Final Deliverable

By: Ernest Wang

Colorectal Cancer

Colorectal cancer is cancer that is found between the areas of the colon and the rectum. It is the third leading cause of death among the cancers with anyone over that age of 50 at risk. Some of the known risk factors include family history, dietary malpractices, lack of physical exercise, and obesity.

Pipeline of Guo's Paper.

1. Found datasets which described the expression of genes for both cancerous and normal tissue
2. Filtered out the DEGs using statistical tools
3. Took phenotypic data into account
4. Screening related mRNA and lncRNA based off the hub genes
5. Construction of a ceRNA network
6. Statistical analysis

Our Pipeline

1. Selected a data set that had an even number of both cancer and noncancer samples.
2. Normalized the data set and determined & excised the outliers.
3. Filtered the genes for differentially expressed genes.
4. Took phenotypic data into account.
5. Mapped the genes to specific functions/pathways using databases such as enrichR and David

Uncertain Execution

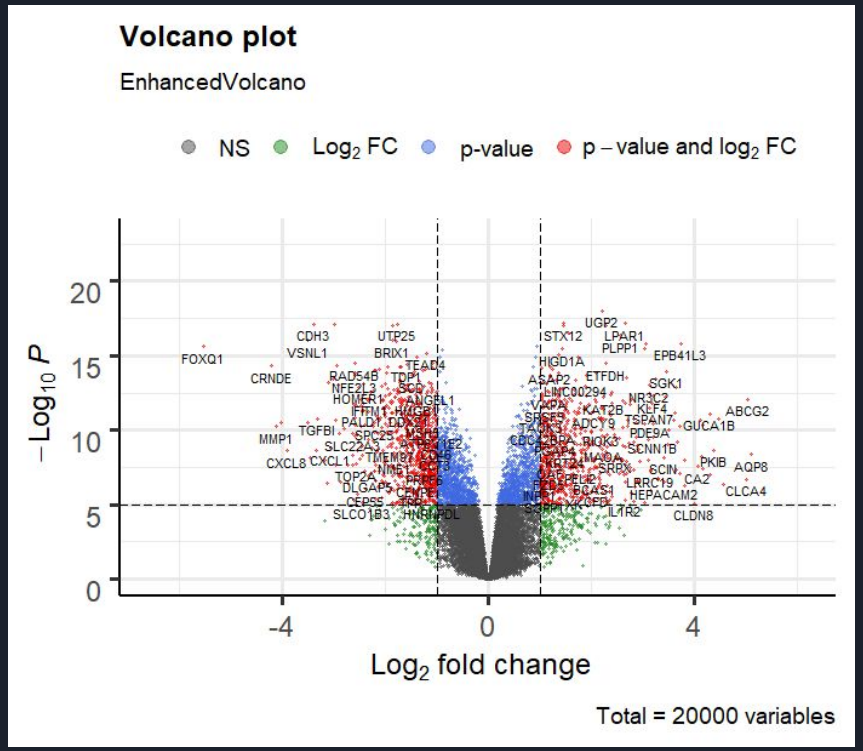
The guo paper gave us an overview of all the steps that they took but they did not include many of the smaller details.

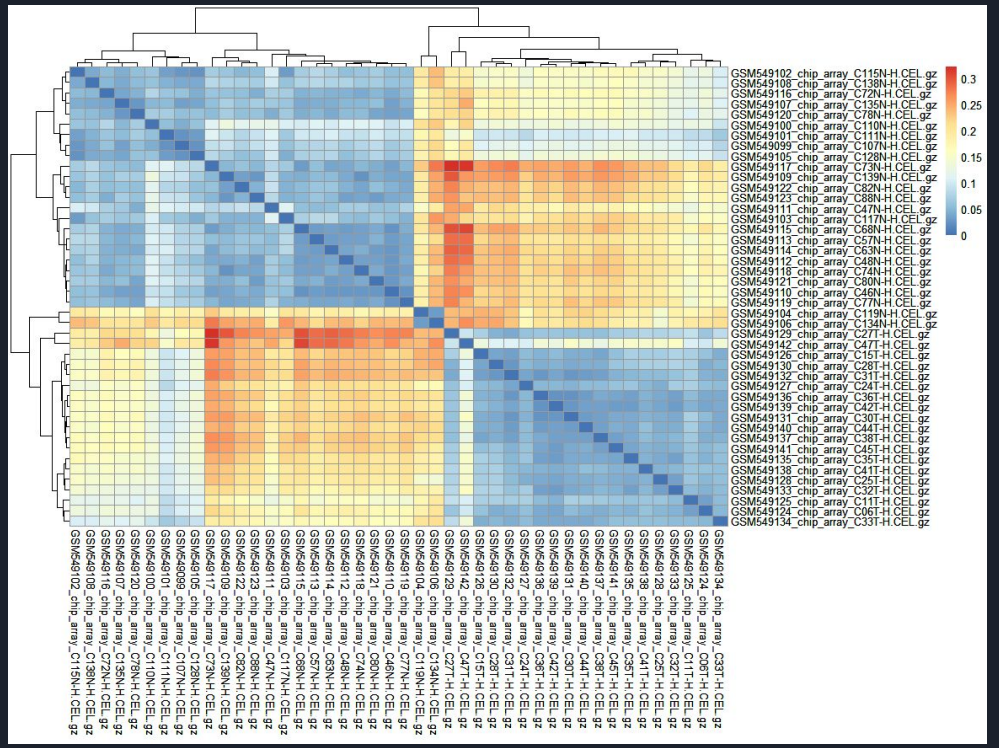
Details such as

1. Normalization method
2. Quality control methods
3. Exact functions used with in the packages

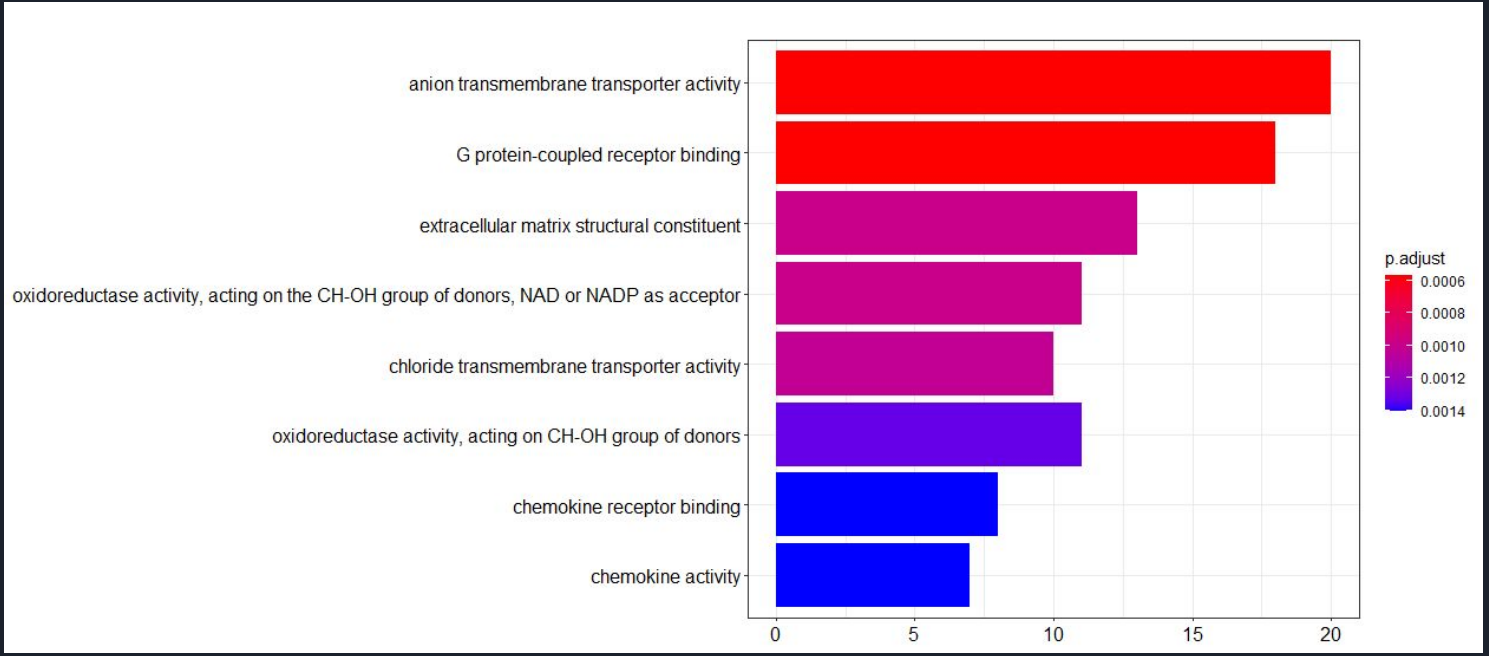
This lack of information makes it a little harder to reproduce to verify the exact results, but it gives us a certain freedom to choose our own way to analyze the data set.

Analysis of GSE21510

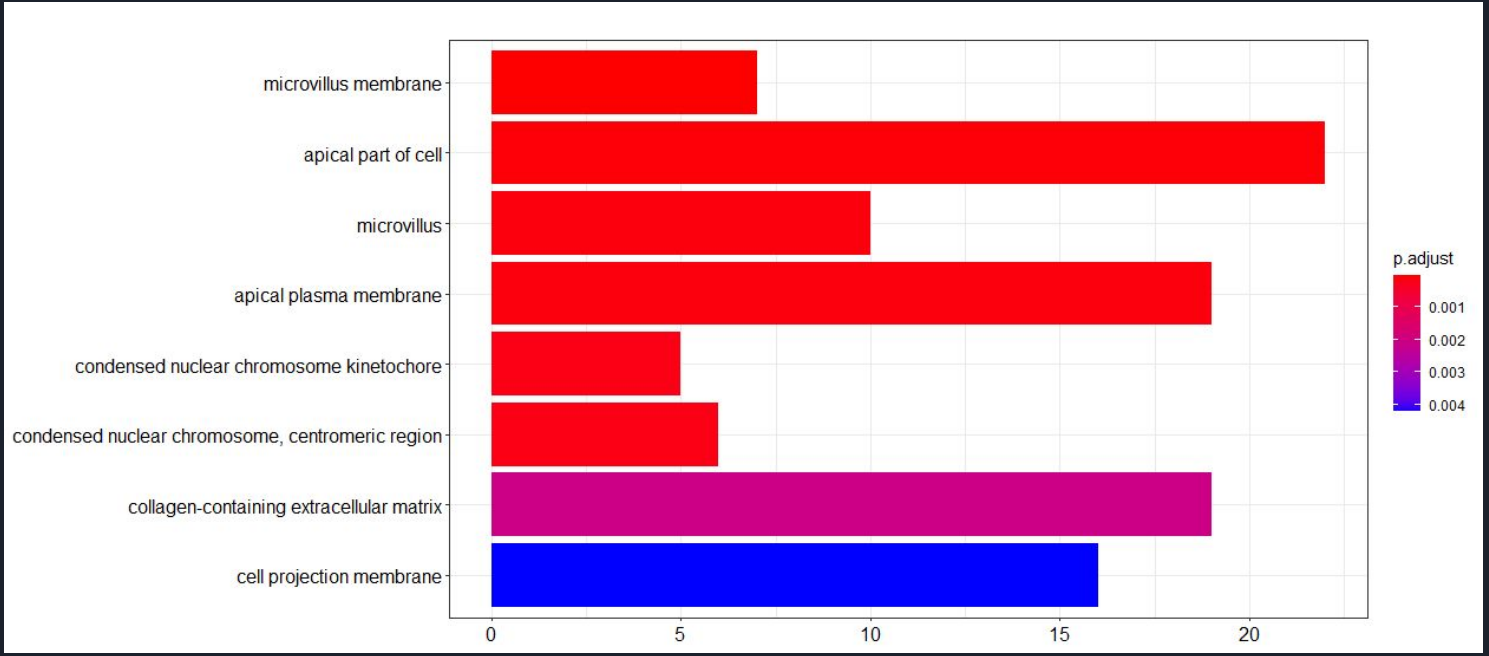




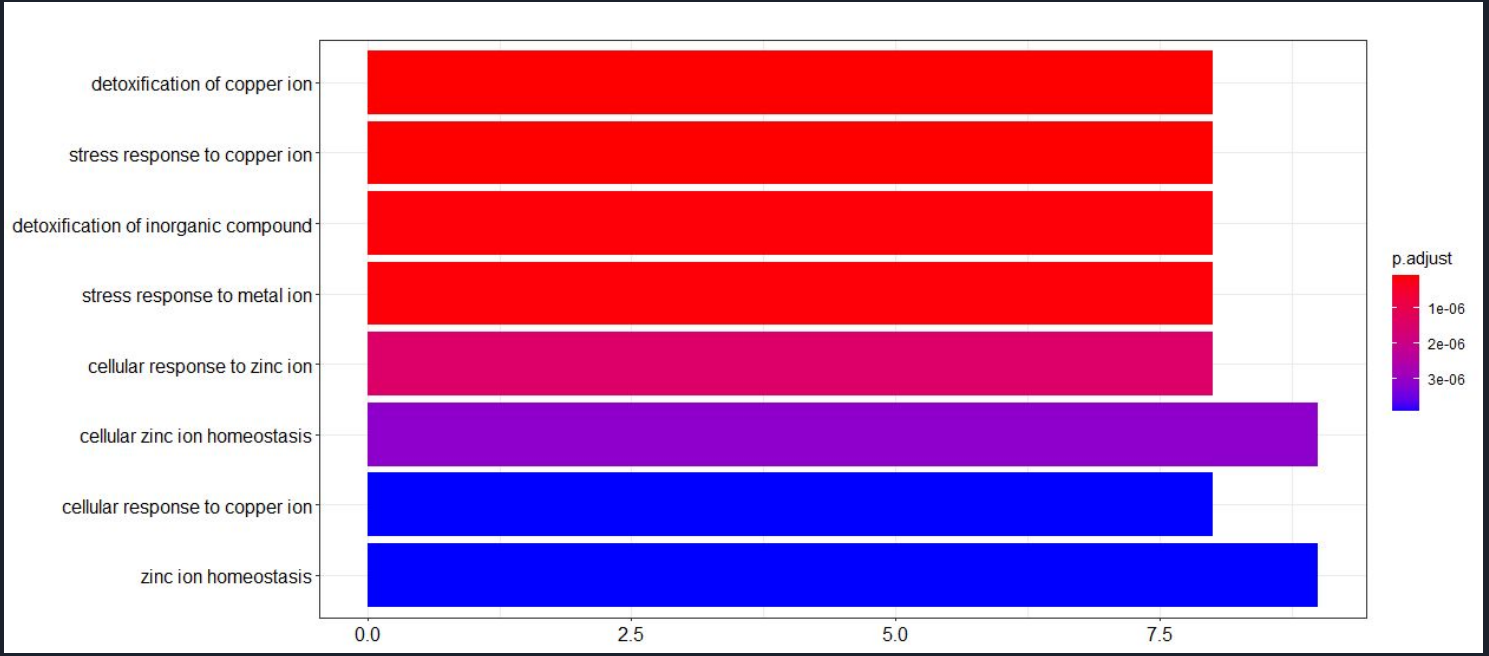
Functional Analysis of GSE21510(MF)



Functional Analysis of GSE21510(CC)



Functional Analysis of GSE21510(BP)



Base R Code

Learning R code is a fundamental aspect within the project, and it was one of the hardest aspects to master.

Challenges:

1. Developing syntax habits
2. Looking for help when working with packages

Problem Solving:

Group 4 solved many of the R problems as a collective. We shared information and possible solutions. Through vetting and trial and error, we completed each task to the best of our abilities. There was not a single task that was too difficult to complete as a group.

Packages

In addition to R code, R packages played and important role in the processing of data and analysis of data. Packages are groups of functions the form the basis of our analysis. We used a plethora of packages, including Affy, Biobase, AnnotationDBI, Limma, and etc…

Challenges:

1. Official documentation of the packages were all professional level and there were little to no secondary sources that could simplify the information.
2. The interplay between different packages

Problem Solving:

Much of the same base R problem solving methods were used, being collaborating. We all had our different interpretation of the documentation. Through meeting and discussing, we developed a final interpretation that could be seen in our code.

Quality Control

The first step of the process is to perform quality control on the dataset. This is to ensure that outliers would not significantly disrupt further analysis. Group 4 had to use Array quality metrics to perform the quality control.

Challenges:

1. Interpreting the graphs generated by array quality metrics.

Problem Solving:

Generating the graphs was relatively simple and streamlined, but interpreting the graphs required us to research each graph's function and what it displayed.

Volcano Plot

Volcano plots demonstrate the genes in respect of the log fold change and p value. It is a simple way to demonstrate significant genes that are above a certain threshold for both log fold change and p value.

Challenges:

1. Preprocessing data using limma

Problem Solving:

The main roadblock when creating the volcano plot was the processing of the data to ready it for plotting. We ran into problems with the formats of the data and the proper input of the limma functions. It was through the documentation and collaboration where we figured out how to properly format the code.

Separating Phenotypic Data

By separating the phenotypic data, we have the potential to look at effects such as size of sample, location sample was taken, and date the sample was taken.

Challenge:

1. Properly setting up the code to extract the phenotypic data

Problem Solving:

Separating the phenotypic data was not as hard compared to other tasks, but it did require some research to properly complete the task.

Functional Analysis of Genes

We used topGO, org.Hs.eg.db, and clusterprofiler to perform a form of functional analysis to determine what roles the genes have within the larger picture of human function.

Challenges:

1. GSEA was extremely difficult to figure out
2. enrichR input was not easy to figure out

Problem Solving:

I relied more on my group as I was not able to figure out GSEA from the given documentation. EnrichR was difficult to figure out but with the limited outputs our code has it was easy to just try all of them and see which ones worked.

Moving Forward

- We did not complete the entire process the Guo paper covered, so the next group could take our data and move forward with it
- This methodology could apply to different cancers, cancers that have not been studied like Guo's take on colorectal cancer

Personal Take

The wide accessibility of data and relatively low operating bar allows for many people to be able to access and interpret lab data without the need of an official licence. This allows for any person to contribute towards the common goal of understanding a disease.