Brianna Campbell

Thesis:

This project focuses on the comparison of expression between high-impact mutations and no mutations in the gene LRP1B in the primary tumor for patients that have lung squamous cell carcinoma, which could provide evidence to aid in certain cancer diagnoses and treatments.

Methods:

To start, we found our desired data sets through the Genomic Data Commons (GDC) by filtering the cases and files to specify what our thesis is examining (Grossman et al., 2016). The project specification used was the Cancer Genome Atlas Program for the Lung Squamous Cell Carcinoma (TCGA-LUSC) along with Somatic Single base Mutations (SSM). Based on this project, where the cases involved patients that had lung squamous cell carcinoma, we chose data on the patients that also had either high impact or no mutations in the LRP1B gene in the primary tumor tissue. From this, the data set for high-impact mutations was created with 31 observations and non-mutated with 115 observations. Both these datasets were uploaded to R Studio for analysis (RStudio Team, 2020).

In order to load data from the GDC into R studio the TCGAbiolinks package was used specifically the query function to only obtain open access gene expression data from the TCGA-LUSC project (Colaprico et al., 2015; Mounir et al., 2019; Silva et al., 2016). Once created, the provided base code from the pipeline was run (Bentley, 2024). To only include the sequence data for the LRP1B gene, the sample IDs based on the previously created non-mutated and mutated data sets were used to filter through the TCGA-LUSC. A final data frame was created based on the filtering with a result of 81 observations. This is what will be used for the rest of the analysis.

From there, more filtering was needed in order to remove loci where 90% or more of the read depths had a mean of 0. A principle component analysis (PCA) was performed using ggplot2 to visualize and remove any outliers in the data (Wickham, 2016). This was done by calculating locus variance for all genes and then subsetted the data for the top 50% most variable genes to make the PCA. The samples were colored by what group they were in (high-impact mutations and non-mutated). Once removed, a differential expression analysis was conducted for the two groups using the DESeq2 package (Love et al., 2014). A cutoff significance was then set by logically removing and keeping the values that had a FDR-adjusted p-value less than 0.01 and a log2(fold change) value greater than or equal to 1.5. From here, a volcano plot was generated, where significant values were also labeled. Here, the cutoffs that were used were -1.5 and 1.5 for the FDR-adjusted p-values and 0.05 for the p-value. The top ten genes from the plot were labeled.

Results:

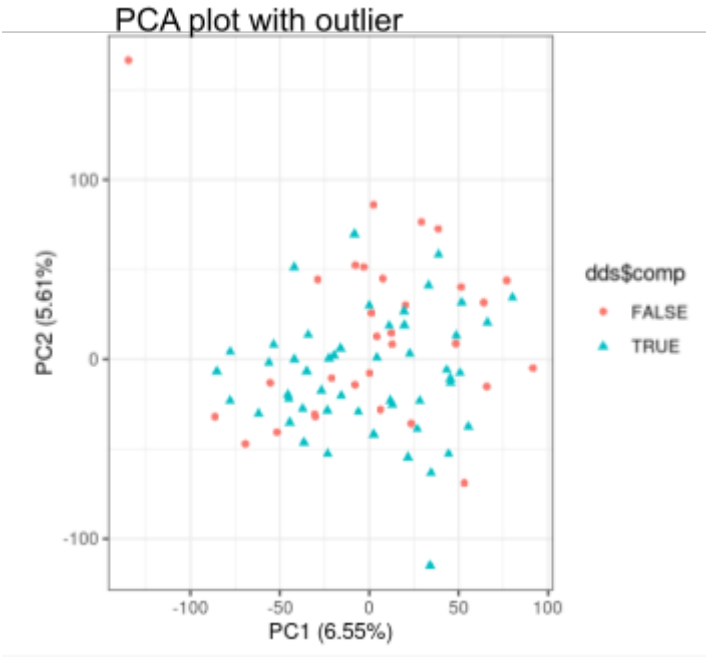

Figure 1: PCA plot with an outlier (81 observations)

When initially plotting our principal component analysis (PCA) plot, our data had 81 observations, where 31 belonged to the group with the high-impact mutations in the LRP1B gene with the other 50 belonging to the non-mutated LRP1B group. When plotting, we noticed a visible outlier that, if not removed, would have potentially messed with our results and analyses. This outlier, which had a sample ID of "TCGA-90-A4ED-01A-31R-A24Z-07" and belonged to our high-impact mutation group, can be seen in the top left corner of the PCA plot in Figure 1. Upon removing this outlier from the data, the observations lowered to 80 due to there now being 29 samples in the high-impact mutation group, and the PCA plot was replotted.

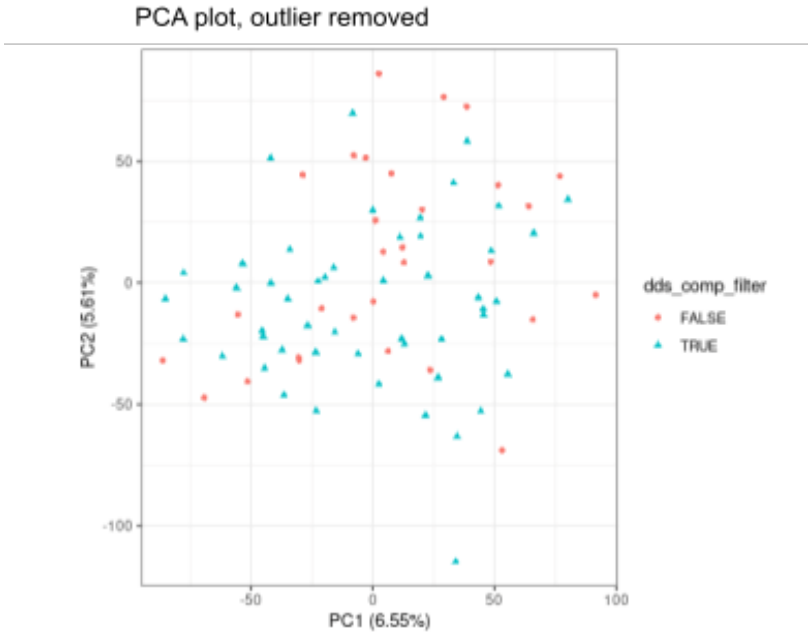

Figure 2: PCA plot with outlier removed (80 observations)

Here, with the outlier removed, it can be seen that the samples within the data share some similarities when compared to the two principal components, with principal component 1 having a variance of 6.55% and principal component 2 having a variance of 5.61%.

To further analyze our data, DESeq2 (Love et al., 2014) was used with default parameters to perform a differential expression analysis, which was visualized in a volcano plot. In this differential expression analysis, 34,634 genes were found to have non-missing significance, which meant that these genes were ones where the adjusted p-values were not missing in the data. Of these 34,634 genes, 114 loci had a false discovery rate (FDR) p-value of less than 0.01, which was about 0.33% of the loci. Additionally, 97 of these loci had a significant and absolute log2 fold change of greater than 1.5. When constructing the volcano plot, the x-axis used the log2 fold change values, while the y-axis was mapped using the -log10 of the FDR-adjusted p-values. Furthermore, 10 genes that were at the top 10 rows after ordering the p-values were labeled because we considered them to be significantly differentially expressed. The cutoff chosen was an adjusted p-value of 0.05.

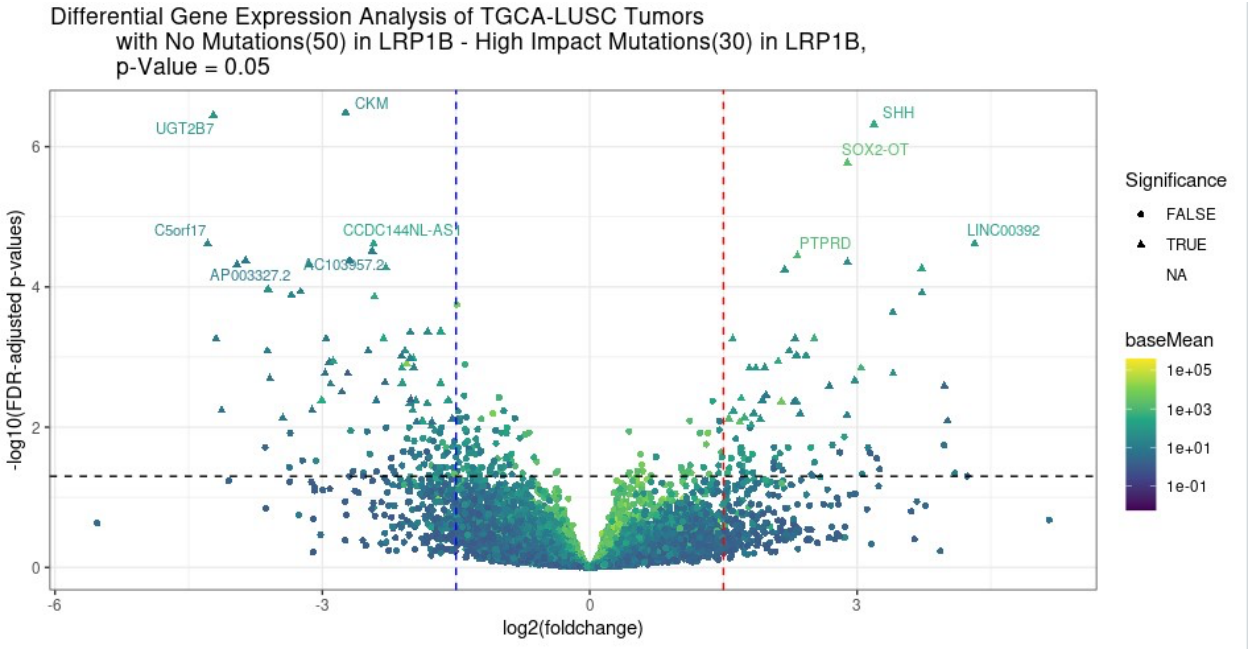

Figure 3: Volcano plot that showcases the top 10 most differentially significantly expressed genes from 34,634 genes, where the cutoff was a p-value of 0.05

The blue and red vertical lines represent the cutoffs for the FDR-adjusted p-value, which were -1.5 and 1.5 respectively. The horizontal line represents the -log10 value of the adjusted p-value, which was mentioned as having a cutoff of 0.05. Out of the 10 labeled significantly differentially expressed genes, the most significant was gene UGT2B7, where the base mean was 73.50248, the log2 fold change was about -4.22, and the FDR adjusted p-value was 3.54e-07**.**

Discussion:

The LRP1B gene only consisted of about 36% of the entire TCGA-LUSC project, which is a considerable amount compared to the others within the same group. It is important to understand how this gene may affect its patient due to its significance. The main functions of this

gene are for signaling pathways and an endocytic receptor, but most importantly this gene acts as a tumor suppressor (Principe et al., 2021). In retrospect this gene has immense importance, so any disturbance of the function of it would most likely not prove to be beneficial. However, the effects of this mutated gene have proved to be the opposite. A study was done to test the effects of mutated LRP1B genes and immune checkpoint inhibitors (ICI) (Brown et al., 2021). It was found through this experiment that 46% of patients with the mutated LRP1B had improved outcomes of ICI therapy after twelve months with them being free from death or progression.

However, this gene LRP1B is known to be most frequently correlated with tumor mutational burden (TMB) along with TP53 (Yu, 2022). From a study that analyzes the difference between co-wild TP53 and LRP1B and mutated TP53 and LRP1B, there was evidence to show that the co-wild had improved survival with immunotherapy (Yu, 2022). These patients had lower TMB and chromosomal instability overall, while also having stronger cytotoxic immune cell infiltrations. These studies prompt the question of whether or not the highly mutated LRP1B gene is as deleterious as initially thought.

The conclusion that was lead to based on the studies and results of the analysis performed, the high-impact LRP1B gene is more than likely a deleterious mutation. It should be stated that this is a broad generalization when the high-impact mutations consist of only 8% of the LRP1B gene within the TCGA-LUSC project. From the first study cited, though some of the mutated genes resulted in beneficial results, it was still only 46% of the population indicating that the non-mutated group had better results over half the time. From Figure 3 the most downregulated gene was UGT2B7 for high-impact mutations. This gene specifically is involved in drug metabolism (U.S. National Library of Medicine). The activity of this enzyme is greatly decreased due to the extreme downregulation which in return lessens the efficacy of the drugs prescribed to certain cancer patients. In addition to this, the genetic variance in a patient's genome can cause fluctuation in the UGT2B7 genotype leading to a difference in the enzymatic activity. For patients who experience a highly mutated LRP1B gene along with the downregulated UGT2B7, it is optimal to investigate different strategies for individualizing therapy treatment. One limitation of this analysis is the extremely small sample group of the high-impact mutation as compared to the non-mutated group. For future implications maybe expanding the thesis to include all mutations vs non-mutations may produce different results allowing for experimentation to be done on the most downregulated gene allowing for a more generalizable approach in dealing with therapy treatment for cancer patients with mutations instead of specifically only high impact.

Supplementary files:

Table Sample Data
(combined high-impact mutated and non-impact mutated groups from the GDC, unfiltered)
Table Sample Data(filtered)
(samDf2, filtered data where additional columns were added, as well as removed cases that did not have primary tumor)
Gene data file

(includes DE results)

Bibliography

Bentley, N. (2024, April 1). *Project Overview Spring 2024*. Bio321G. https://docs.google.com/document/d/13X_3Hxw4M2vjnCNWvDmmcxWsnYLiNRKweXaWfJ705ZM/edit?usp=sharing

Brown, Landon C et al. "*LRP1B* mutations are associated with favorable outcomes to immune checkpoint inhibitors across multiple cancer types." *Journal for immunotherapy of cancer* vol. 9,3 (2021): e001792. doi:10.1136/jitc-2020-001792

Grossman, Robert L et al. "Toward a Shared Vision for Cancer Genomic Data." *The New England journal of medicine* vol. 375,12 (2016): 1109-12. doi:10.1056/NEJMp1607591

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Love, M., Anders, S., & Huber, W. (2014). Differential analysis of count data–the DESeq2 package. *Genome Biol*, *15*(550), 10-1186.

Príncipe, Catarina et al. "LRP1B: A Giant Lost in Cancer Translation." *Pharmaceuticals (Basel, Switzerland)* vol. 14,9 836. 24 Aug. 2021, doi:10.3390/ph14090836

U.S. National Library of Medicine. (n.d.). *UGT2B7 UDP glucuronosyltransferase family 2 member B7 [homo sapiens (human)] - gene - NCBI*. National Center for Biotechnology Information.

Yu, Jiangyong et al. "TP53 and LRP1B Co-Wild Predicts Improved Survival for Patients with LUSC Receiving Anti-PD-L1 Immunotherapy." *Cancers* vol. 14,14 3382. 12 Jul. 2022, doi:10.3390/cancers14143382
