Shafaqat Rahman (Observer)

**Final Deliverables (Bioinformatics June Session)**

**1.**

The first part of the pipeline requires reading in the raw data. In this paper, sequencing data of cancers were obtained from the “TCGAbiolinks” package and over 500 individuals were collected. In terms of determining DEGs common to colorectal cancer (CRC) samples, CRC datasets were obtained from GEO. These datasets are GSE21510 which contains 44 normal and 104 tumor samples and GSE8671 which contains 32 normal and 32 tumor samples.

The DEGs were determined by isolating genes that had log(FC) > 1 and an adjusted p-value < 0.05. However it is not clear what normalization and processing was done to clear out null values and issues in the dataset. DAVID was used to find clusters of biological function for these common DEGs and KEG pathway analysis was also done. Significant GO terms were determined. The PPi networks were constructed using the STRING database and were created with up and down-regulated genes. These parts of the pathway were learned through this internship, and I’m familiar with how to code these parts in R studio.

The next aspects are things that are new to me and were not extensively mentioned in the internship. Survival analyses were used to estimate correlations of the candidate genes with prognostic data such as survival status, cancer stage and grade, survival time, and molecular subtype. We did a similar kind of analysis (phenotypic data analysis) to see correlations between genomic data and phenotypic data but unsure if it's a synonymous term with survival analysis. We did not do this in the internship, but in the pipeline, miRNA:mRNA interactions were collected from a database called StarBase ad miRNet to identify lncRNAs that correlated with cancer prognosis. An R package called “networkD3” was used to construct a ceRNA network based on network relationships between hub genes, interacting miRNAs, and associated lncRNAs.

**2.** The paper fails to mention aspects of quality control that are imperative to analyzing the dataset. Normalization techniques are not mentioned (mas5, rma) and it is unclear whether a linear model was fitted to the dataset (which can be achieved by using the limma package). It’s also unsure how the investigators assessed the quality of array intensities, and fails to mention using tools such as RLE and NUSE.

**3.**

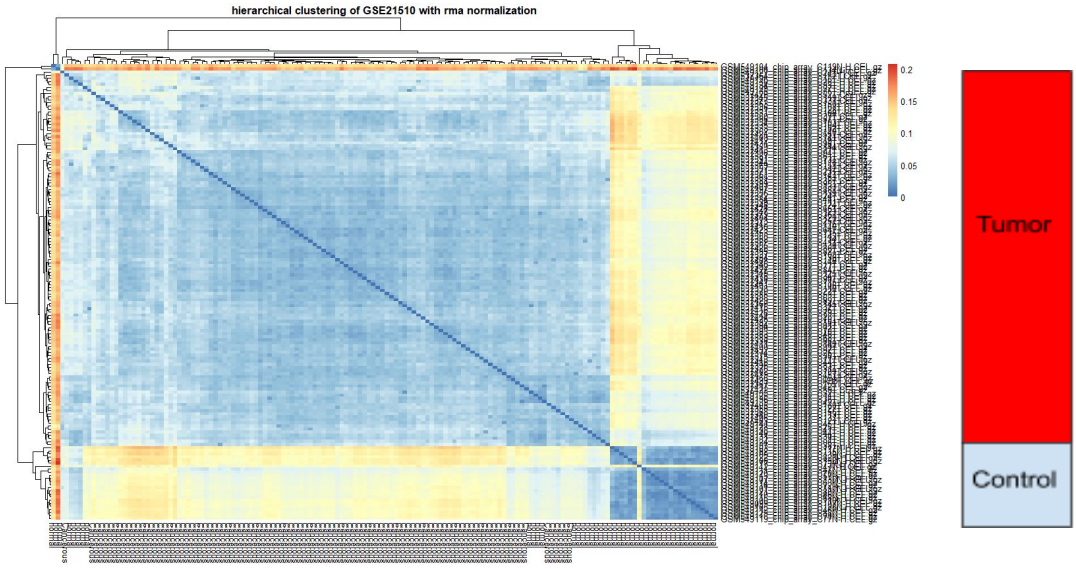


***Figure 1:*** *Hierarchical clustering map of GSE21510 data post rma normalization*

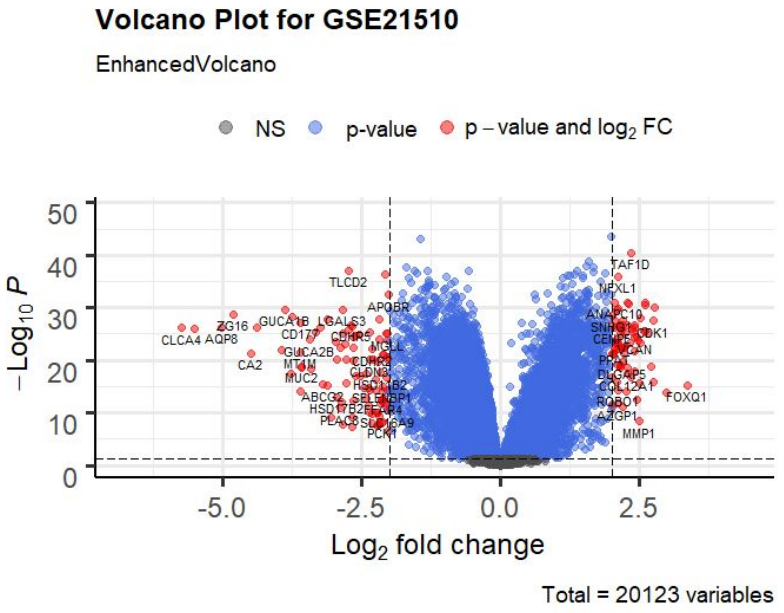


***Figure 2:*** *Volcano plot of GSE21510 data with p-cut off < 0.05 and |log(FC)| > 2. In this data set, 172 differentially expressed genes were determined based on these cut offs, where 75 are upregulated DE genes and 97 are downregulated DE genes.*

**4.**

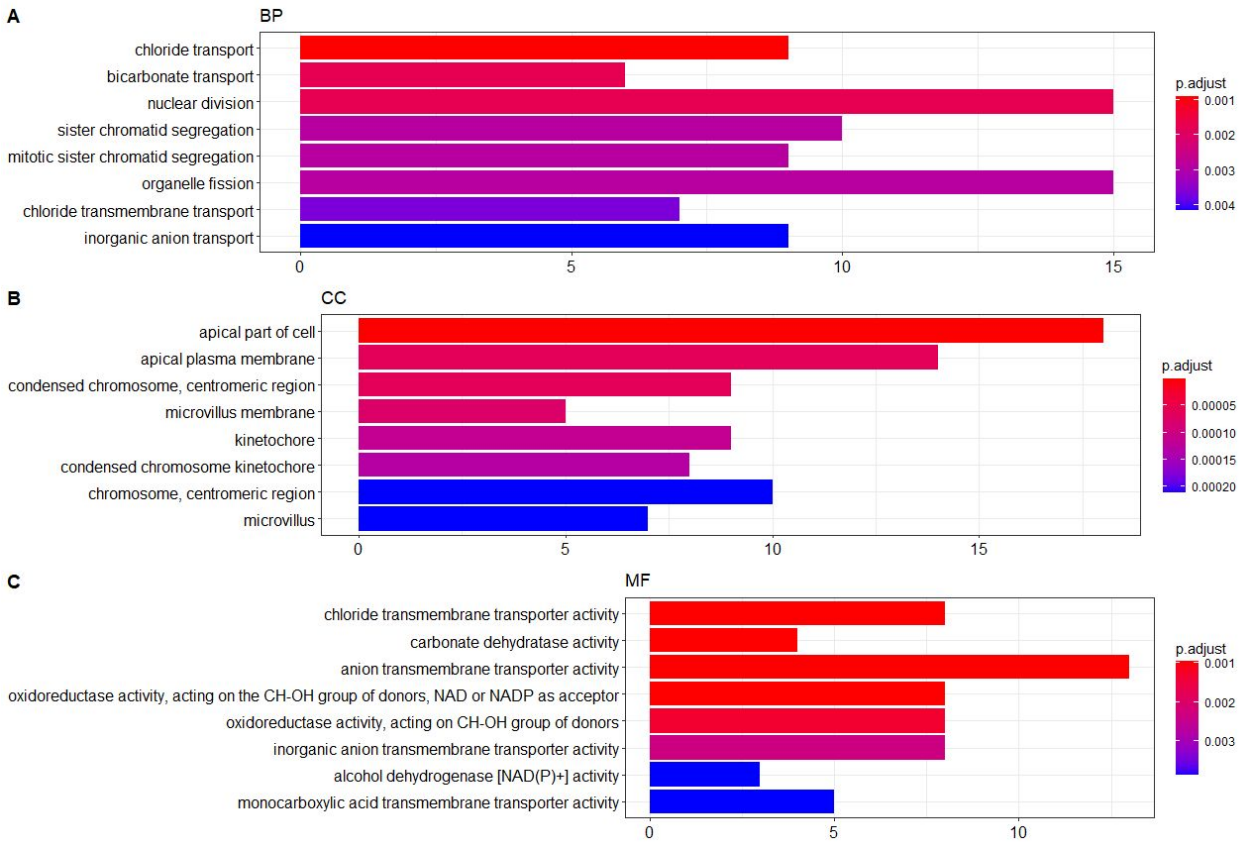


***Figure 3:*** *EnrichGO barplots of GSE21510 data clustered by Biological Process (BP), Cellular Components (CC) and Molecular Function (MF)*

The answer to number 4 isn't an exact replica of Figure 2A from the Guo paper, but it was the closest I could achieve to getting something similar.

**5.**

**Things Learned:**

- More holistic understanding of relevant biological terms (i.e. miRNA, mRNA, lncRNA, ppi networks) as they relate to the discussed Guo et al. paper on CRC prognostic markers
- Stronger understanding of how DEGs are determined and how they're later fed into network analyses for determining higher order trends in biological processes and cellular location.
- Clearer conceptual understanding of how the programming techniques I learned through this internship can translate to determining genomic differences between normal, healthy data and diseased-state data like cancer.

- An awareness for limitations in bioinformatic analyses and how to properly read a scientific paper in bioinformatics.

**Challenges**

- Programming in R has been an issue with my old PC laptop and its limited memory capacity, but I've been able to make it work.
- Some of the higher order analyses techniques to study miRNA interactions are still a bit difficult to understand, but that's also because we didn't have time to properly dissect them through the R exercises.

**6**. It would be beneficial if future sessions of the internship would cover what flow cytometry is and how to analyze flow cytometry data. Another avenue for study would be molecular docking and studying the binding of small molecules with simulation software. In this internship we covered principal component analysis, but there are many other ways to determine groups of data with similar variability. One could also look at non-negative matrix factorization or other source separation techniques.

**7.**

a. I've developed an intermediate level understanding of R as it relates to doing bioinformatics. I also know three ways to make volcano plots in R...
    i. volcanoplot()
    ii. EnhancedVolcano()
    iii. A more tedious method that requires using the plot() function and isolating the logFC values and adjusted p-values from the eBayes structure)
b. Stronger understanding of how the biological interactions/expressions of RNA sequences can be exploited in a chip/array system to detect for macro-level trends in a data set, and how to determine differentially expressed genes post normalization and linear-model fitting.
c. Developed a familiarity with isolating groups of genes based on their biological process, cellular location, and molecular function, through analyses techniques such as DAVID, STRING, and wikipathways GSEA.

**8.** I've learned that the field of Bioinformatics is ever-growing and there's so much more to learn, since the concepts learned in this internship only scratch the surface of what a scientist can do with these tools. I look forward to constantly learning and expanding my bioinformatics background that I've developed through STEMAWAY's internship.