiderman and Martin Wattenberg. Ordered treemap layouts. In *IEEE Symposium on Information Visualization, 2001. INFOVIS 2001.*, pages 73–78. IEEE, 2001.

[16] Cagatay Turkay, Aidan Slingsby, Helwig Hauser, Jo Wood, and Jason Dykes. Attribute signatures: Dynamic visual summaries for analyzing multivariate geographical data. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2033–2042, 2014.

[17] Joe Wandy and Rónán Daly. Graphomics: an interactive platform to explore and integrate multi-omics data. *BMC bioinformatics*, 22(1):1–19, 2021.

[18] Jian Zhao, Zhicheng Liu, Mira Dontcheva, Aaron Hertzmann, and Alan Wilson. Matrixwave: Visual comparison of event sequence data. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 259–268, 2015.