Allegheny College Department of Computer Science and Biology

Senior Thesis Proposal

by Adam Cook

A Bioinformatics Approach to Verify Correlation Between Specific Genes in Cancer Tissue

Abstract

Computer science has allowed for a new approach to searching for cancer treatments. This study covers a bioinformatic approach to searching for interaction between specific genes in cancer tissue. It will be made up of 2 different sections, the first developing a tool in python to search for correlation, and the second testing specific gene interaction in yeast cells. The main goal of the tool will be to pull publically available data and perform statistical analysis to verify the correlation between genes. And in part 2, the main goal will be to determine survivorship advantage genes found with a positive correlation.

Introduction

Cancer is a disease of the genome. The genome can become unstable and expressed differently in cancerous cells. This genome instability can be due to a wide verity of reasons, and although there are several regulatory and reparatory mechanisms in place to manage them. One of the things that can influence whether or not a cell will become cancerous, is how these genes are expressed.

Gene expression is influenced by a number of internal and external factors which can be observed and studied. This is a difficult and time consuming process to determine normally. In order to speed up the discovery time to find gene correlations, there is room for a bioinformatic approach. There are several publicly available databases that can be accessed by anyone at GDC's cancer portal at: https://portal.gdc.cancer.gov. There are over 83,000 cases, and over 500,000 files covering almost 22,872 genes.

With the increase in computational power in recent years, combined in the recent databases with storage of entire genomes, it is becoming more viable to search for correlation between using computers. This senior project proposal aims to create or modify and enhance a tool to test and verify statistical correlation. Due to the recent advance in computing and data collection of gene expression in cancer tissues, new research can be conducted. With the increase of large scale datasets, false or misleading conclusions may be drawn from the data without verification. In order to make use of this information, tools need to be developed to verify the results. The first aim of my project will be to develop a tool that can be used to find correlations between specific genes in cancer tissue and verify it against another study that has done the same, and tested it in the wetlab. A tool like this will be able to provide a platform to do research looking at interactions rather than just how individual genes are expressed in cancer tissue. The second aim of my project will be to test the correlation between two genes that have been found to have a correlation in cancer tissue. This will provide testing to observe the interaction of the gene interaction found and provide further insight into the role that gene interaction plays. In addition, this wet lab also tests the tool developed in aim 1, providing some information that can be used to eliminate some threats of validity. The ultimate goal of a project like this is to advance our knowledge of genes and their relationship with each other. There are many benefits of developing a tool like this because of its potential application to search for gene correlation in cancer tissue, and the potential for future treatments that take this into account and can create targeted drugs.

Gene overexpression is a common trait of many cancers, resulting from the interac- tion of multiple genes that are mutated and mis-regulated. By screening for genes that are exposed in a positive or negative correlation with each other, we can find potential parallel pathways, inhibitors or promoter genes. We do not fully understand how all of our genes and their expression influence the other genes. By finding these correlations, we can then run tests to try to determine their functions. This would fill in part of our gap in knowledge for the relationship between certain genes.

By understanding the correlation between specific genes, it may lead to better outcomes for cancer treatment. This could lead to increased accuracy for gene targeted therapy and a better outcome for cancer patients. Better gene targeted drugs could be developed that specifically target one gene or one protein that is overexpressed in the cancer tissue, and allow treatment to be more targeted and specific. This would mean that a great number of the side effects of current treatments, such as chemotherapy, and irradiation would be minimized, and the damage to the host body would be greatly reduced. While these current solutions may be effective, they are certainly not perfect treatments. Patients would no longer have to risk damaging their body. Instead, they will know specifically what the drug will target that is causing the cells to proliferate. Majority of cancer cell lines rely on multiple mutation normal cells. There are several genes and proteins that are overexpressed and or underexpressed in order to allow them to grow and proliferate abnormally. There are natural regulators that are supposed to inhibit their growth, but in cancer cells this regulatory pathway is either ignored, or bypassed. In order to treat these cancer cells, a multifaceted approach is needed. Once the relation of more genes are known in relation to cancer, and additional pathways are understood, future treatments could be developed for specific patients.

This senior project proposal aims to create a tool that can determine correlations between sets of data using multiple means. Due to the recent advance in computing and genetic testing allows for screening of entire genetic data sets of cancer patients. Although there is still room for improvement in the field of analysis of this data. But in order to make better use of this information, tools need to be developed. The first aim of my project will be to develop a tool that can be used to find correlations between specific genes in cancer tissue. A tool like this will be able to provide a platform to do research looking at interactions rather than just how individual genes are expressed in cancer tissue. The second aim of my project will be to test the correlation between two genes that have been found to have a correlation in cancer tissue. This will provide testing to observe the interaction of the gene interaction found and provide further insight into the role that gene interaction plays. In addition, this wet lab also tests the tool developed in aim 1, providing some information that can be used to eliminate some threats of validity.

The ultimate goal of a project like this is to advance our knowledge of genes and their relationship with each other. There are many benefits of developing a tool like this because of its potential application to search for gene correlation in cancer tissue, and for future treatments that take interaction into account to create targeted drugs.

Hypothesis and specific Aims

Hypothesis: There is a positive correlation between *MMS21* and *SLX5.*

Aim 1: To develop a tool that can be used to verify correlation of gene expression after normalization of housekeeping genes.

Experimental Design

My project proposal will be to design a tool that can test correlation of genes using a multitude of different methods. There are other tools that can do some analysis on what is showing but do not test for correlation between genes in specific cancers. One example of a similar project is DGCA, a comprehensive R package for gene correlation Analysis but does not have any means to pull data from online databases (McKenzie, Katsyv, Song, Wang, & Zhang, 2016). My proposed tool could help identify interactions between genes, providing further insight into cancerous cells, potentially leading to more targeted therapies to be developed.

Statistical analysis will need to be done on the data collected. One approach would be clustering analysis, once the data is classified. This can be done by determining the classification of the forms related to the organization of data and then performed to search for the goodness of fit (Saraçli, Doğan, & Doğan, 2013). Another cluster analysis is based on the findings of Hitchcock et al. Ward's method for testing clustering data and will provide the best platform for searching for average linkage and correlation (Ferreira & Hitchcock, 2009). To optimize Correlation analysis, Deep Canonical Correlation analysis can be done. This process uses two deep networks to search for maximum correlation (Andrew, Arora, Bilmes, & Livescu, 2013). These tools will need to be implemented or similar libraries accessed for python.

Expected outcomes

The outcome of this project will be to develop a tool that is able to complete data pulling, sorting, and performing a host of statistical tests to determine correlation. In addition, it should be able to output to the user results based that show specific gene correlation.

Aim 2: Determining if the overexpression of *MMS21* provides a survival advantage for cells with *SLX5* overexpression

RNF4 and *MMS21* have been found to be overexpressed in derived patient's tumors Bonham-Carter & Thu, 2020). The human ortholog of *SLX5* is *RNF4*, involved in E3 ubiquitin-protein ligase to mediate protein degradation (Van Hagen, Overmeer, Abolvardi, &

Vertegaal, 2010). *MMS21*, another E3 ubiquitin-protein ligase is involved in the DNA repair pathway for the removal of UV-induced DNA damage in Homologous Recombination (Zhao & Blobel, 2005). Both of these have been found to have a positive expression correlation in *SLX5* and *MMS21* genes (Bonham-Carter & Thu, 2020).

Experimental Design

The main goal of the wet lab portion of this study will be to determine if there is a positive correlation between *SLX5* and *MMS21* genes. The positive correlation will be supported if there is a survivorship advantage when *SLX5* and *MMS21* genes are both overexpressed. Overexpression (OE) occurs giving rise to an overabundance of a specific protein. This can be done through a mutation of the gene to promote its OE. These genes have been found in patient tumors, but in order to do testing of OE, another method is needed. Working with genes can be difficult, and that is why we will be using a model organism that allows us to more easily mutate and timely do tests.

Saccharomyces cerevisiae will be used as our model organism because the genes of interest are conserved (serve the same function in yeast) as in humans. *S. cerevisiae* provides the ability to test genes and their roles in a timely manner (Botstein & Fink, 2011). Additionally, it's entire genome has been sequenced and can be found publicly.

In order to alter the expression of the selected gene, transformation will be used. Transformation is the cell's ability to take up foreign DNA, and incorporate them into their genome. Prokaryotes and some eukaryotes have circular DNA. Yeast is one that has this circular DNA, making it easier to Bacteria and yeast. The expression of selected genes (*SLX5*, *MMS21*) will be added through Plasmid: pRS423, and Plasmid: pRS425. This process is done through replicating the genes of interest in bacteria, and then going through a process to transfer the gene of interest into the yeast cells to have it taken up.

Once we have successful transformations for our mutant *Slx* and *Mms21*, we will use a spotting assay to determine the growth rate. A spotting assay uses 10 fold serial dilutions to transfer ranging concentrations of cells onto agar plates with growth media using a 96 pin device called a frogger. This plates the same amount of cells for each phenotype, at a decreasing concentration, allowing for assessment to determine growth rate. Images will be taken after 1, 2, and 3 days.

In order to test for survivorship advantage, cells will be exposed to sources of DNA damage. One test will induce DNA damage through UV radiation, and another test will induce it through Methyl MethaneSulfonate (MMS). Cells will be exposed to varying amounts of UV radiation, ranging from 50 J/M^2 to 150 J/m^2. This will cause DNA damage, and impair their growth. If the cells are not able to repair themselves, we are expecting to see impaired growth and reduced colony counts. One way to check the growth rate between different phenotypes of S. Cerevisiae is by using a spotting assay.

Expected outcomes

It is difficult to predict the outcome of the results, but we are expecting to see a difference between when both are overexpressed and when there is only one gene expressed. The act of the cells taking up foreign DNA and undergoing the process also may differ how the cells react to external DNA damaging factors. In addition, there may be additional pathways that we do not fully understand that also play a role in how these different mutants react. Although, if things are able to be modeled for correlation, then we are expecting that when both genes are overexpressed this will lead to an increase in sensitivity of those mutants. In those that only have *Slx5* or *Mms21* mutation, we are expecting to see a decrease in sensitivity when compared to the double overexpressed.

Background & Significance (Related Work)

Cancer starts when cells in an organism begin to grow out of control. There are large differences in cancer, all starting from the uncontrolled growth of malfunctioning cells. It is a systemic problem of our bodies that millions are currently facing, causing the second number of deaths in the world after cardiovascular disease.

Cancer normally develops from normal cells due to DNA damage that accumulates. Now that people are less likely to die young due to causes such as war and famine, the average length of life has increased dramatically. This gives time for mutations to accumulate. Cells have mechanisms to repair their DNA, or they can kill the cell if the damage is severe enough, but the damage is not always repaired correctly (Sancar, Lindsey-Boltz, Ünsal-Kaçmaz, & Linn, 2004). This can lead to further mutations that can lead it to several of the “Hallmarks of Cancer”. These hallmarks are present in cancer cells, and understanding their interaction between genes can help us suppress the ability of cancer cells to proliferate (Hanahan & Weinberg, 2011).

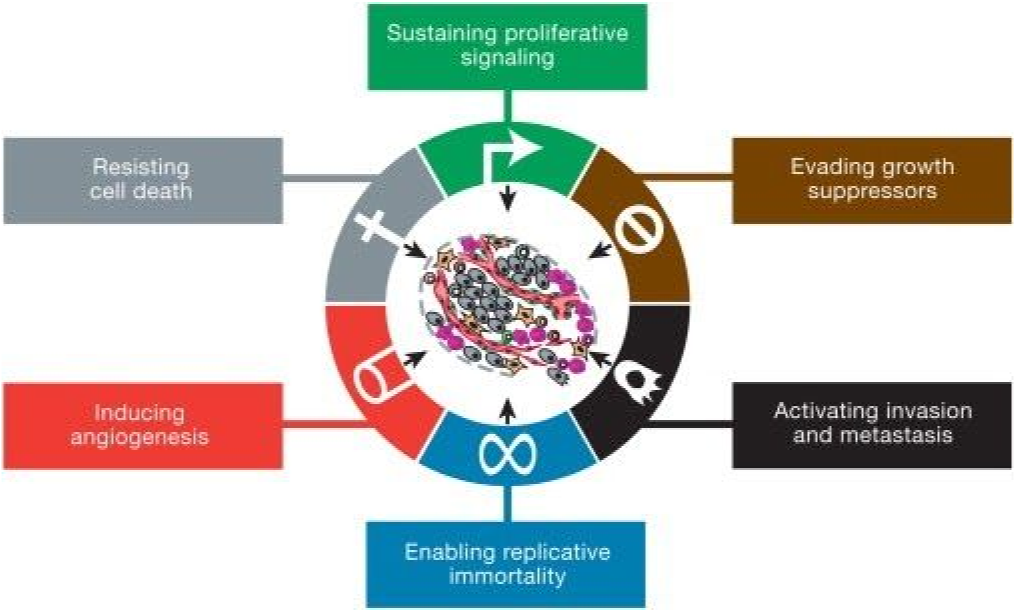
Sustaining proliferative signaling
Evading growth suppressors
Activating invasion and metastasis
Enabling replicative immortality
Inducing angiogenesis
Resisting cell death

Figure 1. Hallmarks of Cancer - Several enabling features of cancer cells. Gene interaction may play a role in how these hallmarks are enabled.

The two genes that we are interested in this study are *SLX5 and MMS21.* These two are involved in evading growth suppressors, as they act as code for cell cycle checkpoint proteins

(Zhang, Roberts, Yang, Desai, & Brown, 2006). They have also been found by OBC and Thu to be found in a positive correlation in breast cancer tissue (Bonham-Carter & Thu, 2020).
When DNA damage occurs, it can lead to stalled replication forks and cell cycle arrest. There are normal checks within the cell cycle to prevent the cells from continuing with damage. One condition of cancer cells is their ability to continue through these checkpoints. Some specific proteins encoded by GS1, TOP3, MUS81, MMS4, SLX1, SLX4, SLX5/HEX3, and SLX8 allow for this progression. We are studying the interaction of *SLX5* and *MMS21* in allowing normal cell cycle progression, one of the hallmarks of cancer, observed by a potential survivorship advantage. (Zhang et al., 2006).

There is currently a gap in our knowledge of how specific genes play a role in cancer. This gap has been shrinking with significant progress made in new technologies that allow us to sequence genes leading to personalized medicine. In addition, high throughput sequencing, database storage, and server computing have brought new testing opportunities. About 20 years ago, the trend of computational methods increased, and the use of databases and the internet began to increase. This laid the groundwork for much of the bioinformatic tools developed today (Perez-Iratxeta, Andrade-Navarro, & Wren, 2007). But in order to advance our understanding of cancer tissue, and develop better treatments we need to understand how specific genes interact with each other.
Going into the future, there will be a need to develop more tools that can be used to study these gene interactions. Computers can allow us to study the expression of hundreds of genes in cancer patients, and databases are now available to do so. Collection of data has increased over recent years, and now that this information is becoming available future treatments can be developed using this information. Tools like this one will provide a fundamental tool for searching and discovering interactions between genes, and developing targeted treatments to help treat cancer. This will be a significant part of this project. Developing a tool to further study interactions. This tool will help fill in this gap of knowledge by providing a creative solution to searching for correlational information in cancer tissue. By utilizing Python and the available libraries, we can use high-level computer algorithms allowing for discovering additional gene interactions that can have a role in future treatments.

Research Design and Methods

Our goal is to create a tool that can be used to determine the correlation between gene expression and to determine if there is a survivorship advantage for cells with *MMS21* and *SLX5* overexpression. This makes sense by allowing us to collect data, do analysis on it, and then test our findings.
In Aim 1 of this project, we are developing the python tool. In order to collect data for this tool, we will be accessing an in-depth publicly available database. The Genomic Data Commons (GDC) is maintained by the US Government and contains information on Cancer patient's cancer information collected. The GDC is a state-of-the-art pipeline that has evolved over time as more information has been added. More and more features are added to help utilize the increase in information. The GDC Data Portal allows the user to view and interact with data. Summaries can be viewed based on objects, Primary sites (hosts), cases, files, genes, and

mutations. In order to work with the GDC, you can view online, or use their API. They have proper documentation about how to interface with their API on their webpage. The GDC uses JSON, a file format that stores its data in simple data structures and objects, as its format and standard HTTP methods. This will make it much easier to interact and work with and opens the possibility to work with the data through JavaScript.

Python will be the primary language that will be used in this project due to its wide adaptability and high functionality. There are a large number of additional features that can be added to it through libraries that allow for additional functionality. In order to handle the data collected, two different libraries can be used. Pandas can be used for data manipulation, analysis, and cleaning of data. Numpy is another package that can be used for scientific computing with python, including features like N-dimensional array objects and linear algebra manipulation. These packages also allow for dynamic manipulation of the data that may come in use for additional calculations. Python was not originally designed for data computing, and so the use of these additional libraries and packages should allow for scientific manipulation and testing. The python library Scikit allows for predictive data analysis and implementation into Matplotlib and NumPy. includes a linear regression model that can be used to determine R^2 values. This approach should also provide an accurate and repeatable testing method. In addition, A focus should be placed on exception handling that way this tool can be used on other data sets as well. The amount of research and data is constantly expanding, so a tool to determine significance can provide additional help for other data sets where it would not be feasible to do calculations by hand.

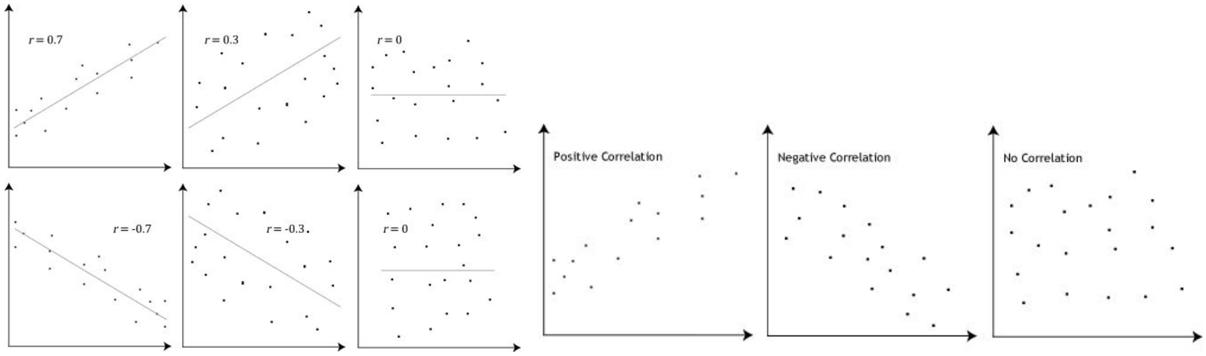

Figure 2a,b. Pearson correlation coefficient - Pearson Product Moment Correlation Coefficient is used to measure the strength of a linear association between two variables. R stands as the Pearson's correlation coefficient (Lard Statistical Guides).

In order to test for correlation, Pearson Product Moment Correlation Coefficient or Spearman’s rank correlation coefficient can be used. Pearson's correlation coefficient can be used when both variables are normally distributed, and Spearman’s when the data is skewed (Mukaka, 2012). Pearson correlation coefficient tests the statistical hypothesis that there is no relationship between the means of two different groups. It will allow us to determine the correlation, or the relationship between two different genes on a scale from no correlation (0), strong positive correlation (+1), or strong negative correlation (-1) or anything in between.

This information only shares with the correlation between genes. It does not provide information with whether or not there is causation between different genes, or why the interaction is taking place, only that it is a correlation.
In order to ensure the data collected through Aim 1 is valid, verification and replication need to be done. One means of testing would be done through Monte Carlo comparison, running through several simulations and testing results of similar statistical analysis to ensure that the results are constant.

For Aim 2 of this project, we are testing the hypothesis to determine if there is a survivorship advantage for cells with *MMS21* and *SLX5* overexpression. In order to do this, we will perform spotting assays testing the growth rate of several mutants of *mms21* and *slx5.* The data collected (colony count) from these spotting assay will be analyzed using JMP Pro, and ANOVA tests. This testing should be able to show if there is a survivorship advantage present.

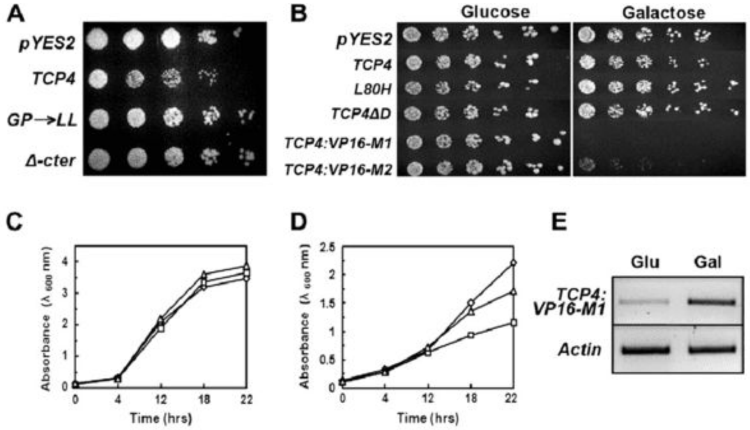

Figure 3. Example Spotting Assay and analysis. Showing assay and example graphs of growth rate connection over time (Aggarwal et al., 2011).

Verification

Verification is a large portion of this project. In order to verify the finding in this study, we will be comparing it to a similar project done by Thomas et. al in “Genetic interaction mapping and exon-resolution functional genomics with a hybrid Cas9–Cas12a platform” This study uses breakthrough technology in gene editing to verify my modifying specific genes. In addition they have created deep learning models and several other screens to map and verify gene pairs.

Expected Outcome

For Aim 1, we are expecting to have developed a fully functional program, either a stand-alone or one that can be executed through the terminal. For Aim 2, we are expecting to see increased sensitivity of cells with *MMS21* and *SLX5* overexpression. This would be observed through less total number of cell colonies when compared to the wild type when exposed to DNA damaging agents. This would support our hypothesis that there is a positive correlation between *MMS21* and *SLX5*, leading to a survivorship advantage. We expect to see this with an increased number of colony count, because if there are more colonies that form, that shows that the cells are more likely to survive despite levels of DNA damage, meaning that they are either able to repair the damage and continue or bypass damage checkpoints.

Conclusion

In conclusion, this project is looking to test correlation studies within genetic datasets, and help the further study of gene interaction in cancer tissue. The second part of this project is to test the relationship of found genes with a positive gene correlation in order to support or refute the relationship correlation.

The possible impact of this research is better treatment for cancer through targeted gene therapy. Done through understanding the interaction between the genes expressed in relationship to each other, drug therapies can be developed that inhibit or promote the expression of specific genes to help suppress the hallmarks of cancer.

Some open issues that we may face while working on this project include the implementation of different libraries in order to search for gene interaction, normalize data, and develop a platform that is usable. In order to create mutants to test our specific genes, there is a difficult process to enable us to create *S. cerevisiae* strains that express each of the different gene expressions. This process will be time-consuming, expensive, and difficult but should be able to be done within a year, following a consistent schedule of work. By having previous experience working in the lab, the majority of time will be spent working on the project, rather than learning the process to follow. Allegheny College has all of the tools needed in order to complete the development of the tool, and the majority of the tools to do the wet lab portion as well. All of the additional tools should be available for an affordable price within the time frame.

Research Schedule

Task	**Begin Date**	**End Date**
GDC Data Collection	Early Sept.	Late Sept.
Data Formatting / OE sequence	Early Oct.	Late Oct.
Correlation Regression Testing	Mid Oct.	Early Nov.
Stat Package Implementation	Mid Dec.	Late Dec.
Verification of Results	Mid Dec.	Early Jan.
Generation of OE Mutants	Early Jan.	Late Jan.
Testing of Mutants & Program	Early Feb.	Late Feb.

Table 1. Research Schedule with proposed dates for project necessary tasks.

Future Direction

Future work on this project could include developing an option to run different statistical tests in addition to correlation regression testing. In addition to the web based platform, an application could be built that could run on Windows, Mac, and Linux. Of course, testing of additional different genes data sets and follow up correlation studies should be done to confirm with other correlation studies to confirm our methods.

Reflections on Improvement

The application of this project can have a great impact, but it still needs further work and development. I have struggled writing this proposal, from working on each of the different sections to putting it all together. Going forward, I should make it a greater priority to design an outline and organize from the beginning. It makes it much easier to follow along and clearly convey a message. I have gotten worse at planning ahead when compared to earlier in the semester at school. Once home, I struggled to sit down and finish the work before it was due. This could be improved by having a specific hour each day to work, rather in between other jobs. Based on feedback of others, I need to work on the transition of thoughts and ideas. This is difficult but is something that I will need to work on throughout this proposal and other writings. I feel like there is a lack of transitions between the majority of ideas and paragraphs. Transitions are something I can develop more, but for outside tools that have helped, there have been a few that helped immensely for organizing and writing. Using RefWorks, and GoogleDocs helped for keeping track of articles and citing them. GoogleDocs has been instrumental in writing and collaborating, with its comment and collaboration features. In the future, I plan on using these two more in-depth, along with better time management and deadline following.

For changes made on the proposal, I have changed the way I want to approach the data analysis. At first, I thought it would be fine to write something to test for correlation myself, but this would be very time consuming and would produce worse results than using available scientific python libraries.

References

Aggarwal, P., Padmanabhan, B., Bhat, A., Sarvepalli, K., Sadhale, P. P., & Nath, U. (2011). The TCP4 transcription factor of arabidopsis blocks cell division in yeast at G1→ S transition. *Biochemical and Biophysical Research Communications, 410*(2), 276-281.

Andrew, G., Arora, R., Bilmes, J., & Livescu, K. (2013). Deep canonical correlation analysis. Paper presented at the *International Conference on Machine Learning,* 1247-1255.

Bonham-Carter, O., & Thu, Y. M. (2020). Systematic normalization with multiple housekeeping genes for the discovery of genetic dependencies in cancer. *bioRxiv,*

Ferreira, L., & Hitchcock, D. B. (2009). A comparison of hierarchical methods for clustering functional data. *Communications in Statistics-Simulation and Computation, 38*(9), 1925-1949.

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell, 144*(5), 646-674.

McKenzie, A. T., Katsyv, I., Song, W., Wang, M., & Zhang, B. (2016). DGCA: A comprehensive R package for differential gene correlation analysis. *BMC Systems Biology, 10*(1), 106.

Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal, 24*(3), 69-71.

Perez-Iratxeta, C., Andrade-Navarro, M. A., & Wren, J. D. (2007). Evolving research trends in bioinformatics. *Briefings in Bioinformatics, 8*(2), 88-95.

Sancar, A., Lindsey-Boltz, L. A., Ünsal-Kaçmaz, K., & Linn, S. (2004). Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. *Annual Review of Biochemistry, 73*(1), 39-85.

Zhang, C., Roberts, T. M., Yang, J., Desai, R., & Brown, G. W. (2006). Suppression of genomic instability by SLX5 and SLX8 in saccharomyces cerevisiae. *DNA Repair, 5*(3), 336-346.

Efron, Bradley. "Correlation and large-scale simultaneous significance testing." Journal of the

American Statistical Association 102.477 (2007): 93-103.

Oliver Bonham-Carter and Yee Mon Thu, GenExSt: A Tool to Identify Correlation of Gene Expression after Normalization with Housekeeping Genes (TBD), Allegheny College

Pearson Product-Moment Correlation. (n.d.). Retrieved May 6, 2020, from https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

GDC. (n.d.). Retrieved May 6, 2020, from https://portal.gdc.cancer.gov/

NumPy¶. (n.d.). Retrieved May 6, 2020, from https://numpy.org/

Learn Scikit. (n.d.). Retrieved May 8, 2020, from https://scikit-learn.org/stable/