Analysis of Gene Expression in HNSCC Tumor Cells

BS831 Final Project

Sara Shiv Denner

Due 7 May 2025

I. Introduction

Head and neck squamous cell carcinomas affect nearly 600,000 patients per year worldwide and can be caused by smoking. The TCGA head and neck squamous cell carcinoma (HNSCC) dataset contains RNA-seq gene expression profiles from tumor cells collected from either the oral cavity, oropharynx, or laryngeal sites from mainly male patients who are heavy smokers. The aim of this study is to understand the gene expression patterns associated with grade 1 HNSC tumor cells, characterized by cells that look mostly healthy and grow and spread slowly, versus grade 3 HNSC tumor cells, which appear more abnormal and grow and spread more quickly.

The unfiltered dataset contains 34,422 genes across 120 samples, and has already gone through summarization, which converted raw expression counts in the form of .CEL files to expression matrix format. Differential expression analysis with DESeq2 will be used to understand which genes, if any, have significantly differentiated expression levels between phenotype G1 vs. G3. Next, gene set enrichment analysis (GSEA) will be used to investigate if any of the differentially expressed genes are significantly enriched within two human gene sets related to cancer from the Human Molecular Signatures Database (MSigDB). Unsupervised learning in the form of hierarchical clustering is used to understand if grouping by gene expression profiles and similarities alone can successfully differentiate samples belonging to grade G1 against grade G3. Finally, supervised learning in the form of classification analysis will be leveraged to determine the accuracy of a Naïve Bayes classifier to differentiate samples of grade G1 between grade G3.

II. Methods

a. Preprocessing, Quality Control, and Running Differential Expression Analysis

As mentioned in the introduction, the original Expression Set had a total of 34,422 genes and 120 samples. Before any analysis occurred, the Biobase and dplyr packages in R were used to conduct several pre-processing and quality control measures on the raw RNA-seq data. First, the Expression Set was first limited to include just the phenotype of interest by filtering out samples that did not belong to grade G3 or G1. Genes with total counts equal to zero were removed.

The DESeq2 package was used to create a DESeqDataSet (dds) object using the DESeqDataSetFromMatrix() command. Next, a pre-filtering workflow from the Bioconductor

DESeq2 vignette ‘Analyzing RNA-seq data with DESeq2’ was utilized to remove genes with very low counts. This step is not required, but it is useful since it reduces the memory size of the dds data object and therefore increases the speed of differential expression analysis within DESeq2. Only rows that had a count of at least 10 for the minimal number of samples were kept. The minimal number of samples used for this analysis was 40, which was chosen since it is the smallest group size (i.e., there are 40 samples in each of the G1 and G3 grades).

After filtering zero and low count genes, differential expression analysis using DESeq2 was utilized to determine which genes had significantly different expression levels between G3 vs. G1. DESeq2 utilizes a generalized linear model to determine significance, since the distribution of the RNA-seq count data is negative binomial. In the test, G3 was treated as experimental and G1 was treated as the control. The null hypothesis of the test is that there is no differential expression between G1 and G3 and that the log Fold Change is 0. Since we are testing multiple samples, we use a false discovery rate of 0.05 for the entire experiment. The magnitude and direction of the differential expression is quantified by the log2 Fold Change. The top 3 most significantly differentiated genes between G3 and G1 were identified and interpreted, and the DESeq2 built-in MA Plot function was used to plot the mean of the normalized counts against the log Fold Change.

b. Normalization

DESeq2 includes built in normalization functions in its analysis of raw RNA-seq counts, which is why no normalization steps were carried out before differential expression analysis. However, for downstream tasks such as clustering and classification, the dataset needs to be normalized to account for any differences in counts that are not due to gene expression. To do this, we can again utilize DESeq, which uses the median of ratios method of normalization, using the estimateSizeFactors() function. The estimateSizeFactors() function is used to show the counts before normalization and after normalization, plotted with boxplots.

To be able to extract the normalized expression matrix from DESeq, the variance stabilizing transformation (vst()) function is used. VST() computes a variance stabilizing transformation, which is similar to putting the data on the log2 scale, and deals with sampling variability of low counts. The blind = TRUE argument is used to calculate the across-all-samples variability, which is useful for downstream use in clustering, since we will be comparing gene expression across all samples. This type of transformation is necessary so that the variance of a gene does not depend on its mean. The output is a normalized expression matrix and a phenotype annotation matrix to accompany it. We can visualize the variance stabilization using the meanSdPlot() function from the vsn package in R, which plots the rank of the mean count versus the standard deviation for each gene. The red trend line in the resulting plot should be relatively flat with respect to the scale on the y-axis.

c. Gene Set Enrichment Analysis

The gene signature used in gene set enrichment analysis (GSEA) is the results from the differential expression analysis. Two gene sets were studied for enrichment within the gene signature, both from the HALLMARK gene sets from the Human Molecular Signatures Database (MSigDB), since these summarize specific well-defined biological states and display coherent expression in humans. The hypeR package was used to extract the gene sets from the MSigDB database. First, the Epithelial Mesenchymal Transition gene set is used, as it consists of genes relating to epithelial-mesenchymal transition (EMT). We are interested in this gene set because Type 3 EMT is associated with cancer progression and metastasis, and genes from this gene set are involved in initiation and early growth of primary epithelial cancers such as HNSCC.

The next gene set used is the P53 Pathway gene set. This gene set is interesting, because it contains genes known to contribute to p53 pathways and networks. P53 is a protein involved in tumor suppression, and the TP53 gene is the most commonly mutated gene in human cancer with over 100,000 literature citations in PubMed. If this gene set is enriched in the gene signature, it will provide interesting insight into whether tumor suppression genes are being expressed in the G1 or G3 HNSCC tumors.

Gene set enrichment analysis was conducted by first matching each gene set to the DESeq2 results via the HGNC symbol, and then determining the number of genes from each gene set that overlapped with the gene signature. The overlapping genes were ranked in the direction of G3 vs. G1, meaning that genes most upregulated in G3, and therefore with the most positive test statistics, were ranked at the top of the list. The ksGenescore() function introduced in the demo sections was utilized to perform GSEA, generate a Kolmogorov-Smirnov test statistic and p-value, and plot the result to visualize upregulation or downregulation patterns of enrichment.

d. Hierarchical Clustering

Hierarchical clustering was conducted to understand whether G1 and G3 grades could be distinguished from one another using unsupervised learning methods. To do this, the normalized expression matrix and accompanying phenotype annotation matrix outputted from normalization using DESeq2 and VST() functions were used, because normalized data will allow more informative clustering to take place. If non-normalized data was used, the clustering analysis may not yield as meaningful results. The normalized data was further subsetted to the top 1,000 genes with top variability using median absolute deviation (MAD) score. This step ensures that only the top expressed genes are being used to cluster, and reduces the computational load needed to perform hierarchical clustering. Hierarchical clustering on both the rows and columns was done using helper functions introduced in the demo sessions (hcopt.R). Next, the ComplexHeatmap package from Bioconductor was used to generate a heatmap using hierarchical clustering on the rows and the columns.

e. Classification Analysis

Classification analysis was conducted to build a predictive model that could classify tumor grade based on provided labels and gene expression data (this is a supervised approach). A Naïve-Bayes classifier was built using the caret package in R. First, the normalized expression matrix and associated phenotype data were split into training and test sets based on a 50% train/test split. Next, the training set was used to fit a Naïve Bayes model using caret. Feature selection methods were explored via cross validation using the trainControl() function. The best classifier identified using cross validation and the model was fit using caret::train(). The classification accuracy of the training model was evaluated using a confusion matrix and using the caret package. Finally, the Naïve Bayes model was evaluated using the test set using the predict() function and its accuracy was evaluated using caret to create an accuracy measurement and a confusion matrix.

III. Results

After removing all samples that did not belong to grade G3 or G1 and genes with a total count equal to zero, there were 32,208 genes and 80 samples left in the dataset. After removing rows that had less than 10 counts per 40 samples, 16,967 genes remained across the 80 samples in the dataset. After running the DESeq() function on the filtered RNA-seq data, it was found that 4,007 genes were significantly differentially expressed between the G1 and G3 tumor grades. The top 3 most significant genes, their ID's, symbols, raw p-values, FDR adjusted p-values, and log2 Fold Changes are shown in the table below.

Gene ID	Gene Symbol	Raw p-value	FDR Adjusted p-value	Log2FoldChange
ENSG00000197915	HRNR	4.35E-27	7.38E-23	-5.33
ENSG00000134765	DSC1	1.18E-25	1.00E-21	-5.44
ENSG00000176075	LINC00302	1.12E-21	6.33E-18	-5.13

Table 1. Top 3 Most Highly Significantly Differentially Expressed Genes from DESeq2

The following MA Plot was generated from the DESeq2 MA Plot function:

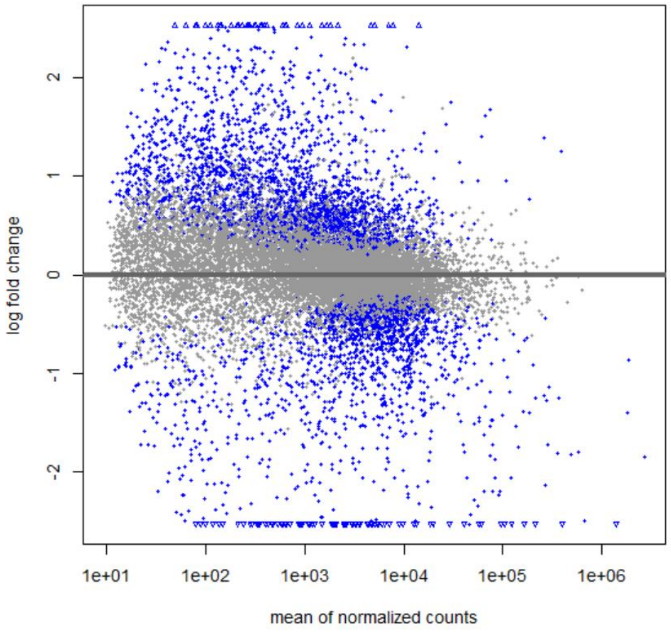

Figure 1. MA Plot of Differentially Expressed genes in G3 vs. G1 phenotype

Internal functions from DESeq2 were used to normalize the RNA-seq data for downstream analyses such as hierarchical clustering and classification. The plots below show boxplots of the raw counts and boxplots of the normalized counts using the estimateSizeFactors() function, which is applied internally when calling both VST() and DESeq() functions.

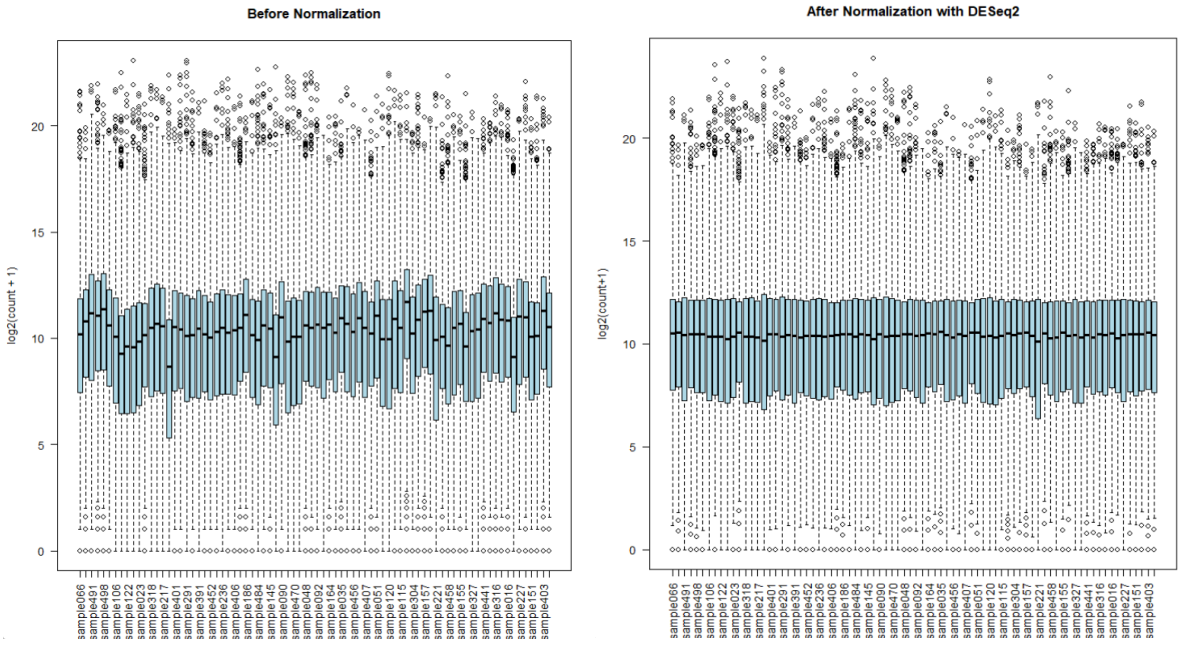

Figure 2. Boxplots of raw RNA-seq counts before and after normalization using internal functions from DESeq().

After applying the VST() function to create the normalized expression matrix and accompanying phenotype matrix, the number of genes was checked to ensure no internal filtering was applied. After applying VST(), there were still 16, 967 genes across 80 samples. When applying normalization and extracting the normalized expression matrix, we can show that the variance across samples is stabilized using the VST() function. Below is the resulting plot from the meanSdPlot() function.

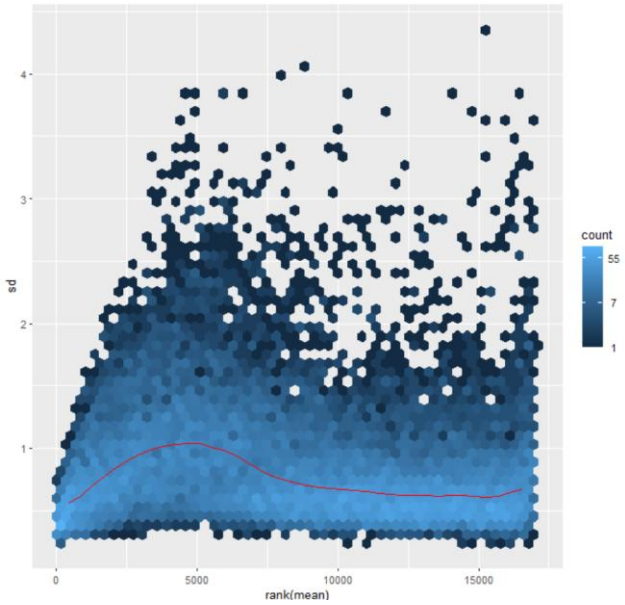

Figure 3. Plot of mean count versus standard deviation for each gene illustrating variance stabilization from the vst() normalization function in DESeq2.

During GSEA, it was found that 197 genes in the gene signature overlapped with genes in the EMT gene set. It was found that 192 genes in the gene signature overlapped with genes in the P53 gene set. These results make sense – since our dataset involves tumor cells, both cancer-related gene sets should have high overlap with our data. The GSEA test using the EMT gene set showed that the EMT gene set is significantly enriched in our data (Kolmogorov-Smirnov p-value = 3.85E-06), and the max enrichment score is 0.18. The enrichment with the EMT gene set is up-regulated in G3. The GSEA test using the P53 gene set showed that the P53 gene set is significantly enriched in our data (Kolmogorov-Smirnov p-value = 1.01E-07), and the max enrichment score is -0.21. The enrichment with the P53 gene set is up-regulated in G1. The following plots were found for both of the enrichment tests.

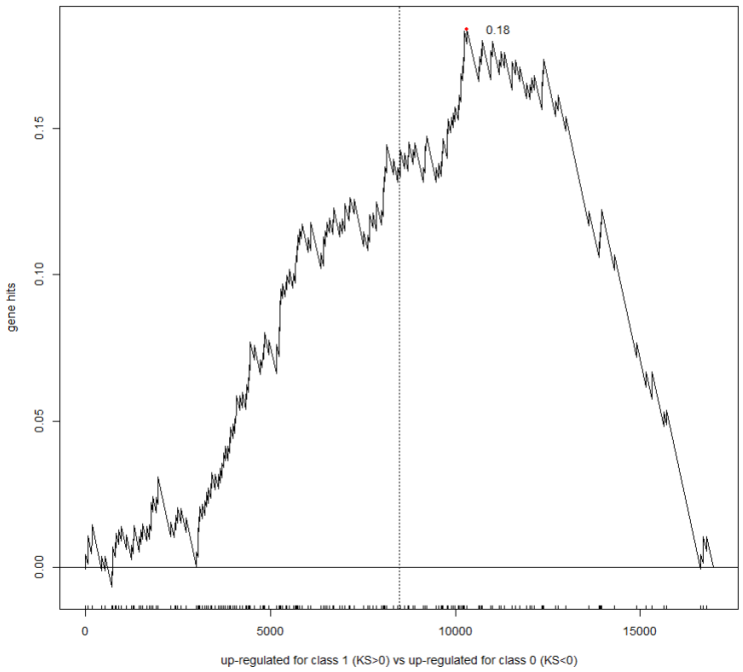

Figure 4. Plot showing enrichment in G3 with the EMT gene set.

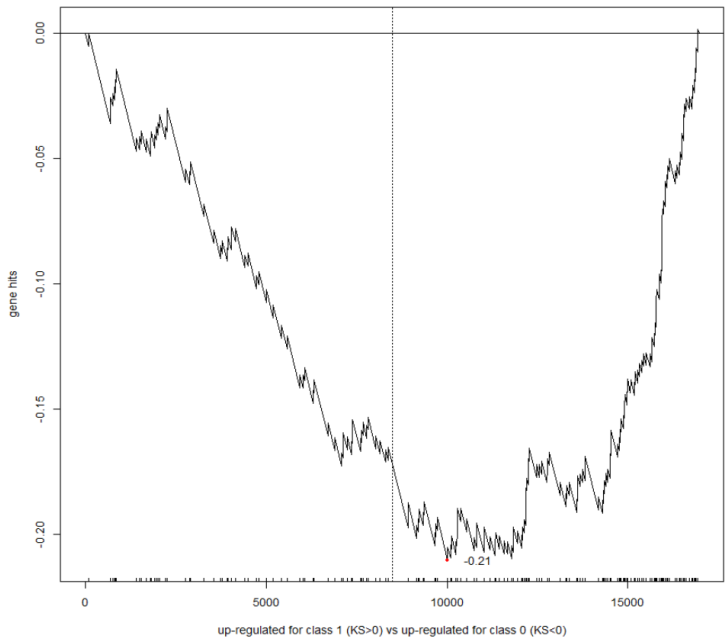

Figure 5. Plot showing enrichment in G1 with the P53 gene set.

Results from the hierarchical clustering step were in the form of a heatmap that compared the G1 and G3 grades. Dendrograms on the columns show clustering across samples and show which condition (either grade “G3” or grade “G1”) the samples belong to. Dendrograms on the

rows show clustering in gene expression, which is shown as a function of the VST() normalized expression counts. The resulting heatmap is found below.

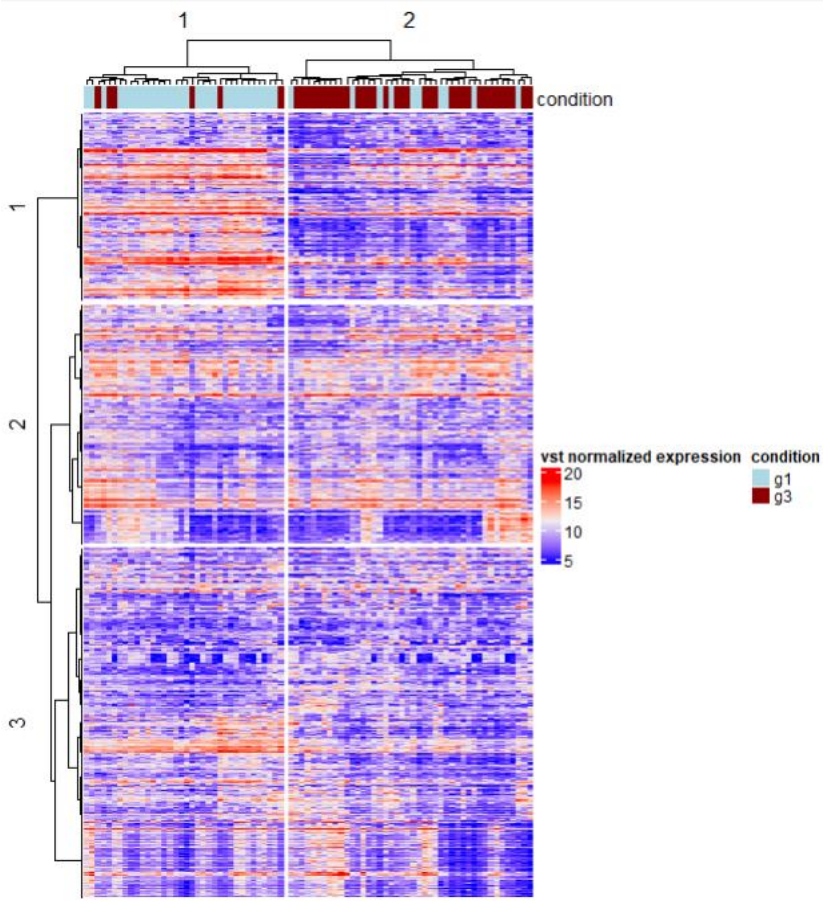

Figure 6. Heatmap illustrating results from hierarchical clustering across the samples and genes.

When fitting the Naïve-Bayes classifier, the accuracy of the training model was found using a confusion matrix and using the caret package. In analyzing the confusion matrix, it was found that the training model correctly predicted 19 observations as G1, and 1 observation was misclassified as G3. Similarly, 19 observations were correctly classified as G3, while 1 observation was misclassified as G1. The caret package output showed that the accuracy of the training classifier was 0.95. When evaluating the accuracy of the classifier using the test set, it was found that 10 observations were correctly predicted as G1 and 10 were misclassified as G3, while 19 were correctly predicted as G3 and 1 was misclassified as G1. The accuracy of the classifier on the test set was 0.725.

IV. Discussion

Interpretations

From the differential expression analysis, it was found that for gene ENSG00000197915 (HRNR), expression is 5.33 times lower in the G3 grade compared to the G1 grade. At an overall

significance level of 0.05, the differential expression is significant, because the adjusted p-value is 7.38E-23 < 0.05. It was found that gene ENSG00000134765 (DSC1) is expressed 5.44 times lower in the G3 grade compared to the G1 grade. At an overall significance level of 0.05, the differential expression is significant, because the adjusted p-value is 1.00E-21 < 0.05. For gene ENSG00000176075 (LINC00302), expression is 5.13 times lower in the G3 grade compared to the G1 grade. At an overall significance level of 0.05, the differential expression is significant, because the adjusted p-value is 6.33E-18 < 0.05. The MA plot shows that there are significantly differentiated genes both upregulated in G3 and downregulated in G3.

HRNR is a gene involved in hepatocellular carcinoma (HCC) progression, and high expression of HRNR in HCCs has been found to be associated with vascular invasion and poor tumor differentiation. DSC1 is a gene that has shown to be expressed higher in HNSCC and patients with higher DSC1 had unfavorable prognosis. As shown in Wang et al, DSC1 was significantly higher expressed in HNSCC cells than in normal tissue. LINC00302 is a long intergenic non-coding gene not involved with protein coding, though associated with epithelial cancers.

Curiously, all three of these genes, which are known to be significantly expressed in tumor cells, are significantly downregulated in grade 3 versus grade 1 HNSCC tumors in the dataset. These findings could illustrate that while genes like HRNR, DSC1, and LINC00302 are involved in early tumor development, reflected in earlier stage G1 grade tumors, their expression may decrease in more aggressive tumors such as G3 tumors. The loss of expression could be due to mechanisms such as epithelial-mesenchymal transition (EMT). While these genes are highly expressed in lower grade tumor cells that are still developing – i.e., more epithelial and tumor-promoting – they may be shut off in later stage tumors who have lost their epithelial identity and have transitioned to a mesenchymal – or more invasive – phenotype.

The gene signature used in the enrichment analysis was the set of differentially expressed genes that were generated using DESeq2. The first enrichment test tested whether the Hallmark Epithelial Mesenchymal Transition (EMT) gene set was enriched in the gene signature. It was found that the EMT gene set was significantly enriched in the gene signature (Kolmogorov Smirnov p = 3.85E-06) and that the enrichment is upregulated in the G3 tumors, meaning that more genes related to EMT are upregulated in G3 than not. Earlier in the analysis, it was found that several genes that were significantly differentially expressed, namely DSC1, HRNR, and LINC00302, were downregulated, suggesting that EMT be more likely to be expressed in G1 rather than G3. These results may not be contradictory. The GSEA results show us that most of the genes in the gene signature that overlap with EMT genes tend to be more highly expressed in G3 than G1. This could suggest that the EMT genes that overlap with the gene signature may cause more mesenchymal and invasive (more aggressive) tumors, while others in the gene set could be more epithelial, such as DSC1, HRNR, and LINC00302.

The second enrichment test tested whether the P53 gene set was enriched in the gene signature. It was found that the P53 gene set was significantly enriched in the gene signature (Kolmogorov Smirnov p = 1.01E-07) and that the enrichment is upregulated in the G1 tumors, meaning that more genes related to p53 pathways are upregulated in G1(low grade tumors) rather than in G3 (high grade tumors). This result could be reflective of the fact that the p53 gene set involves genes involved in tumor suppression. When the genes in the p53 pathway are functioning normally, tumor progression may be limited due to the function of the genes in the pathway. In high grade (G3) tumors, genes in the p53 pathway may be mutated or suppressed, allowing more aggressive tumors to proliferate. In low grade (G1) tumors, genes in the p53 pathway may still be active, which would cause more expression and explain the result found.

Results of the hierarchical clustering step show that most samples belonging to grade "G3" were placed in cluster 2 in the columns, while most samples belonging to grade "G1" were placed in cluster 1 in the columns. This result is expected – unsupervised methods were able to distinguish sample groups based on expression values, but not perfectly, as we can see there are some misclassified samples of grade "G3" in cluster 1 and of grade "G1" in cluster 2. By looking at the clustering in the genes, we can see that cluster 1 consists of mostly high expressed genes, while clusters 2 and 3 consist of mainly low expressed genes. We can see that highly expressed genes in cluster 1 in the genes tended to exist mostly in cluster 1 in the columns, showing us that the highest expressed genes in the dataset mostly belong to the G1 grade, while lower expressed genes tended to be in grade G3.

The Naïve Bayes classifier had an accuracy of 0.95 on the training set, while an accuracy of 0.725 when used on the test set. The reduction in accuracy is expected, since the test set includes samples that were initially not used to train the classifier. This shows that the model's performance drops substantially on the test set, and it is concerning that 10 G3's were misclassified as G1. This means that the classifier could underestimate tumor severity – if used in a medical setting, using this classifier could be risky, as failing to detect aggressive tumors could have consequences.

Future directions of this analysis could expand on the classification portion to further validate the accuracy of the classifier generated. For example, further feature selection could be employed to further improve the accuracy of the classifier, and ROC curves could be generated to quantify the accuracy. A next step from the hierarchical clustering step could include identifying the high expression genes belonging to grade G1 that clustered in cluster 1 in the columns and cluster 1 in the rows and determine if any of them are of specific biological interest.

Strengths of this analysis come mainly from the use of literature and online vignettes to verify steps taken. The dataset also had many samples and many genes, allowing for high power to detect significant results if significant results existed. The analysis did have limitations. For example, normalization of the RNA-seq counts was conducted using internal functions from DESeq2, namely the VST() function. The plot of the rank of the mean versus the standard

deviation for each gene shows that there appears to be slightly higher variance between ranks 2000-5000 in our dataset. Therefore, a potential limitation of the analyses conducted using the normalized expression matrix outputted from these functions could be that VST() did not properly stabilize the variance. This may cause the variance of certain genes to depend on their mean expression. A future analysis could explore other normalization techniques to determine if a different technique may yield more reliable results. Another limitation of this analysis is the inaccuracy of the Naïve Bayes classifier. Further steps should be taken to increase the accuracy of the classifier.

Certain challenges in the analysis included determining a method of filtering out low-expressed genes before performing differential expression analysis. There were certain filtration attempts made that failed to filter out low expressed genes, causing biased MA-plots and strange outliers. I also attempted to filter to the top 10,000 MAD genes before doing differential expression analysis, which yielded strange interpretations of results as well. I believe this happened because filtering out the top MAD genes filters out genes with low variability, however does not take into consideration the gene expression, so it is possible that doing this could have filtered out highly expressed genes between G1 and G3 that happened to have low variability. I read further into the Bioconductor DESeq2 vignettes and followed an example RNA-seq analysis and QC pipeline to decide on my final QC and pre-filtering technique.

Appendix A.

Works Cited

"Cancer Grade vs. Cancer Stage." *MD Anderson Cancer Center*, www.mdanderson.org/patients-family/diagnosis-treatment/a-new-diagnosis/cancer-grade-vs--cancer-stage.html. Accessed 3 May 2025.

The Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576–582 (2015). https://doi.org/10.1038/nature14129

Kalluri R, Weinberg RA. The basics of epithelial-mesenchymal transition. J Clin Invest. 2009 Jun;119(6):1420-8. doi: 10.1172/JCI39104. Erratum in: J Clin Invest. 2010 May 3;120(5):1786. PMID: 19487818; PMCID: PMC2689101.

Love, Michael I. "Analyzing RNA-Seq Data with DESeq2." *Analyzing RNA-Seq Data with Deseq2*, Bioconductor, 15 Apr. 2025, bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#principal-component-plot-of-the-samples.

Wang Y, Chen C, Wang X, Jin F, Liu Y, Liu H, Li T, Fu J. Lower DSC1 expression is related to the poor differentiation and prognosis of head and neck squamous cell carcinoma (HNSCC). J Cancer Res Clin Oncol. 2016 Dec;142(12):2461-2468. doi: 10.1007/s00432-016-2233-1. Epub 2016 Sep 6. PMID: 27601166.

Fu SJ, Shen SL, Li SQ, Hua YP, Hu WJ, Guo B, Peng BG. Hornerin promotes tumor progression and is associated with poor prognosis in hepatocellular carcinoma. BMC Cancer. 2018 Aug 13;18(1):815. doi: 10.1186/s12885-018-4719-5. PMID: 30103712; PMCID: PMC6090597.

Rombaut, D., Chiu, HS., Decaesteker, B. *et al.* Integrative analysis identifies lincRNAs up- and downstream of neuroblastoma driver genes. *Sci Rep* **9**, 5685 (2019). https://doi.org/10.1038/s41598-019-42107-y

"VSN." *Bioconductor*, bioconductor.org/packages/release/bioc/html/vsn.html. Accessed 3 May 2025.

Meeta Mistry, Radhika Khetani. "Count Normalization with Deseq2." *Introduction to DGE - ARCHIVED*, Harvard Chan Bioinformatics Core, 26 Apr. 2017, hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html.

Ziebell, Frederik. "RNA-Seq Quality Control." *RNA-Seq Quality Control*, R CRAN, cran.r-project.org/web/packages/RNAseqQC/vignettes/introduction.html. Accessed 3 May 2025.

Hernández Borrero LJ, El-Deiry WS. Tumor suppressor p53: Biology, signaling pathways, and therapeutic targeting. Biochim Biophys Acta Rev Cancer. 2021 Aug;1876(1):188556. doi: 10.1016/j.bbcan.2021.188556. Epub 2021 Apr 29. PMID: 33932560; PMCID: PMC8730328.

Appendix B.

Comprehensive genomic characterization of head and neck squamous cell carcinomas | Nature

Cancer Grade vs. Cancer Stage | MD Anderson Cancer Center

The basics of epithelial-mesenchymal transition - PMC

Analyzing RNA-seq data with DESeq2

Lower DSC1 expression is related to the poor differentiation and prognosis of head and neck squamous cell carcinoma (HNSCC) - PubMed

Hornerin promotes tumor progression and is associated with poor prognosis in hepatocellular carcinoma - PubMed

Integrative analysis identifies lincRNAs up- and downstream of neuroblastoma driver genes | Scientific Reports

Bioconductor - vsn

Count normalization with DESeq2 | Introduction to DGE - ARCHIVED

RNA-seq Quality Control

Tumor Suppressor p53: Biology, Signaling Pathways, and Therapeutic Targeting - PMC