John Saigusa (jms22278)
Roman Martinez (ram6489)
Hee Seung (Lina) Choi (hc28557)

Differentially Expressed Genes Associated with Age in Stomach Adenocarcinoma Patients

ABSTRACT

TCGA-STAD refers to stomach adenocarcinomas which are responsible for 90-95% of stomach cancers and develop from gland cells in the mucosa (innermost lining) of the stomach. STAD was chosen due to the decrease in case rates in the US for reasons not entirely known. However, its prevalence throughout parts of the world drives researchers' efforts to lower case rates globally. The average age of diagnosis for STAD is 68 years old (Cancer.org, 2022).

Differential gene expression analysis was used to compare TCGA-STAD data. The data used in this study filtered patients by short letter code and age at index. The control group for this project focused on primary solid tumor cells of patients that were younger than 68 years while the manipulated group focused on primary solid tumor cells of patients aged 68 and older. The younger group consists of 12 people, while the older group consists of 15 people. This analysis determined whether individuals diagnosed with STAD at the average age of diagnosis or older have genes differentially expressed from those diagnosed younger than the average age. A DESeq analysis was performed and results were visualized in a volcano plot to which differentially expressed genes can be identified and further investigated (Love et al., 2014).

Fourteen genes were differentially expressed in the older patient group. Two genes were upregulated while the remaining twelve were downregulated, with A2ML-1 and ST8SIA6-AS1 being the most significantly downregulated genes. The one-gene plot of ST8SIA6-AS1 demonstrates significant downregulation in the older patient group.

METHODS

We decided to use the Cancer Genome Atlas' RNA-Seq Data on Stomach Adenocarcinoma (TCGA-STAD) and extract it from the NIH Genetic Data Commons Portal (TCGA Research Network). TCGA-STAD data was downloaded and prepared using the TCGAbiolinks package (Colaprico et al., 2016). We obtained GDC data with the use of multiple Bioconductor packages . The GDCquery function in the TCGAbiolinks package was utilized to search for data sets regarding TCGA-STAD in The Cancer Genome Atlas Program database. This function aided in narrowing the data for differential expression analysis. The parameters in the GDCquery function involved searching for the project, data.category, data.type, experimental.strategy, and workflow.type. "STAR - Counts" was the algorithm we used to produce the number of mapped reads per gene (National Cancer Institute).

After TCGA-STAD was downloaded from the database, samples containing NA values were removed in age_at_index when the data was subsetted and saved as a new vector named cnt_woNA. The data was subsetted again based on shortLetterCode to only involve patients that have a primary solid tumor (TP). This subset data was all saved to a new vector named cnt_PT.

cnt_PT was then grouped based on the age_at_index category in which the constant group was assigned by age values less than 68 in a new column called cnt_sub_below68. The manipulated group contained samples of patients 68 years of age and above in a new column named cnt_sub_68nabove. The age value 68 was used as a cutoff to ensure that we would have a sufficient amount of cases (at least 12) for each group used in the analysis (Schurch et al., 2016). The data was further subsetted to exclude any loci that had a high frequency of 0 counts in either of the two groups. The genes where less than 50% of the samples were at 0 counts were kept in each of the two groups. Grouping the subset by mean count further removed any samples that were very considered outliers and too different from the target grouping of data points.

DESeq2 analysis (Love et al., 2014) was then conducted to observe the differential expression analysis of tumorous genes from the age group younger than 68 compared to the age group of patients 68 years and older. The DESeq function was utilized from the DESeq2 package to calculate fold changes and p-values, which allows us to determine how significant the gene expressions were. The parameters were set to default and the results were organized with the gene data into a vector called resOutput.

Afterwards, a volcano plot was created to visualize the statistical significance and magnitude of change for the expressed genes. The two cutoffs set to the maximum false discovery rate adjusted p-value and the minimum log fold difference were 0.01 and 1.50, respectively (Burgman, 2022). The genes were then categorized based on if they were above or below the two cutoffs. The volcano plot uses ggplot2 (Wickham, 2016) to visualize results and ggrepel (Slowikowski, 2021) to plot labels.

Furthermore, a box plot was created using ggplot2 (Wickham, 2016) to visualize the distribution of the single gene, ST8SIA6-AS1, due to significant differential expression levels shown in the volcano plot to further evaluate differences between the normalized counts of the two groups. In addition, a principal component analysis (PCA) was performed to reduce the dimensionality of our dataset for easier visualization and analysis of the differences between the significantly differentially expressed genes. The function prcomp

RESULTS

The DESeq analysis compared 15 samples of patients with primary solid tumors (TP) 68 years of age and older, and 12 samples of TP patients below 68 years of age. This analysis found 14 genes that were highly upregulated/downregulated and were significant statistically compared to other genes.

A volcano plot is interpreted as having the most upregulated genes towards the right of the plot and the most downregulated genes towards the left of the plot. Genes with significant p-values are located near the top of the plot. Thus, highly differentially expressed genes are those in which the FDR-adjusted p-value is lower than the cutoff and the absolute log2FoldChange is greater than the cutoff.

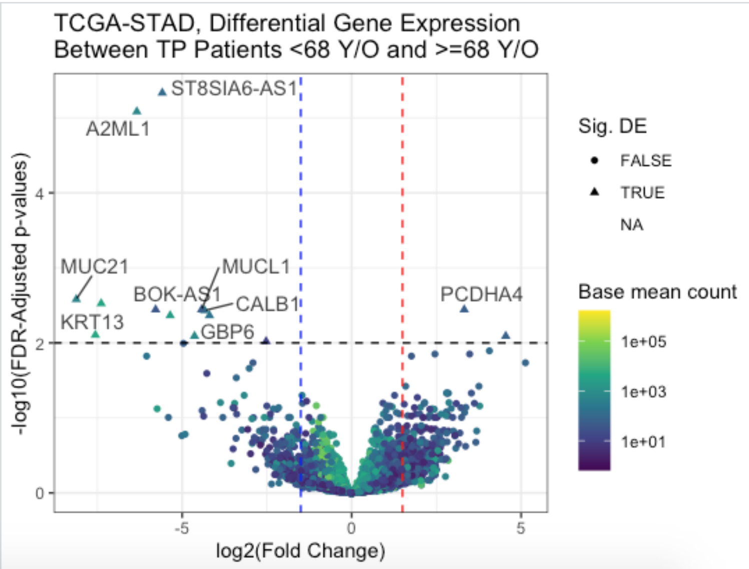

Figure 1. The volcano plot above displays the differential expression analysis of genes between TP patients less than 68 years of age and TP patients 68 years of age and older. The vertical lines represent the positive and negative lfc.cut.off. The black horizontal line represents the fdr.cut.off.

The methods used to produce the volcano plot filtered the TCGA-STAD dataset to include patients with a primary solid tumor, which was further divided into age groups containing patients with <68 years of age and >=68 years of age. Additionally, the DESeq results further categorized the genes based on statistical significance and whether they were strongly upregulated or downregulated. The data was interpreted by whether the genes were above or below the adjusted p-value cutoff and log fold cutoff, indicating their regulation and significance.

Table 1. This table shows a summary of the filtering process prior to DESeq analysis.

	Before Filtering	After Filtering
Number of Genes	60660	32914
Number of Patients Younger than 68	24	12
Number of Patients 68+ Years Old	30	15

Table 2. This table shows the number of genes that had p-adjusted scores deemed significant/or not/or NA, and genes that were differentially expressed/expressed within cutoffs.

Significant_DE	Amount of Genes
FALSE_FALSE	30858
FALSE_TRUE	765
NA_FALSE	1260
NA_TRUE	17
TRUE_FALSE	0
TRUE_TRUE	14

A principal component analysis was performed on each gene after normalizing their RNA expression count data. As can be seen in Figure 2, the gene clusters have somewhat of a wide spread, ranging from around -150 to 150 for PC2 and -150 to 100 for PC1. Because each gene tends to have a fair amount of distance from one another in the PCA plot, there are no obvious outliers present, just a sparsely spread out cluster of genes.

An additional analysis had been performed on the most significantly differentially expressed gene from Figure 1. The gene, ST8SIA6-AS1, had a one-gene plot conducted as shown in Figure 3. As can be seen in the one-gene boxplot, the TRUE group with patients aged younger than 68 had a higher mean value for normalized counts of RNA expression. In contrast, the FALSE group on the chart of patients aged 68 and higher had a lower average amount of RNA expression, with little to no variance in their normalized counts.

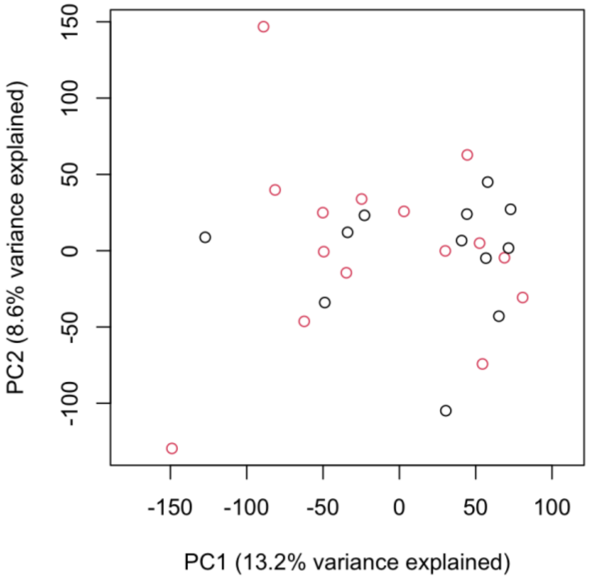

Figure 2. This scatter plot depicts principal component analysis. This plot simplifies differences between different dimensions of data, clustering each RNA-Seq sample into a more predictable array. The black circles are samples from TP patients less than 68 years old, while the red circles are samples from TP patients 68 years and older.

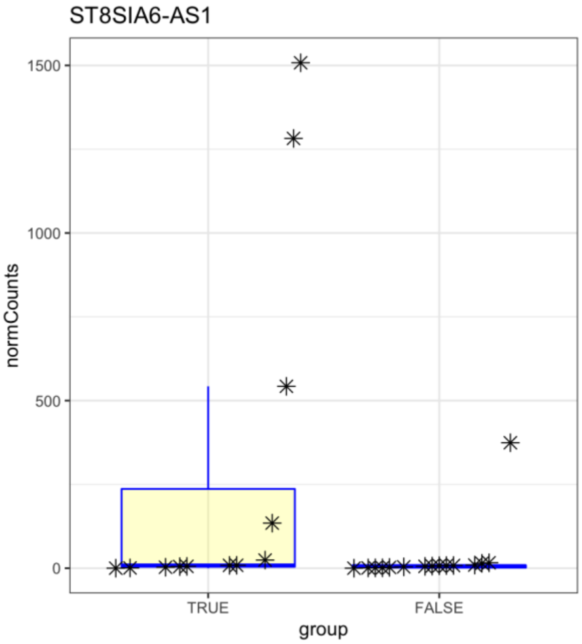

Figure 3. This boxplot depicts a one-gene plot for the gene ST8SIA6-AS1. This gene was determined to be the most differentially expressed gene because it was highly downregulated. The boxplot depicts the quartile ranges of normalized RNA-Seq counts for the ST8SIA6-AS1 gene. Samples from people under 68 years old fall under the TRUE group, while samples from people 68 and older fall under the FALSE group.

ANALYSIS

DESeq analysis was conducted in order to determine which genes were significant, upregulated, and downregulated (Figure 1). The control group was taken from primary solid tumors of adults younger than 68 years old. The treatment group included primary solid tumor data from adults 68 and older. From Table 1 and Figure 1, the treatment group only had 14 genes with significant FDR-adjusted p-values, but all of the significant genes were differentially expressed. Only 2 of the significant and DE groups were overexpressed and the other 12 were downregulated. The cause of this small sample of genes that were significant and differentially expressed may be due to the fact that only patients with primary solid tumors for TCGA-STAD were studied. Because of this, most patients' samples had overactive gene expression in their cells because cancer already involved the overactivation of genes in the cell division pathway. So, it would be less likely for one gene in the tumor group to be differentially expressed from the others whenever they are all under the same overactive, proliferating, tumorous state . One explanation for why there are 12 genes that are significant and downregulated is because these genes may already be downregulated for older patients, so that trend may remain true in cancer patients.

Focusing on one gene may provide more clarity on the differences between patients younger than 68 and patients that are 68 and older. ST8SIA6-AS1 was the most significant differentially expressed gene in Figure 1. ST8SIA6-AS1 is a type of long non-coding RNA (lncRNA) that plays a significant role in the progression of various adenocarcinomas apart from stomach adenocarcinomas (Cao et al., 2020). The function of many lncRNAs like ST8SIA6-AS1 involves the regulation of gene expression at the epigenetic level (He et al., 2021). In Figure 3, a one-gene boxplot was created for ST8SIA6-AS1 in order to plot the normalized RNA-seq data counts from both patients younger than 68 in the TRUE group and patients 68 and older in the FALSE group. This plot shows that the 68 and older population of patients tend to express ST8SIA6-AS1 at lower levels, even with normalized counts. Other studies have shown that the overexpression of ST8SIA6-AS1 is linked to the overactive cell-proliferation observed in lung adenocarcinomas (Cao et al., 2020). Another study has shown that the overexpression of ST8SIA6-AS1 suppresses the regulatory activities of miR-145-5p microRNA (miRNA) and exacerbates the progression of cholangiocarcinoma (He et al., 2021). The main conclusion from other research papers shows that the overexpression of ST8SIA6-AS1 can cause cancer to progress. However, no specific research on STAD has been done in regards to ST8SIA6-AS1 expression levels. It may be possible that the underexpression of ST8SIA6-AS1 could cause stomach adenocarcinoma to progress. Future studies can investigate how the differential expression of ST8SIA6-AS1 affects the progression of STAD, and the cell-proliferation biological pathways that ST8SIA6-AS1 affects in regards to STAD.

After conducting a principal component analysis on the STAD genes, it was determined that there was no obvious correlation between PC1 and PC2. This was because the PCA plot had a wide spread with no apparent pattern. Although there was no simple correlation, at least a known range of values were established in PCA. So, it may not be easy to pinpoint an exact value for a gene in the PCA test because there's no obvious slope in Figure 2, but it would be easy to predict a general range that a value would fall under in the PCA test. So, Figure 2 has provided some semblance of an explanation for multidimensional data.

CONCLUSION

The expression profiles were compared between TP patients below 68 years of age and patients 68 years and older in stomach adenocarcinoma using a DESeq analysis. This was done using the TCGA-STAD project data. Results showed that the significantly differentially expressed genes may or may not be associated with the age of the patient as follow-up research may be done to include a greater number of samples as well as compare TP patients to patients who do not have primary solid tumors (those with normal tissue). Age linked biological differences in STAD may not be well explained by the most significant genes found in this study and other studies regarding ST8SIA6-AS1. Further research may provide insight for improved methods in the way we can treat STAD patients for various age groups.

Works Cited

Becker, R. A., Chambers, J. M. & Wilks, A. R. (1988) The New S Language. Wadsworth & Brooks/Cole.

Burgman, B. (2022) Personal Communication. *University of Texas at Austin*.

Cancer.org. (2022) Key Statistics About Stomach Cancer. *The American Cancer Society*. https://www.cancer.org/cancer/stomach-cancer/about/.

Cao, Q., Yang, W., Ji, X., & Wang, W. (2020). Long Non-coding RNA ST8SIA6-AS1 Promotes Lung Adenocarcinoma Progression Through Sponging miR-125a-3p. *Frontiers in genetics*, *11*, 597795. https://doi.org/10.3389/fgene.2020.597795

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T., Malta, T., Pagnotta, S. M.,Castiglioni, I., Ceccarelli, M., & Noushmehr, G. B. H. (2016). TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data *Nucleic Acids Research 44*(8): e71. (doi:10.1093/nar/gkv1507)

He, J., Yan, H., Wei, S., & Chen, G. (2021). LncRNA ST8SIA6-AS1 Promotes Cholangiocarcinoma Progression by Suppressing the miR-145-5p/MAL2 Axis. *OncoTargets and Therapy, 14*, 3209–3223. PubMed. https://doi.org/10.2147/OTT.S299634

Love, M.I., Huber, W., Anders, S.(2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 *Genome Biology 15*(12):550 A BibTeX entry for LaTeX users is @Article{, title = {Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2}, author = {Michael I. Love and Wolfgang Huber and Simon Anders}, year = {2014}, journal = {Genome Biology}, doi = {10.1186/s13059-014-0550-8}, volume = {15}, issue = {12}, pages = {550}, }

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G. G., Owen-Hughes, T., Blaxter, M., & Barton, G. J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?. *RNA 22*(6), 839–851. https://doi.org/10.1261/rna.053959.115

Slowikowski, K. (2021). ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. R package version 0.9.1. https://CRAN.R-project.org/package=ggrepel

TCGA Research Network. (2022). TCGA-STAD RNA-Sequencing Data. *National Institutes of Health Genetic Data Commons*

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York* https://ggplot2.tidyverse.org