12-2019

Customizable Data Visualization of Wnt Signaling

Morgan Lee King
Trinity University, mlking0505@gmail.com

Customizable Data Visualization of Wnt Signaling

BY

MORGAN KING

TO

THE DEPARTMENT OF COMPUTER SCIENCE

IN PARTIAL FULFILLMENT OF REQUIREMENTS
FOR THE DEGREE OF
BACHELOR OF SCIENCE
IN THE SUBJECT OF
COMPUTER SCIENCE

TRINITY UNIVERSITY
SAN ANTONIO, TX
NOVEMBER 2019

Customizable Data Visualization of Wnt Signaling

Morgan King

A departmental senior thesis submitted to the Department of Computer Science at Trinity University in partial fulfillment of the requirements for graduation with departmental honors.

November 11, 2019

Matthew A. Hibbs	Yu Zhang
Thesis Advisor	Department Chair

Michael Soto, AVPAA

Student Agreement

I grant Trinity University (“Institution”), my academic department (“Department”), and the Texas Digital Library (“TDL”) the non-exclusive rights to copy, display, perform, distribute and publish the content I submit to this repository (hereafter called "Work") and to make the Work available in any format in perpetuity as part of a TDL, digital preservation program, Institution or Department repository communication or distribution effort.

I understand that once the Work is submitted, a bibliographic citation to the Work can remain visible in perpetuity, even if the Work is updated or removed.
I understand that the Work’s copyright owner(s) will continue to own copyright outside these non-exclusive granted rights.

I warrant that:

1) I am the copyright owner of the Work, or

2) I am one of the copyright owners and have permission from the other owners to submit the Work, or

3) My Institution or Department is the copyright owner and I have permission to submit the Work, or

4) Another party is the copyright owner and I have permission to submit the Work.

Based on this, I further warrant to my knowledge:

1) The Work does not infringe any copyright, patent, or trade secrets of any third party,

2) The Work does not contain any libelous matter, nor invade the privacy of any person or third party, and

3) That no right in the Work has been sold, mortgaged, or otherwise disposed of, and is free from all claims.

I agree to hold TDL, DPN, Institution, Department, and their agents harmless for any liability arising from any breach of the above warranties or any claim of intellectual property infringement arising from the exercise of these non-exclusive granted rights.”

I choose the following option for sharing my thesis (required):

[X] Open Access (full-text discoverable via search engines)
[ ] Restricted to campus viewing only (allow access only on the Trinity University campus via digitalcommons.trinity.edu)

I choose to append the following Creative Commons license (optional):
N/A

ABSTRACT

As technology advances, biologists are able to obtain more genetic information from experiments than ever before. As the amount of data they produce continues to increase, it is becoming more difficult to process the information and produce results that can be used in biological and medical research. One of the simplest ways to parse information quickly is through visualization. This project aims to improve the readability and utility of graph visualizations for biological pathway analysis by adding interactive and customizable components that allow biologists to determine what view of a dataset is most helpful to answer their questions.

This visualization strategy is being applied specifically to the Wnt signaling process, which is a form of cellular communication that is an area of active research. There is a lot of data about this process available, yet the fine details of Wnt signaling are often overlooked in common visualizations, despite its importance in cancers and other human diseases.

This project produced approximately 450 high level visualizations of Wnt signaling in human genetic datasets that are publicly available for use, and the code was made publicly available as well, so it might be extended to other pathways.

TABLE OF CONTENTS

Introduction

Objective

This project began as a research project analyzing a wide breadth of topics in the field of bioinformatics. From these readings, I developed a specific interest in Wnt Signaling and how it changes in different parts of the human body. To get a closer look at this process, I explored different data visualization tools. However, there is not a biological data visualization tool today that is easy-to-use, accurate, and highly readable. Therefore, I created a data visualization tool that allows users to choose the amount of detail they want to see in the graph they are looking at. Currently, this graph visualization is available for over 450 human Wnt Signaling datasets. In the future, it can be easily extended to other signaling processes and gene groups as well.

Relevance to Biology

As biologists are able to obtain more data from individual experiments, they require better visualization tools that minimize the crowding effect of the large amounts of data they are attempting to analyze. Currently, even the best biological graph visualizations become incredibly complex when examining all of the values found in large genomic datasets. Providing a visualization with “an overview first” that allows biologists to pick “details on demand” means that they can be more thorough in analyzing the specific proteins that they are researching and the data related to them (Shneiderman 1996).

Furthermore, the overview this visualization provides can aid biologists in quickly determining how relevant a particular dataset is to their research as certain colors indicate a

strong presence of Wnt signaling. In reducing the amount of time needed to choose and analyze datasets, biologists are able to focus on the research they actually want to do.

I chose the Wnt signaling process as the starting point for this project because there is a lot of research to be done in this topic. Wnt signaling has been widely studied as a mechanism of cellular communication, but the fine-grained details of this complex process are still under intense scrutiny. Wnt signaling is implicated in a range of cancers and heart failures, which are some of the most common causes of death in America. Additionally, it is highly complex, involving over 40 gene groups made up of over 100 individual genes. In different cellular and environmental contexts, Wnt signaling employs different combinations of these genes. Both cleaning human datasets and developing Wnt-specific visualizations for them makes these data more readily available for scientists researching Wnt Signaling in hopes of better understanding the fundamentals of how it works and potentially how it can be manipulated in treatments for cancer.

Relevance to Computer Science

The type of data I aim to display can be best described as a large dense graph where each individual gene is represented as a node and the correlation between a pair of genes is the weight of an edge between those two nodes. Because so many genes work directly together in the Wnt signaling process, there are many edges to and from each one of the nodes in this graph.

Due to the sheer quantity of edges in this graph, it is incredibly difficult to visualize in a way that is both easy to read and still accurate. While there are some graph visualization tools that address this issue (see the Graph Visualizations section of my Background), most are still very difficult to read, and none provide the ideal “detail on demand” approach (Shneiderman

1996). Current visualizations either display an overview of a general process or the results of visualizing a dataset at a very detailed level. The main goal of this visualization is to provide both options in a single application. Starting with the general view a user should be able to choose how much detail they see and where that detail is focused.

Background

Wnt Signaling

Cell signaling is a process through which cells can receive information about their environment or from other cells. They can then respond to that information by taking action or creating additional signals that are then sent to other cells nearby or at a distance (Dyson 1978). While some signals that cells process are mechanical (e.g. sensory cells in the skin responding to touch), most are chemical and received via receptors (Dyson 1978).

One type of cell signaling in particular, Wnt Signaling, has been implicated in a variety of human diseases and stages of development, including cancer, cardiac disease, and embryogenesis (Logan & Nusse 2004 and Meyer & Leuschner 2018). Wnt knockout phenotypes include a wide range of defects and deficiencies in the female reproductive system, the respiratory system, and human development (Logan & Nusse 2004). Multiple studies have shown that Wnt signaling also has a significant role in the development of many tumors and cancers and a high presence after myocardial infarction (Logan & Nusse 2004 and Meyer & Leuschner 2018).

Despite its apparent role in major health issues, Wnt signaling is often considered from a high-level perspective. Analyzing from this perspective could potentially cause researchers to overlook important subtle details about the process. In general, during canonical Wnt signaling, Wnt proteins from outside a cell interact with Frizzled receptor proteins in the cell wall. This activates the Dishevelled protein, which in turn reduces the inhibition of B-Catenin as seen in Figure 1 below and enhances its role as a transcription factor, which can greatly alter the gene

expression of a cell (Logan & Nusse 2004). However, this high-level view obscures the fact that humans have 19 Wnt genes, 9 Frizzled genes, and 3 Dishevelled genes, and the combination of these genes may be important for the various biological functions that Wnt signaling performs. Beyond these combinatorics, Wnt secretion, how Wnt proteins travel to and from cells, how they bind to Frizzleds, and how the process is regulated are all still unclear (Meyer & Leuschner 2018).

Currently, researchers are developing Wnt inhibitors aimed to treat cancers (Meyer & Leuschner 2018). However, because so little is known about the process as a whole and about how the process changes in different types of cells, very little progress is being made in that domain. Research into these processes and further understanding of the Wnt signaling process as a whole could lead to the development of therapeutics and drugs that help treat cancers and heart disease.

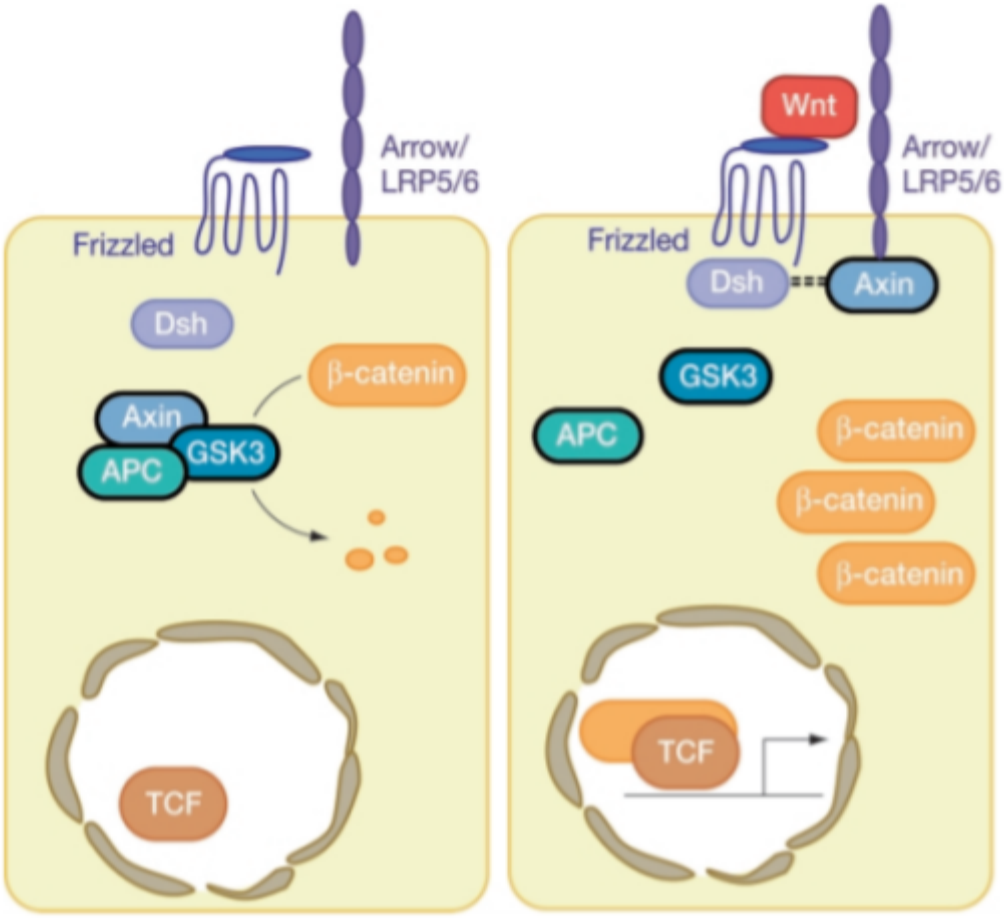

Figure 1. Reprinted from Logan & Nusse 2004. Depicts the effect of a Wnt signal in the canonical Wnt signaling pathway. In the right panel, Wnts bind to Frizzleds or LRP, signaling to Dishevelled, and degrading β-Catenin inhibition.

Microarray Data and Data Cleaning

The data being used in this project is human gene co-expression data found in microarrays. A DNA Microarray is a tool used by biologists to observe the presence of thousands of genes in a sample. It consists of a flat surface and thousands of probes. Once exposed to a target sample, the probes change color according to how present particular genes are in the sample (Conzone & Pantanot 2004). From there, the coloration of a microarray can be converted

into numerical data. The results of many experiments involving the same genes can then be output into a gene expression matrix where each row represents an individual gene and each column represents an individual experiment. Beyond gene expression microarrays, many other technologies produce similar matrices of data. The visualization methods discussed and presented here could potentially be applied to these other data types as well.

While this technology has allowed biologists to obtain large amounts of genetic data extremely quickly, microarray data is currently not standardized (Conzone & Pantanot 2004). Matrices often have missing values and do not adhere to any one numerical range. Different datasets can contain entries between 0 and 1, -1 and 1, 0 and n, $-n$ and n, or even m and n, where n is a positive integer and m is an integer. This lack of standardization makes comparing datasets much more complex. Further, missing values cause clustering algorithms to become less accurate and can cause programs to crash if left unhandled (Troyanskaya *et al* 2001). As such, microarray data is incredibly difficult to analyze once collected, and there is currently no standard tool to read it in or analyze it.

As such, when using microarray data, all data cleaning must be done by hand. The first step in cleaning this type of data is replacing missing values. This process is known as imputation, and there are multiple ways to perform this task. One of the more common forms of imputation is mean imputation, or row averaging, which involves replacing a missing value with the mean of the row that value is found in. In microarray datasets, this involves averaging the presence of the gene across experiments and replacing the missing values with this average (Troyanskaya *et al* 2001). This fails to take into account the fact that genes will be far more or

less present in different experiments depending on the environment and processes taking place (Troyanskaya *et al* 2001).

To reduce some of that error, other types of imputation such as Singular Value Decomposition (SVD) imputation and K-Nearest Neighbors (KNN) imputation are often used. SVD imputation begins by performing the row average imputation. It then repeatedly maximizes its estimates until the best fit for the missing value is found. KNN imputation involves finding *k* genes that have a similar overall presence to the gene with missing values. It takes the average value of these genes in an experiment and replaces the missing value with this number (Troyanskaya *et al* 2001).

For this project, I used KNN imputation. While both SVD imputation and KNN imputation produce dramatically better results than a simple row average, this form of imputation is ideal for preserving precision and reducing error (Troyanskaya *et al* 2001). KNN imputation outperforms SVD imputation in noisy datasets, time series, and non-time scaled datasets (Troyanskaya *et al* 2001). Because KNN imputation has been shown to produce the best results for microarray data, it is also a form of imputation that has been automated.

Graph Visualizations

The type of data described above can also be described as a dense graph. Every node, or gene, is linked to most, if not all, of the other genes being represented. At a scale of even only 30 genes, this type of graph becomes impossible to read when displayed as a graph (see Figure 2A).

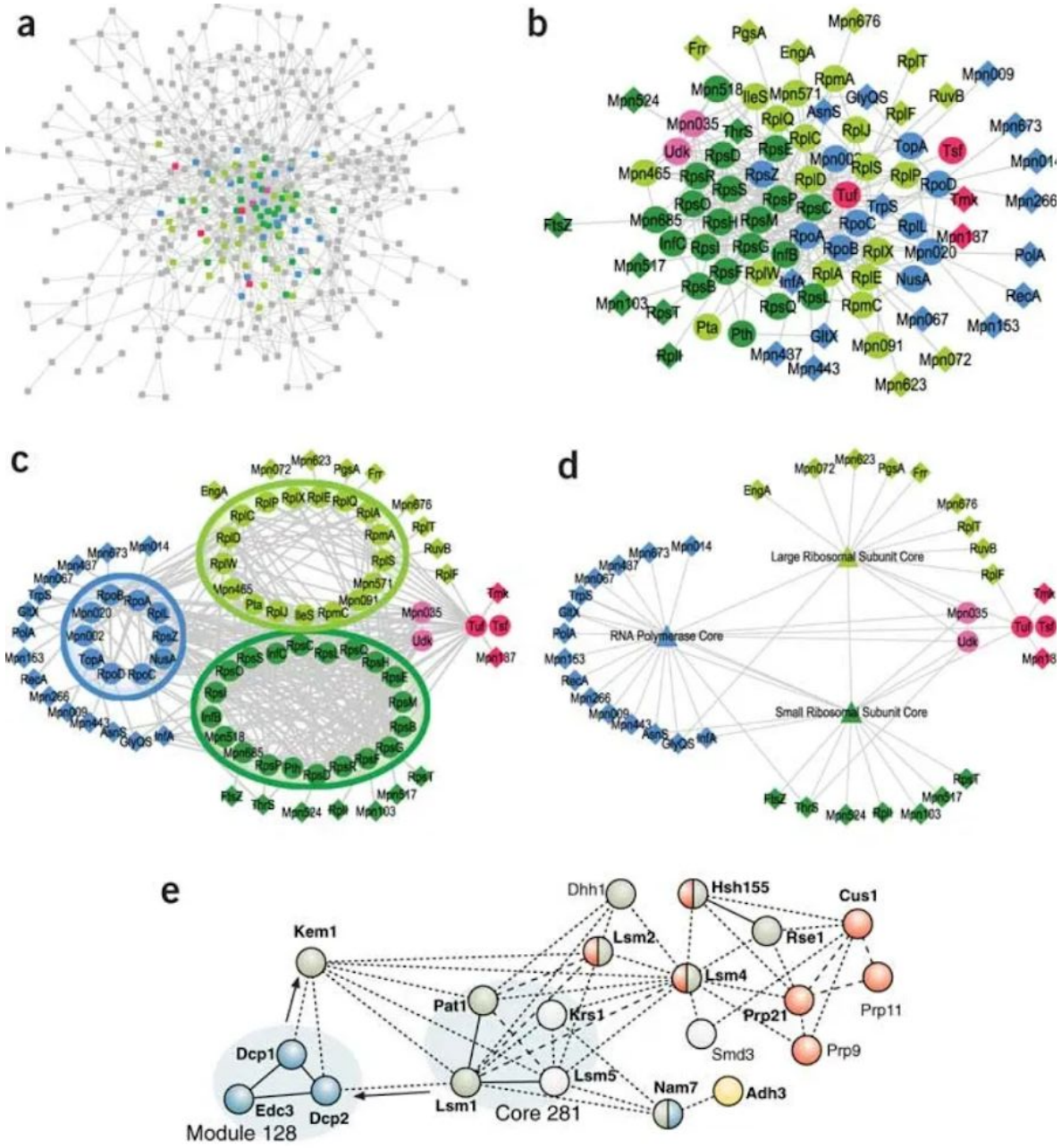

Figure 2. Reprinted from Gehlenborg *et al* 2010. Some examples of different visualization techniques being used in biological research today. Part A shows a simple graph. Part B shows a similar graph where nodes are labeled and shaped and colored by role and type. Parts C and D depict different clustering techniques meant to improve readability.

Simple improvements include coloring nodes by the complex they belong to and labeling them with the genes they represent (Figure 2B). Other improvements include allowing users to group proteins they'd like to compare to one another, clustering closely related proteins, or filtering the visualization by location (Figure 2CDE). All of these improvements dramatically improve graph readability and allow users to make insights they would otherwise be unable to see (Gehlenborg *et al* 2010).

Another type of visualization, the KEGG pathway, eliminates all of the variable links between nodes and simply displays the pathway of a given signaling process. While this contains no physical data, it does provide a clear description of the system it is visualizing as seen in Figure 3 (Kanehisa & Goto 2000).

Each of the above types of visualizations have their own advantages: The data visualizations help viewers parse large amount of complex data, and KEGG pathways provide a simplistic overview of complicated pathways without additional detail that might confuse the viewers. However, there is no single model that allows users to choose the level of detail they see. Current visualizations either provide a static view of a general pathway or the results of visualizing a dataset.

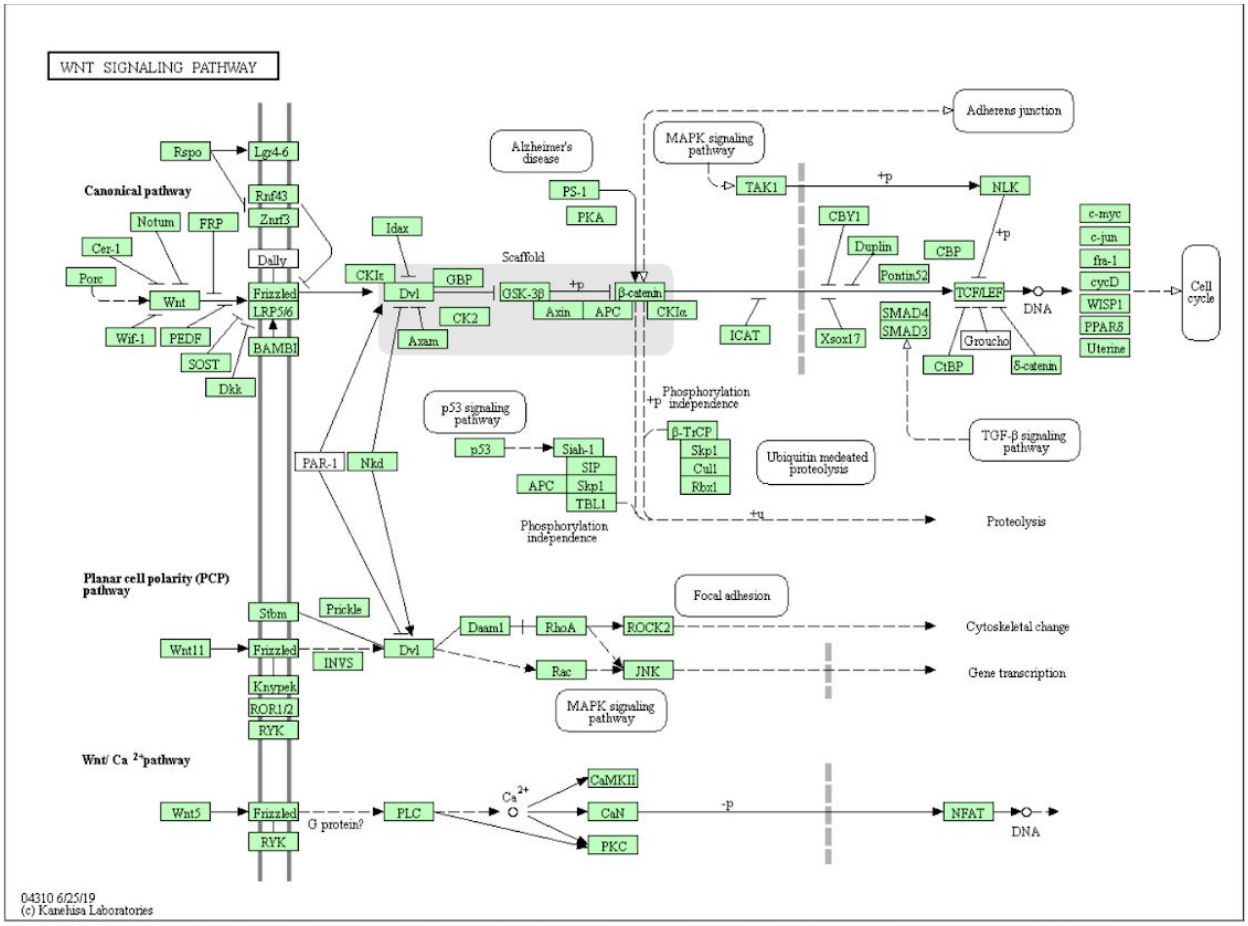

Figure 3. Reprinted from Kanehisa & Goto 2000. Canonical and non-canonical Wnt signaling pathway overview.

Methods

Open-Source Software and Other Tools Used

I obtained the data used in this project using Python and the HumanBase API. The HumanBase API is a RESTful API that allows users to find and search for human genetic data (https://hb.flatironinstitute.org/). All of the data cleaning, histogram building, and JSON generation for this project was done using Python and some of its associated libraries (json, pandas, numpy, math, os, re, matplotlib, and scipy stats). The visualization itself was built in HTML and Javascript using the open source library force-graph by vasturiano (https://github.com/vasturiano/force-graph). This library is built on top of the D3 javascript library, and it allowed me to use D3's force simulation capabilities to generate graphs quickly and in a way that was easy to read.

Obtaining Data

All datasets were downloaded from ftp://ftp.ncbi.nlm.nih.gov/geo/datasets/ using Python. I specifically selected those datasets that were known to contain human genetic data in the form of co-expression. Which datasets met those qualifications was determined using the HumanBase API.

The datasets were originally downloaded as .soft.gz files. Using Python code borrowed from a prior project, they were finally converted to a usable .pcl file format and remained in this format for the duration of my project (Hibbs 2007). The .pcl format is a simple tab-delimited plain text format suitable for all downstream analyses. In total, I found and used approximately

450 human genetic co-expression datasets spanning a range of biological contexts and manipulations.

Cleaning Data

The bulk of this project involved cleaning data as a variety of issues with the datasets caused multiple delays. Before I could even begin to standardize the data I had, I first had to run imputations on all of the datasets as many of them had null or empty rows, which caused any standardization efforts to crash. Imputation was done using a KNN imputer built in the 2001 project, "Missing value estimation methods for DNA microarrays," and took multiple hours to run due to the sheer amount of data being cleaned (Troyanskaya *et al* 2001).

Once I could verify that there were no null values in any of my datasets, I could start to standardize the data. As mentioned in the Data Cleaning section of the earlier chapter on Background research, these datasets can be highly variable and inconsistent. Using Python code that I wrote and software from prior projects, I was able to take the now imputed datasets and standardize them over multiple hours (Hibbs 2007).

Once the project was started was I able to find areas where imputation was necessary. Further still, the lack of standardization in these datasets was not evident until that imputation was complete.

Building Histograms

Once the data was in a usable form, I wanted to see which would be most useful in studying Wnt Signaling. One way to do this is by examining the distributions of correlations

between genes in the datasets and whether or not these distributions change when only comparing genes found in Wnt Signaling.

To do this, I built overlapping histograms for each individual dataset. The bottom layer of the histogram displayed the distribution of correlations between random genes in the dataset. For most datasets, this distribution was normal with a mean of about 0 (no correlation). The next layer of the histogram contained only correlations between genes that are a part of the Wnt Signaling process. If a dataset contained Wnt Signaling, this layer of the histogram was much less likely to be normal, or its mean would be closer to 1 (positive correlation). Figure 4 displays an example of a dataset that is likely to contain Wnt Signaling.

From there, to gain an even more focused understanding of how Wnt Signaling affected the dataset, I added a third layer to these histograms that specifically compared correlations of genes that directly interact with one another in the Wnt Signaling process. Because these pairs occur much less frequently in datasets than the other two types of correlations, this third layer tends to appear less normal. However, in datasets where Wnt Signaling is present, this third layer is much more likely to have a positive mean (compare Figure 4 and Figure 5).

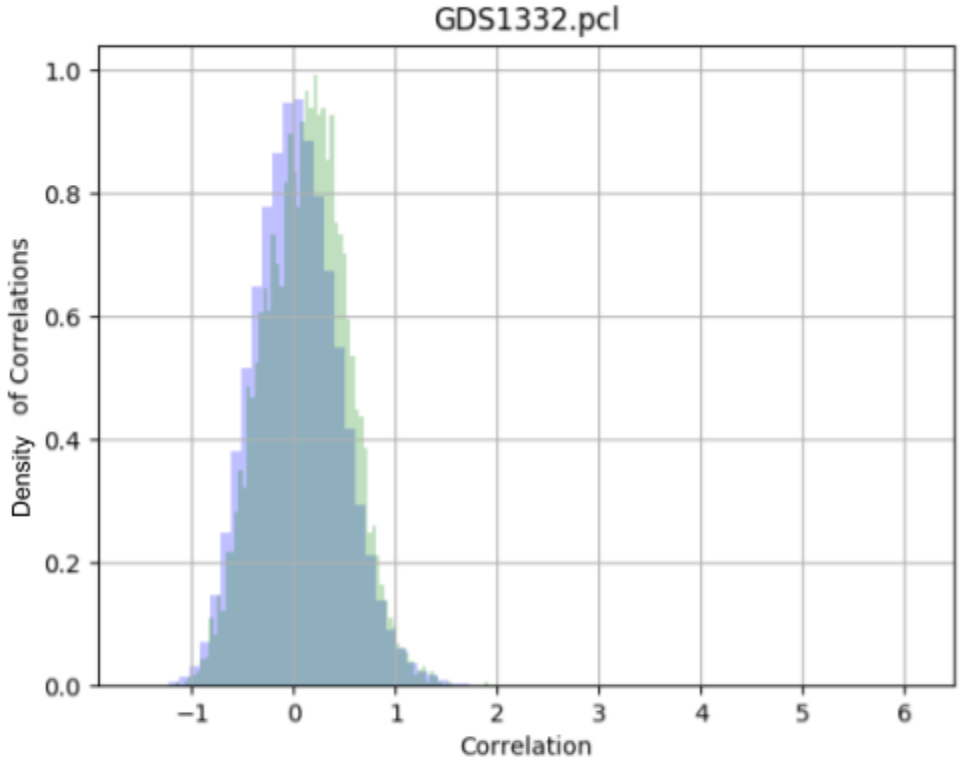

Figure 4. A histogram of the correlations of gene-pairs in a given dataset. The underlying blue layer is the distribution of all gene-pair correlations, and the green layer is the distribution of Wnt signaling gene-pair correlations. Because the green layer has a mean that is greater than 0, it is likely that Wnt signaling exists in this data.

To generate the correlations used in these histograms, I randomly selected pairs of genes from either the entire dataset, the entire Wnt Signaling chain, or connected pairs in the Wnt Signaling chain depending on the histogram layer. I then used the Pearson correlation function found in the scipy.stats Python library. Pearson correlation was used because it determines the linear relationship between two genes (i.e. gene A appears more often when gene B appears, and gene B is less likely to appear when gene A is less prevalent). From there, I applied the inverse

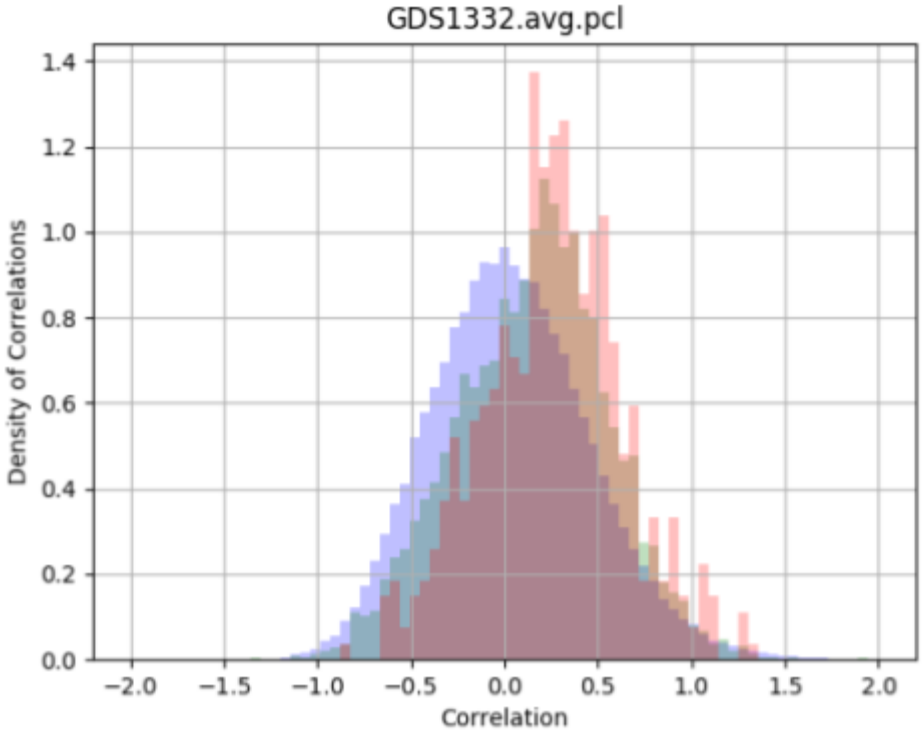

Figure 5. A histogram of the correlations of gene-pairs in a given dataset. The underlying blue layer is the distribution of all gene-pair correlations, the middle green layer is the distribution of Wnt signaling gene-pair correlations, and the red layer s the distribution of Wnt signaling gene-pair correlations where the gene-pairs directly interact during Wnt signaling. Because both the green and red layers have a mean that is greater than 0, it is likely that Wnt signaling exists in this data.

hyperbolic tangent to normalize the shape of the bell curve found in the histogram. This process is known as the Fisher Z-transformation (Fisher 1915). I also took a 2-sample Kolmogorov-Smirnov statistic to determine whether each of the datasets had a statistically significant amount of Wnt signaling correlations and only generated histograms for those datasets that did have statistical significance. Statistically significant datasets had a p-value of 0.05 / 769.0 (alpha = 0.05 over the total number of datasets to account for random chance). In total, I created histograms for 473 datasets.

Building Initial Graphs - Specific & General

For testing purposes, I selected a dataset I knew looked very normal and contained many of the genes involved with Wnt Signaling (GDS 3571). While building the initial graph designs, I only used this dataset.

Before I could even look at the graph itself, I had to create JSON files that could be easily parsed into the visualization I was creating. To do this, I used Python to find: all the genes that exist in Wnt signaling, which of those genes also exist in the dataset I was visualizing, the correlations between these genes and the genes they interact with in the Wnt signaling process, and the correlations between these genes and every other Wnt signaling gene present in the dataset. Once these data were found and computed, I then adjusted it to a standardized JSON template.

Once the JSON was built, it was easily incorporated into a visualization created using a Javascript script in an HTML file and an open-source library called force-graph. Force-graph allows for highly readable and highly customizable graph visualizations. The first fully-informative and accurate version of this visualization is depicted in Figure 6.

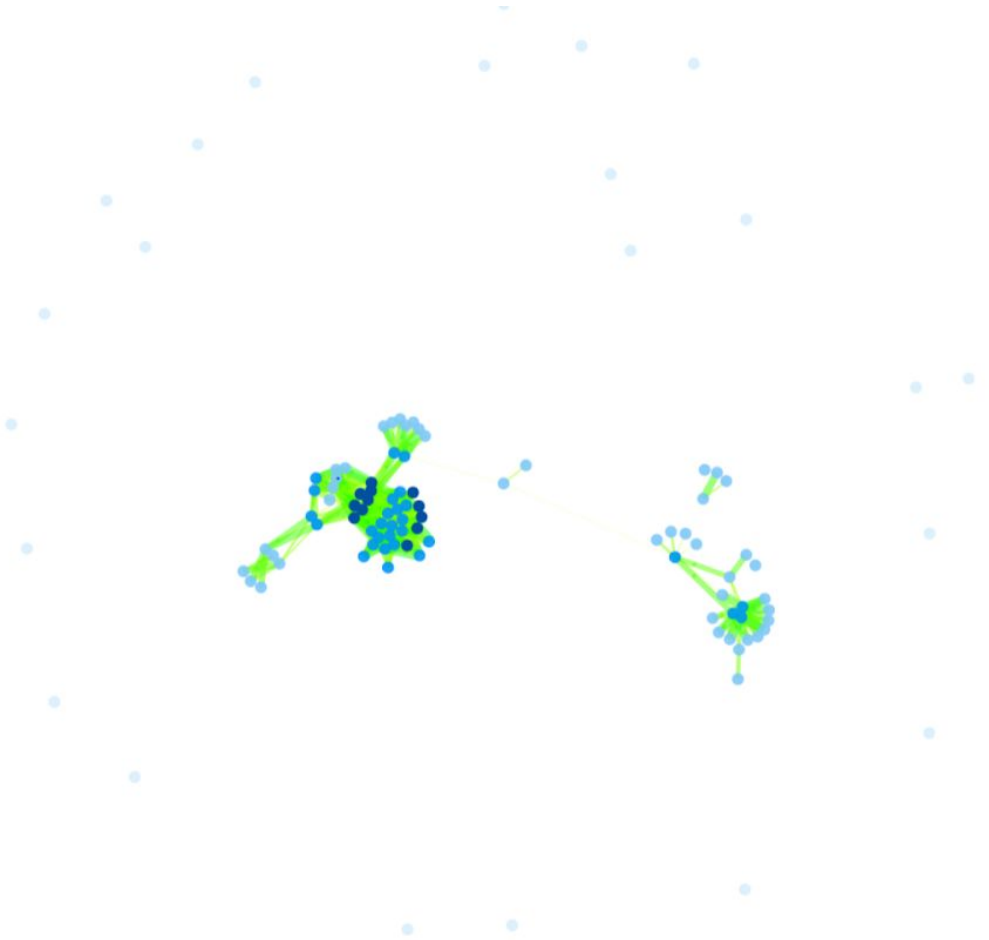

Figure 6. Wnt Signaling Visualization before readability improvements.

The color of the nodes in the graph encode how many links are connected to each node. A darker blue signifies that it is more connected than lighter blues. Hovering over a node would cause a pop-up to display the name of the gene that the node represents. The color and thickness of a link signify how correlated the two nodes it connects are (green being more correlated and yellow being less). Hovering over a link would cause it and the two nodes it connects to turn red. This color scheme choice was partly based on pre-existing visualizations (see Figure 13). Some of these capabilities can be seen in Figures 7 and 8.

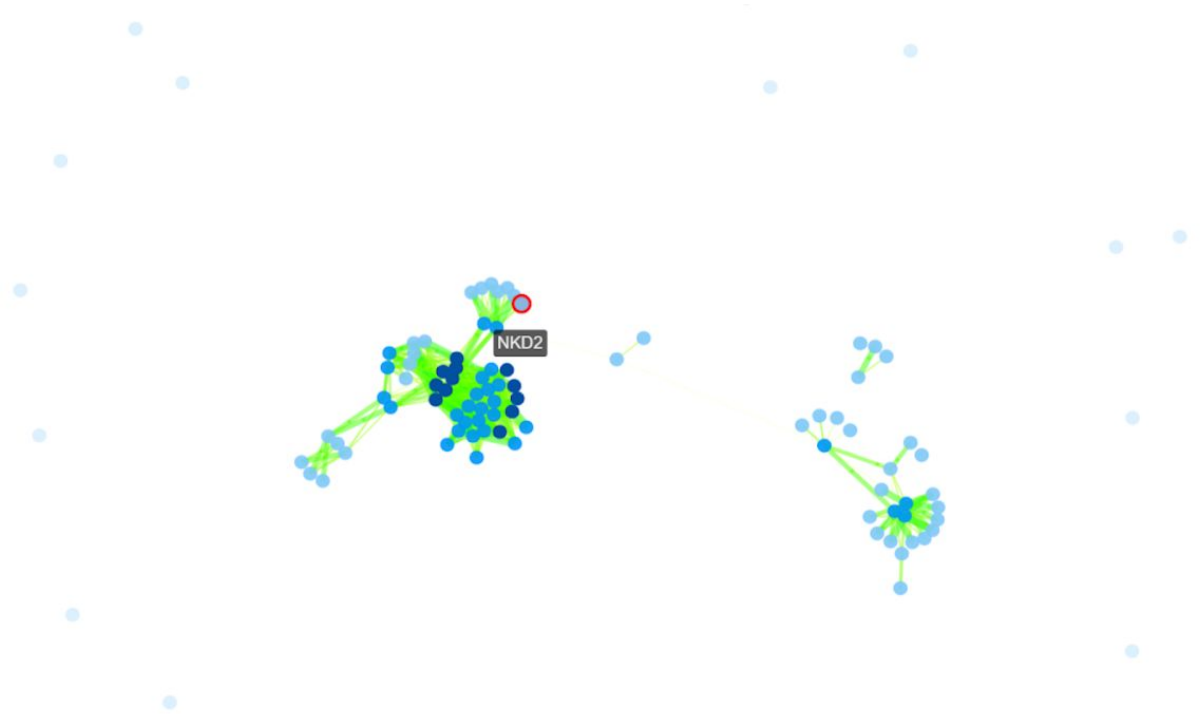

Figure 7. Hovering over an individual nodes highlights the node in red and prints the name of the gene the node represents in the Wnt signaling visualization without readability improvements.

At this point, the visualization was in a fully stable stage. I then created a second JSON file that compared general groups of genes against other general groups of genes that exist in the Wnt Signaling process (e.g. Wnt and Frizzled instead of Wnt11 and Fzd4). This JSON template followed the same standard as the previous one. However, it contained the correlations between groups of genes that are connected in Wnt Signaling. This was computed by averaging the correlations between every pair of genes between one group and another group. The end result

was a similar looking graph with many fewer nodes.

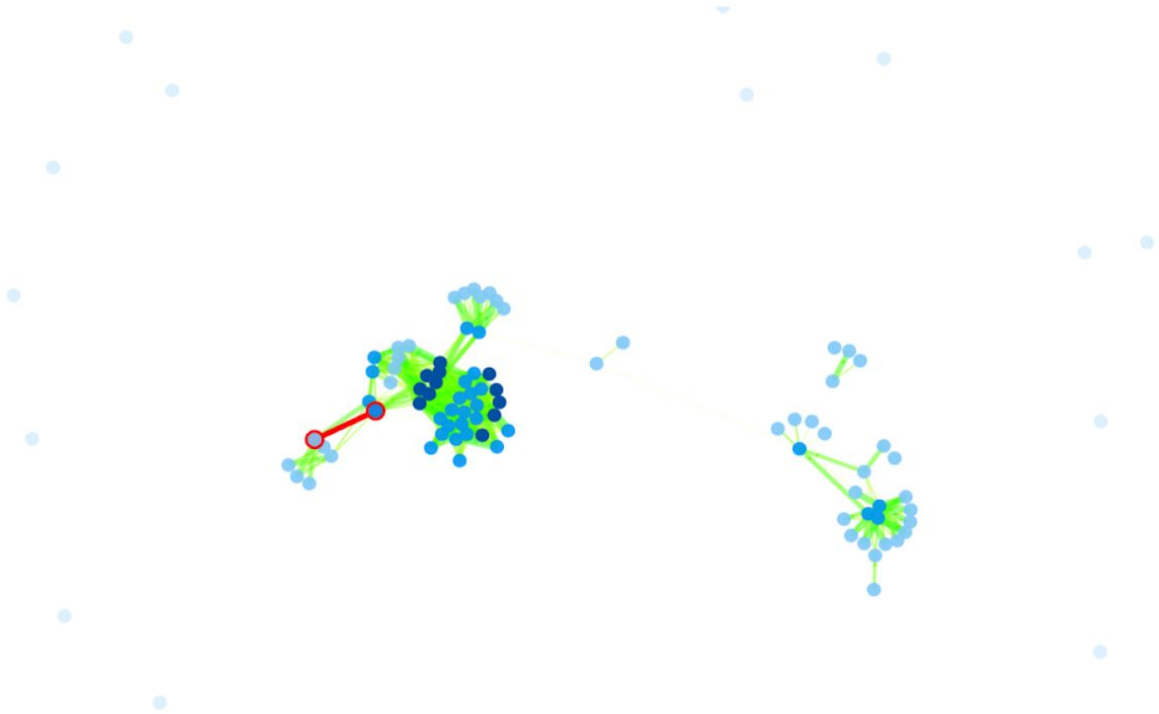

Figure 8. Hovering over a link highlights the link itself and the two nodes it connects in the Wnt signaling visualization without readability improvements.

Improving Graph Readability

While this was accurate, the graphs I initially produced were highly unreadable. The link color was difficult to discern, the nodes and links between them were indistinguishable from one another, and from a single glance, a user could not tell what the visualization was even for.

The first step in producing a more readable graph was to replace the circular nodes with text to display the name of the gene/gene group the node represents. This way, from a glance, a user could instantly determine what genes were being compared and which genes played the biggest role.

From there, the next step was to make the entire visualization color-blind friendly. In its current stage, the links were a color between yellow and green. This color combination was difficult to differentiate and invisible to anyone with certain forms of color-blindness. From the zoomed out perspective in Figures 6, 7, and 8, the yellow lines appear completely invisible. Therefore, I switched the links to a teal range, which was more discernible, visually appealing and more accessible.

Next, to reduce the amount of overlap between links, I increased the amount of separation between nodes and decreased the thickness of all of the links. These improved proportions along with smarter color choices resulted in a much more user-friendly visualization as seen in Figure 9.

Finally, I wanted users' attention to be drawn to the most important links and nodes in the visualization first. Therefore, I added an additional moving link element and a color element: text background. The moving link element is a ball that moves up or down a link at a speed based on how correlated the two nodes are. This moving ball goes in the direction of the correlation (e.g. if the presence of gene A causes more of gene B to occur, the ball moves from node A to node B). This both draws a user's eyes towards those links and improves the understanding of the pathway.

Those nodes which were most correlated with every other node (and, by extension, most important to the graph as a whole) were given an orange background. The more vibrant the orange, the more connected it is to the graph as a whole. The orange, teal, and navy blue text combination is color-blind friendly regardless of the combination. The final makeup of the general visualization can be seen in Figure 10.

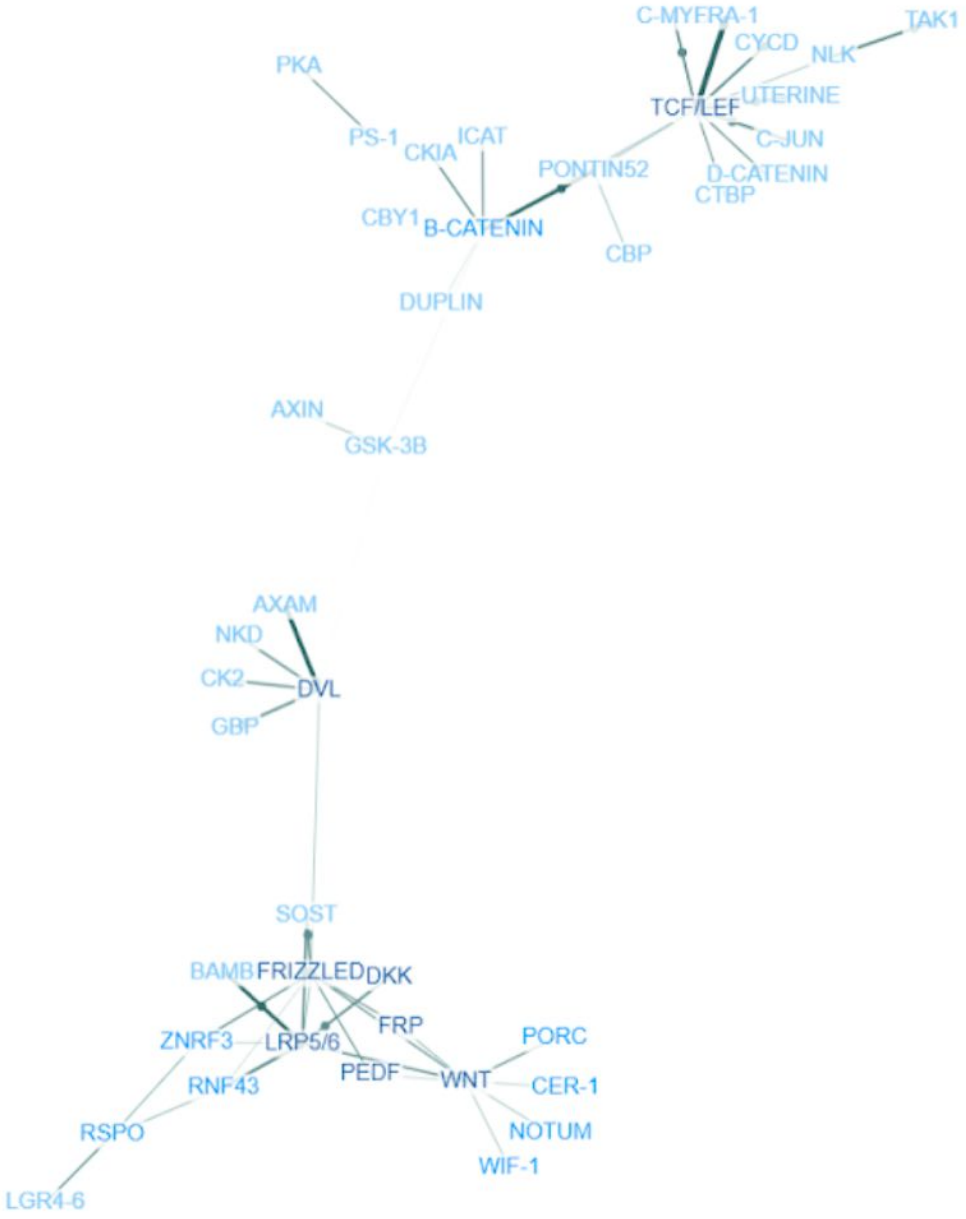

Figure 9. Wnt signaling visualization with first readability improvements in a zoomed out view.

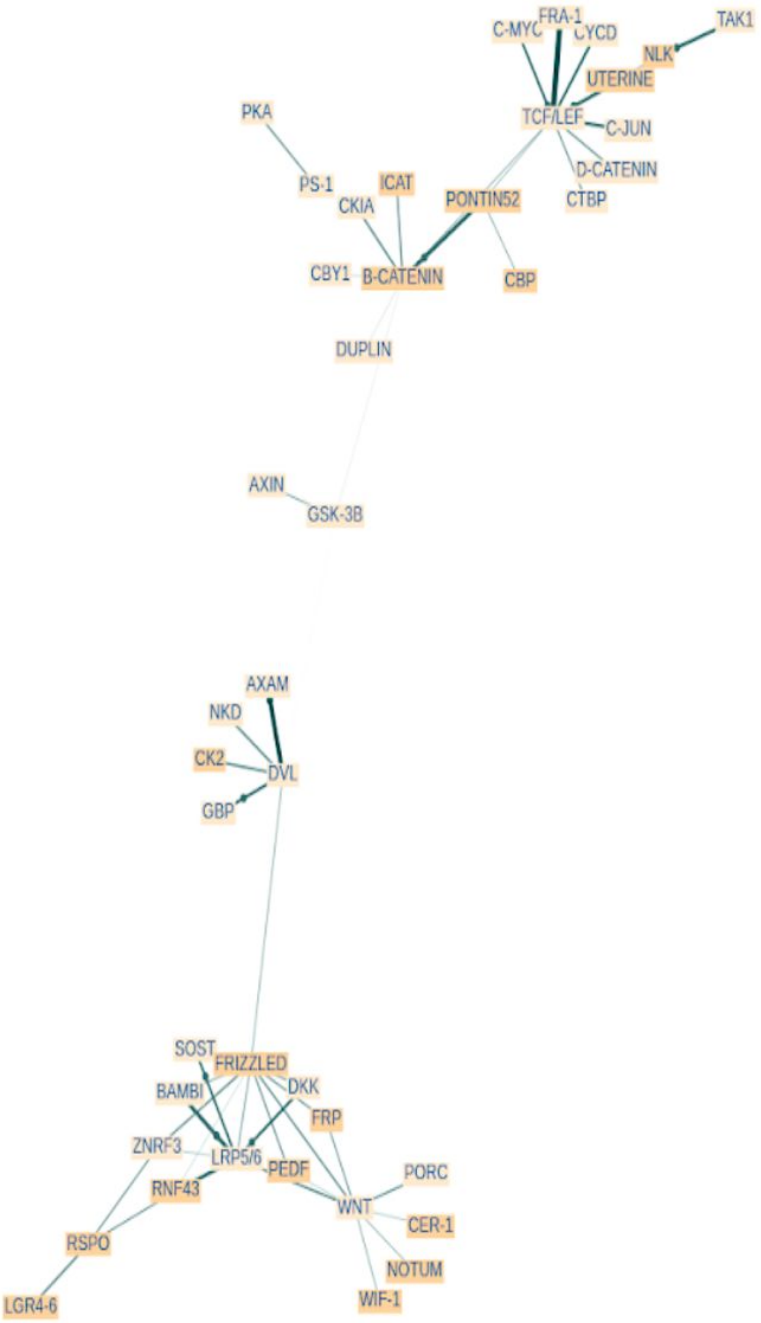

Figure 10. Wnt signaling visualization with all readability improvements in a general view. The darkness and thickness of a link represents how correlated the two connected nodes are, and the orange background on nodes represents how correlated that node is to the system as a whole.

A New Visualization Tool

Once the general and specific versions of this visualization were in their most readable states, I could start to bring them together into the final visualization tool. To begin, the graph would appear like the general visualization found in Figure 10. From there, the user can click on any of the gene groups, and the view would switch from general to specific for that individual gene as seen in Figure 11.

This was accomplished by first combining the nodes and links from both the general and specific visualizations. I then added a third “parents” section to the JSON standard which containing links from gene groups to the genes in those groups. It also included a link from the gene groups to a single “ROOT” node, which pulls the entire graph together. From there, I added additional links that symbolize the correlation between individual genes and groups of genes, so the user can pick the level of correlation they wish to view.

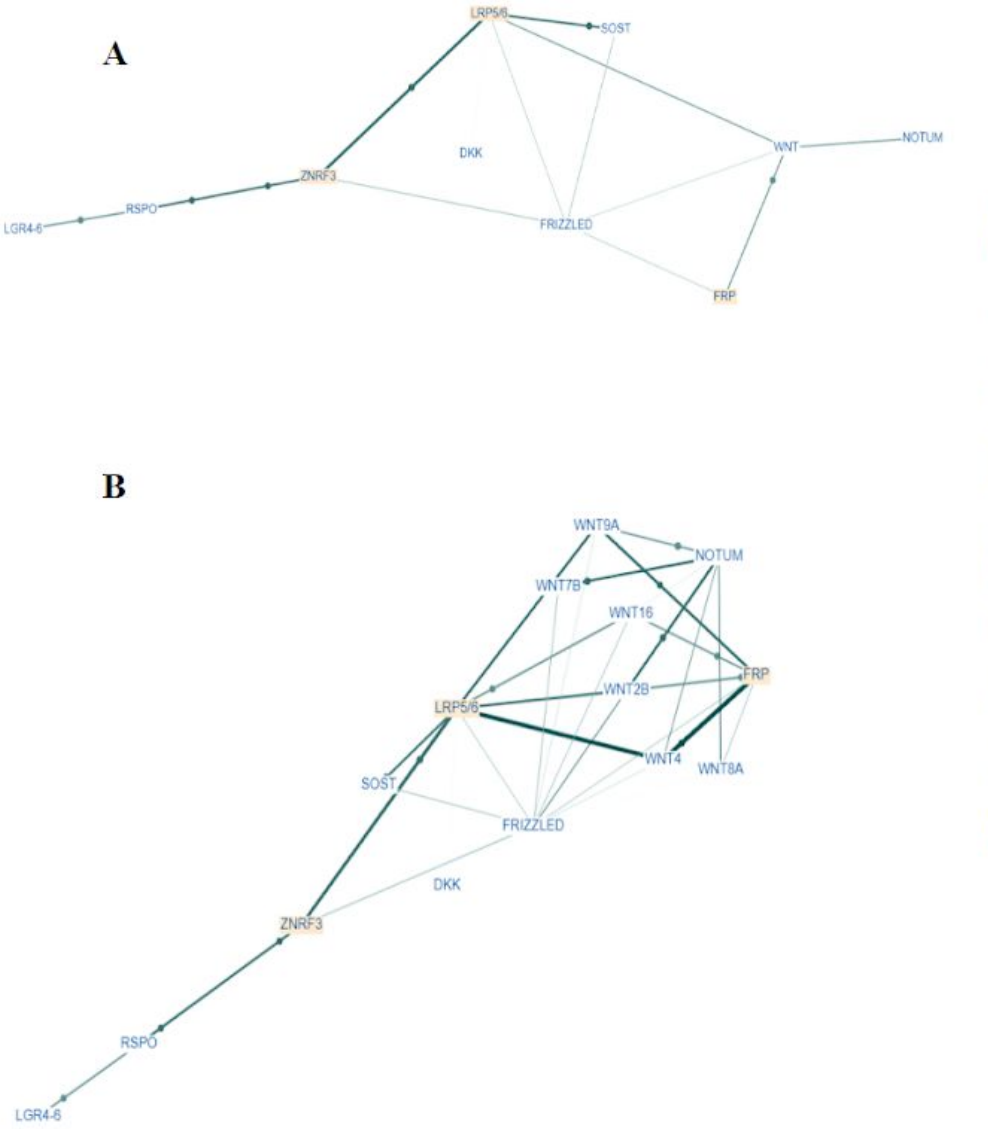

Figure 11. Part A shows the Wnt signaling visualization in the zoomed out view. If a user were to click on the WNT Node in this diagram, it would expand into the view depicted in Part B

Results

Once the visualization was finished for the first dataset, I applied the visualization to every other human gene expression dataset I had standardized by running a final Python script that created every JSON file, combined them, and duplicated the HTML and javascript needed. It ran on about 450 datasets in under an hour. All of these visualizations are available at lumos.cs.trinity.edu/mking4/wntViz/.

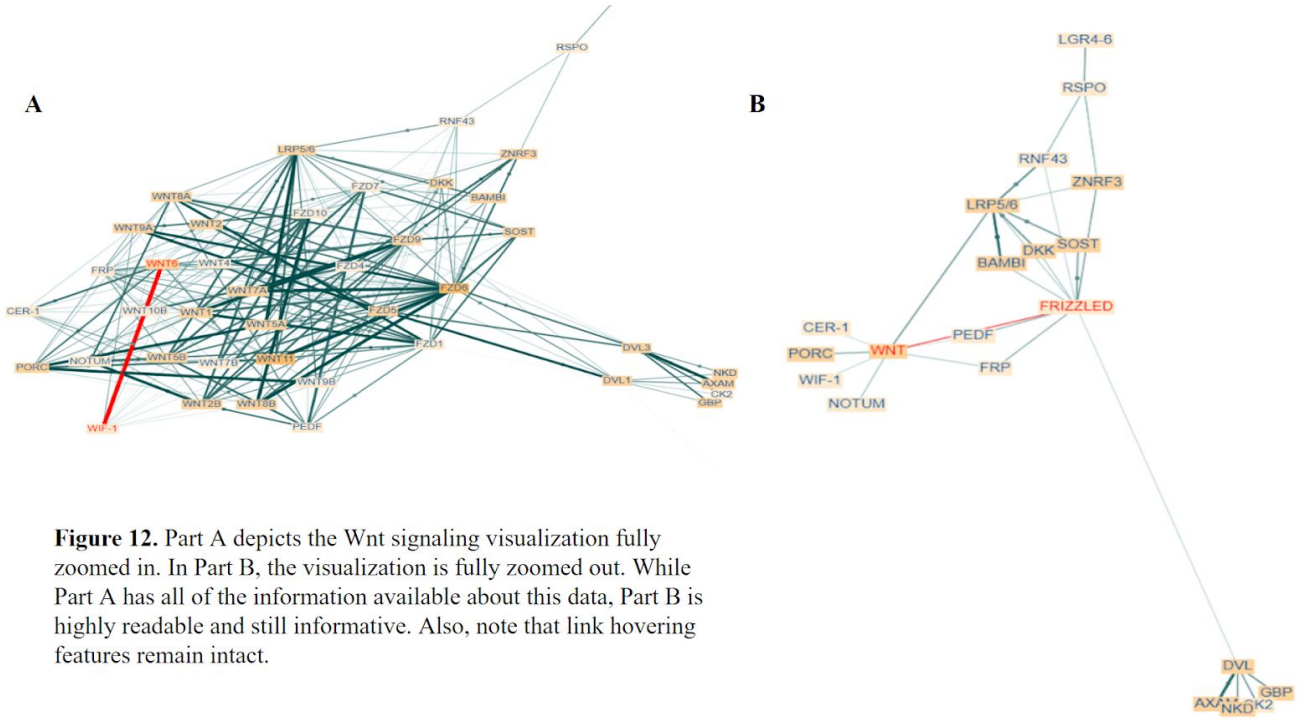

Figure 12. Part A depicts the Wnt signaling visualization fully zoomed in. In Part B, the visualization is fully zoomed out. While Part A has all of the information available about this data, Part B is highly readable and still informative. Also, note that link hovering features remain intact.

Figure 12 (shown above) is an example of a densely populated dataset using my visualization. While comparing specific genes to other specific genes is still difficult to read, the

highlighting tool improves overall readability, and the most zoomed out version of the graph makes the visualization as a whole highly readable (compare Figure 12A to Figure 12B).

Other Wnt signaling visualizations (such as the one shown below in Figure 13) lack the ability to zoom and filter gene groups out. As such, there are limits to how readable the visualization can be. While node stacking is a dramatic improvement over clustering and randomly distributed nodes, the interactive capabilities my visualization provides takes readability one step further. Furthermore, the stacked nodes only visualize 4 total gene groups whereas my visualization places those 4 gene groups in the broader context of the entire signaling process without losing readability.

Additionally, because my visualization is based off of the KEGG pathway for Wnt Signaling, it maintains the overall shape of the KEGG pathway as shown in Figure 3. The user can see which genes directly interact during the Wnt Signaling process. Unlike the KEGG pathway, my visualization also allows the user to see how specific genes interact during Wnt Signaling as well as gather information about how correlated those genes are in a given experiment or dataset.

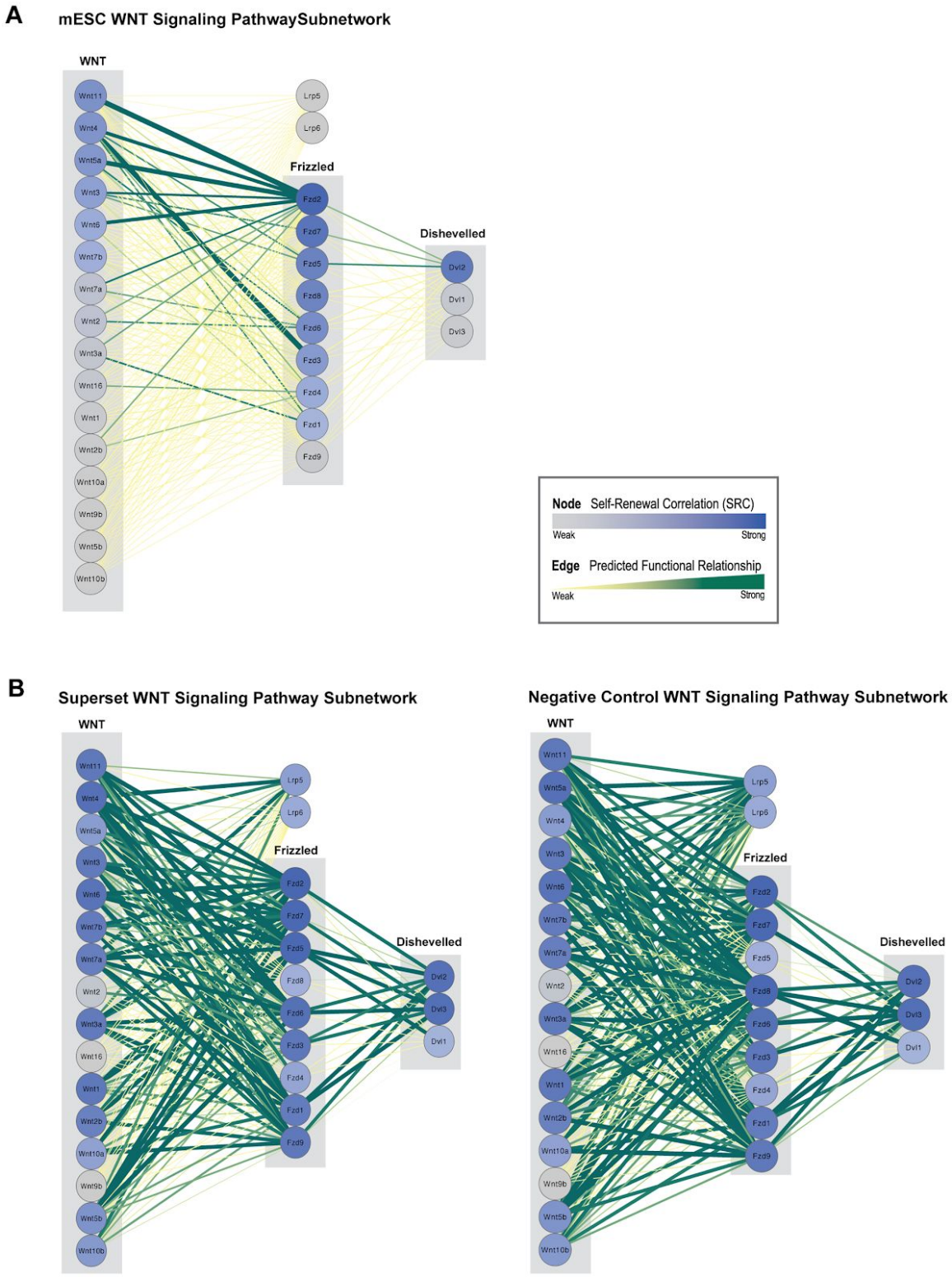

Figure 13. Reprinted from Dowell *et al* 2013. A visualization of Wnt signaling featuring stacked nodes. While this model is certainly more readable than other previously mentioned visualizations, this only contains four gene groups and is still overwhelming.

Conclusions & Future Work

As described in Results, I successfully completed my goal of creating a customizable data visualization that allows an end user to choose the amount of detail they see. This tool has been applied to hundreds of datasets and is available for viewing at: lumos.cs.trinity.edu/mking4/wntViz/.

Some of the high points of this project include the highly readable color scheme and its overall scalability. The color scheme is both color-blind friendly and attractive. The monochromatic color ranges compliment one another and clearly convey which nodes and edges are most important. This readability was further improved by adding hover over features which allow users to focus on individual links. Additionally, once the visualization was finalized, it was incredibly easy to scale to every dataset I had available. This visualization can be easily extended to include any other cleaned KEGG datasets.

However, there were some shortcomings during this project as well. Cleaning datasets took an incredible amount of time, which meant I was unable to extend this project as far as I wanted it to go. I hoped to gain further insights on Wnt Signaling through machine learning, which I could not even begin. Further, the layout of the specific genes is still more condensed than I would like it to be.

A first future first project would be improving the layout of the graph as a whole. I think the general shape of the graph should be angled rather than linear, and many of the nodes should be pushed further apart. I would also try to extend to more datasets and create an automatic dataset cleaning tool for gene-co expression microarray datasets from KEGG. These additional

steps would be positive contributions towards standardizing microarray data and improving how insights are ascertained from this type of input.

From here, this visualization tool can be easily extended to other genetic pathways and other types of dataset. Though this tool is currently designed for KEGG datasets involving Wnt Signaling specifically, new files can be built replacing the Wnt Signaling pathway with other options. How they are built can be modified depending on the type of datasets received.

Beyond this, future work on this project could also include automatically converting KEGG pathways to the proper format to be read in by this tool. In accomplishing this, a biologist could easily upload a KEGG pathway to the visualization tool and look at any dataset they choose.

While there is still much that can be accomplished for this project, it is a successful new visualization tool for Wnt Signaling that can be used today. It allows users to customize the amount of detail they see and in what areas they want that level of detail. This visualization is highly customizable and extendable with minimal improvements.

References

Angermueller, Christof, Heather J. Lee, Wolf Reik, and Oliver Stegle. "DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning," Genome Biology, vol. 18, no. 67, 2017.

Chen, Kathleen M., Evan M. Cofer, Jian Zhou, and Olga G. Troyanskaya. "Selene: a PyTorch-based deep learning library for sequence-level data," Princeton, New Jersey and New York City, New York: Princeton University and Flatiron Institute, 2018.

Ching, Travers, Daniel S. Himmelstein,, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, Brian T. Do, Gregory P. Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M. Hoffman, Wei Xie, Gail L. Rosen, Benjamin J. Lengerich, Johnny Israeli, Jack Lanchantin, Stephen Woloszynek, Anne E. Carpenter, Avanti Shrikumar, Jinbo Xu, Evan M. Cofer, Christopher A. Lavender, Srinivas C. Turaga, Amr M. Alexandari, Zhiyong Lu, David J. Harris, Dave DeCaprio, Yanjun Qi, Anshul Kundaje, Yifan Peng, Laura K. Wiley, Marwin H. S. Segler, Simina M. Boca, S. Joshua Swamidass, Austin Huang, Anthony Gitter, and Casey S. Greene. "Opportunities and obstacles for deep learning in biology and medicine," J. R. Soc. Interface, vol. 15, no. 141, April 2018. Available: https://royalsocietypublishing.org/doi/pdf/10.1098/rsif.2017.0387.

Chintala, Soumith. *Deep Learning with Pytorch: A 60 Minute Blitz.* New York, New York: PyTorch, 2017.

Conzone, Samuel D. and Carlo G. Pantanot. "Glass Slides to DNA Microarrays," Materials Today, vol. 7 no. 3, March 2004, pp. 20 - 26.

Dowell, Karen G., Allen K. Simons, Zack Z. Wang, Kyuson Yun, and Matthew A. Hibbs. "Cell-Type-Specific Predictive Network Yields Novel Insights into Mouse Embryonic Stem Cell Self-Renewal and Cell Fate," PLoS ONE, vol. 8, no. 2, February, 2013.

Dyson, Robert D. "How Do Cells Sense Their Environment?" *Essentials of Cell Biology.* United Kingdom, Allyn & Bacon, 1978.

Fisher, R.A. "Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population," Biometrika, vol. 10, no. 4 May, 1915, JSTOR, www.jstor.org/stable/2331838. [Accessed November 4, 2019].

Gehlenborg, Nils, Sean O'Donoghue, Nitin S Baliga, Alexander Goesmann, Matthew A Hibbs, Hiroaki Kitano, Oliver Kohlbacher, Heiko Neuweger, Reinhard Schneider, Dan Tenenbaum, and Anne-Claude Gavin. "Visualization of omics data for systems biology," Nature Methods Supplement, vol. 7, no. 3s, March, 2010. Available: https://www.nature.com/articles/nmeth.1436.

Greene, Casey S., Arjun Krishnan, Aaron K. Wong, Emanuela Ricciotti, Rene A. Zelaya, Daniel S. Himmelstein, Ran Zhang, Boris M. Hartmann, Elena Zaslavsky, Stuart C. Sealfon, Daniel I. Chasman, Garret A. FitzGerald, Kara Dolinski, Tilo Grosser, and Olga G. Troyanskaya. "Understanding multicellular function and disease with human tissue-specific networks," Nature Genetics, vol. 47, no. 6, June 2017.

Hibbs, Matthew A., David C. Hess, Chad L. Myers, Curtis Huttenhower, Kai Li, Olga G. Troyanskaya. "Exploring the functional landscape of gene expression: directed search of large microarray compendia," Bioinformatics, vol. 23, no. 20, October 2007. Available: https://academic.oup.com/bioinformatics/article/23/20/2692/229926 [Accessed: October 31, 2019].

Kanehisa, Minoru and Susumu Goto. "KEGG: Kyoto Encyclopedia of Genes and Genomes," Nucleic Acids Research, vol. 28, no. 1, January 2000. Available: https://academic.oup.com/nar/article/28/1/27/2384332. [Accessed November 4, 2019].

Le, Quoc V. *A Tutorial on Deep Learning Part 2: Autoencoders, Convolutional Neural Networks and Recurrent Neural Networks*. Mountain View, CA: Google Brain, 2015.

Logan, Catriona Y. and Roel Nusse. "The WNT Signaling Pathway in Development and Disease," Annu. Rev. Cell Dev. Biol., vol. 29, July 2004. [Accessed: November 8, 2018].

Luber, Jacob. *Improved Prediction of Mouse Pathways Related to Bone Maintenance Through Machine Learning Utilizing Diverse Genomic Data*. San Antonio, TX: Trinity University, 2016.

Meyer, Ingmar Sören and Florian Leuschner. "The role of Wnt signaling in the healing myocardium: a focus on cell specificity," Basic Research in Cardiology, vol. 113, no. 44, October 2018. [Accessed: November 8, 2018].

Myers, Chad L., Drew Robson, Adam Wible, Matthew A Hibbs, Camelia Chiriac, Chandra L. Theesfeld, Kara Dolinski, and Olga G Troyanskaya. "Discovery of biological networks from diverse functional genomic data," Genome Biology, vol. 6, no. 13, December, 2005. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1414113/pdf/gb-2005-6-13-r114.pdf [Accessed: September 6, 2018].

Shneiderman, Ben. "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations," Proceedings 1996 IEEE Symposium on Visual Languages, September, 1996. Available: https://www.cs.umd.edu/~ben/papers/Shneiderman1996eyes.pdf. [Accessed November 4, 2019]

Troyanskaya, Olga G., Kara Dolinski, Art B. Owen, Russ B. Altman, and David Botstein. "A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)," PNAS, vol. 100, no. 14, July, 2003. Available: http://www.pnas.org/content/pnas/100/14/8348.full.pdf.

Troyanskaya, Olga, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert

Tibshirani, David Botstein, and Russ B. Altman. “Missing value estimation methods for DNA microarrays,” Bioinformatics, vol. 17, no. 6, February, 2001. Available: https://academic.oup.com/bioinformatics/article/17/6/520/272365.

Zhou, Jian, Chandra L. Theesfeld, Kevin Yao, Kathleen M. Chen, Aaron K. Wong, and Olga G. Troyanskaya. “Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk,” Nature Genetics, vol. 50, August 2018.

Zhou, Jian and Olga G. Troyanskaya. “Predicting effects of noncoding variants with deep learning–based sequence model,” Nature Methods, vol. 12, August 2015. Available: https://www.nature.com/articles/nmeth.3547.