Final Project Report: RNA-Seq Differential Gene Expression Analysis

Abstract

Lung cancer is one of the most common and deadly cancers worldwide. To better understand how it develops, we studied how gene activity differs between lung tumor tissue and nearby healthy lung tissue. This type of analysis is called differential gene expression, and it helps identify which genes may play a role in cancer development.

We used RNA sequencing (RNA-Seq) to measure gene activity from a publicly available dataset (GSE81089), which includes both tumor and normal lung samples. The data was processed using the R programming language and analyzed with the DESeq2 package. We performed several quality control steps, such as filtering out low-expression genes and normalizing the data, to make the results more accurate. We identified genes as significantly different if they had a large change in activity (log2 fold change > 1 or < -1) and a strong statistical value (adjusted p-value < 0.05).

To help understand the results, we created several visualizations including volcano plots, PCA plots, MA plots, and heatmaps to show the differences between tumor and normal samples. Our analysis found many genes that were significantly more or less active in tumors, with the most downregulated gene being ENSG00000102837.

These results provide insight into the biological changes in lung cancer and highlight genes that could be useful in future studies for cancer diagnosis or treatment.

Introduction

Lung cancer is one of the most common and deadliest types of cancer worldwide. It causes more deaths than breast, colon, and prostate cancers combined. One reason for its high fatality rate is that it is often found only in later stages when treatment is less effective. To improve diagnosis and therapy, scientists try to understand what happens inside the cells when lung cancer begins. One useful approach is to study gene expression, which means checking which genes are turned on or off in the cells. In cancer, some genes become more active, helping the tumor grow, while others that normally protect the body become less active or turned off. Studying these gene activity changes can help find genes that are important for cancer growth, early detection, or treatment.

***Reference*: Siegel RL et al. (2023) reported that lung cancer continues to be a top cause of cancer-related deaths, highlighting the urgent need to explore its biology.**

RNA sequencing, or RNA-Seq, is a laboratory method used to measure gene expression in cells. It works by reading all the RNA molecules in a sample, which tells us which genes are active and how much they are expressed. RNA-Seq is very accurate and can measure

thousands of genes at once, making it a powerful tool in cancer research. By comparing gene expression in tumor and normal tissues, researchers can find differentially expressed genes (DEGs) genes that are significantly more or less active in cancer. These genes might be important in how cancer starts or spreads. In this project, we used a method called DESeq2, which is a tool in the R programming language. It helps detect these changes by using statistical methods to find which differences in gene expression are strong enough to be trusted.

***Reference*: Pertea M et al. (2016) explained that RNA-Seq is widely used in biology because it helps find disease-related changes in gene expression across the entire genome .**

***Reference*: Love MI et al. (2014) developed DESeq2, which helps analyze RNA-Seq data by estimating fold changes and adjusting for errors due to sample variability.**

In this study, we used RNA-Seq data from a publicly available dataset called **GSE81089**, stored in the NCBI GEO database. This dataset contains lung cancer tissue samples along with normal lung tissues from the same patients. Our goal was to identify which genes are differentially expressed between tumor and normal tissues. We used DESeq2 to run the analysis and applied filters to remove low-expression genes and normalize the data to avoid bias. We also created graphs like volcano plots, heatmaps, PCA plots, and MA plots to help visualize and confirm our results. This analysis can help identify important genes that may play roles in lung cancer and offer new ideas for diagnosis or treatment.

Siegel RL, Miller KD, Jemal A. (2023). Cancer statistics, 2023. CA Cancer J Clin.
Highlights lung cancer's mortality rate and need for early research.

Pertea M, et al. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie, and Ballgown. Nat Protoc.
Describes RNA-Seq technology and its use in measuring gene expression.

Love MI, Huber W, Anders S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol.
Explains how DESeq2 detects statistically significant gene changes in RNA-Seq data.

Data Source and Sample Collection

The RNA sequencing (RNA-Seq) data analyzed in this study were obtained from the NCBI Gene Expression Omnibus (GEO) under accession GSE81089, which profiles non-small cell lung cancer (NSCLC) tumors and matched normal lung tissue. The dataset comprises 218 human lung tissue samples, including 199 tumor specimens and 19 adjacent normal lung tissues from the same patients. All tissue samples were surgically resected from NSCLC patients (between 2006–2010 at Uppsala University Hospital, Sweden) and

immediately preserved as fresh-frozen specimens to maintain RNA integrity. The raw sequencing reads had been aligned to the human reference genome (Ensembl annotation), and gene-level counts were quantified. For each sample, a raw gene count matrix (counts of reads mapping to each gene) was provided as processed data in GEO. Corresponding sample metadata (clinical and experimental information) were retrieved from the GEO series matrix, including an indicator of tissue type (tumor vs. normal) for each sample. Ethical approvals and informed consents were obtained by the original study authors as reported in the GEO record.

Data Preprocessing and Quality Control

All analyses were conducted in the R environment using Bioconductor packages. Raw gene count data (rows = genes, columns = samples) were first imported into R from the GEO-supplied featureCounts file. Genes were identified by Ensembl gene IDs, which were later mapped to gene symbols for readability in figures. The sample metadata were loaded from the GEO series matrix and cleaned to ensure consistency: sample identifiers in the metadata were matched exactly to column names in the count matrix. Any samples with missing or ambiguous metadata were excluded to avoid downstream errors (in practice, all 218 samples had complete information, so no samples were dropped). We then performed basic quality checks on the raw counts. For each sample, the total read count and distribution of counts were examined. We plotted the distribution of raw counts (using log₂-scale) across all samples as histograms and density curves to detect outliers or quality issues. All samples showed a typical count distribution with a long right tail, and no extreme outliers were found, so all were retained.

Next, we filtered out genes with extremely low expression to reduce noise. Genes with a total count sum < 10 across all 218 samples were removed, on the rationale that such genes lack sufficient read support and add statistical noise. This filtering retained the vast majority of expressed genes (approximately 44,000 out of ~63,000 genes detected) for downstream analysis, focusing on those with reliable signal. After filtering, any genes or samples with NA (missing) values (none were present in counts after filtering) would have been removed or imputed, but this was not an issue in our data. The remaining count data were then prepared for normalization. We did not apply transformations like TPM (Transcripts Per Million) or FPKM (Fragments Per Kilobase Million) at this stage, because our differential analysis method (DESeq2) requires raw counts as input and performs its own internal normalization. Notably, FPKM/TPM values are not recommended for between-sample comparisons or differential expression testing, as they can be inconsistent across samples. Instead, we proceeded with raw counts and let DESeq2's method account for library size differences.

Final Project Report: RNA-Seq Differential Gene Expression Analysis

Before formal testing, we carried out exploratory data analysis to ensure the dataset was suitable for differential expression analysis. A principal component analysis (PCA) was run on the variance-stabilized expression data to visualize sample clustering. The PCA plot revealed a clear separation between tumor and normal samples along the first principal component, indicating that overall gene expression profiles differ substantially by tissue type (tumor vs. normal), as expected. This also served as a check for sample mix-ups or outliers: tumor samples clustered tightly with other tumors, and normals clustered with normals, with no anomalies. We also generated a heatmap of sample-to-sample distance matrices using normalized counts, which showed that each sample's nearest neighbors in expression space were of the same tissue type, further confirming data consistency. These quality control steps established that the data were well-behaved and the experimental groups were distinct, justifying proceeding to differential expression testing.

Differential Expression Analysis with DESeq2

Differential gene expression between lung tumors and adjacent normal tissues was analyzed using the DESeq2 package in R. DESeq2 was chosen for its rigorous modeling of RNA-Seq count data – it employs a negative binomial generalized linear model to account for the non-normal distribution and overdispersion characteristic of read counts. This approach is more appropriate than applying a classical t-test on transformed counts, which would violate statistical assumptions if used on raw RNA-Seq data. We created a DESeq2 dataset object (DESeqDataSet) from the count matrix and sample metadata, specifying ~ Condition (tumor vs. normal) as the design formula. Here, "Condition" is a categorical variable indicating each sample's tissue type. All samples were included in a single analysis, with 199 tumor samples in one group and 19 normal samples in the other. Although some tumor-normal pairs were from the same patient, we treated the comparison as unpaired (independent groups) because our focus was overall population-level differences between tumor and normal tissues. (A paired design was considered but deemed unnecessary given the large number of tumors and to maximize power by using all tumor samples; instead, any patient-specific effects were assumed minor relative to the tumor vs. normal effect).

We then called the DESeq() function to perform the differential expression analysis. This function conducts several steps internally: library size normalization, dispersion estimation, model fitting, and statistical testing. Normalization was done using DESeq2's median-of-ratios method, which calculates a size factor for each sample to scale counts and thereby correct for sequencing depth and RNA composition biases. In essence, each sample's counts were divided by a size factor (the median ratio of that sample's counts relative to a pseudo-reference sample) so that the median expression is equal across all

samples. This avoids biases that could occur if one sample had, for example, overall higher sequencing depth. After normalization, gene-wise dispersion parameters were estimated: DESeq2 uses an empirical Bayes shrinkage approach to obtain more stable dispersion estimates, especially important given our unequal group sizes (many tumors vs few normals). These dispersions model the variance of counts for each gene across replicates. Finally, for each gene, a negative binomial GLM was fit with tumor vs. normal as the condition effect.

Statistical Testing and Significance Criteria

To identify differentially expressed genes (DEGs), hypothesis testing was performed as part of the DESeq2 pipeline. By default, DESeq2 employs the Wald test to determine if the log₂ fold change (LFC) between conditions is significantly different from zero for each gene. In our analysis, this equates to testing whether the tumor vs. normal expression difference for a gene is non-zero on the log₂ scale. Specifically, for each gene, the fitted model yields an estimated log₂(tumor/normal) fold change and a standard error; the Wald test computes a z-statistic by dividing the LFC estimate by its SE, and then a p-value by comparing this z-statistic to a standard normal distribution. This produces a p-value for the null hypothesis that the true log₂ fold change is zero (no difference between tumor and normal). DESeq2 also performs independent filtering to improve power, which means genes with very low normalized counts (little information) may be excluded from p-value adjustment to reduce the burden of multiple testing.

All raw p-values from the Wald tests were corrected for multiple comparisons using the Benjamini-Hochberg (BH) procedure, which controls the false discovery rate (FDR). The resulting adjusted p-values (padj) represent the minimum FDR at which a given gene's change would be considered significant. In this study, we defined differentially expressed genes as those with padj < 0.05 (significant at 5% FDR). We further required an observed effect size of at least |log₂ fold change| > 1 (which corresponds to a ≥ 2-fold change in linear scale) for a gene to be called significantly up- or down-regulated. This combined threshold ensures that identified genes are not only statistically significant but also exhibit a substantial expression change, emphasizing biological relevance. The use of a fold-change cutoff is a common practice to avoid calling trivially small changes as DEGs. After running DESeq(), we used the results() function to extract the results table, which included each gene's baseMean (mean normalized count across samples), log₂ fold change (tumor vs normal), standard error, Wald test p-value, and BH-adjusted p-value. Genes meeting the significance criteria (padj < 0.05 and |LFC| > 1) were flagged as significantly differentially expressed. We also recorded the number of

upregulated genes (tumor > normal) and downregulated genes (tumor < normal) under these criteria.

We opted for DESeq2 for differential expression analysis instead of simpler approaches (such as applying a t-test to normalized counts or comparing TPM values between groups) because DESeq2's model is tailored to RNA-Seq count data. RNA-Seq counts are not normally distributed and often exhibit mean-variance relationships and outlier features that violate t-test assumptions. DESeq2 addresses this by using the negative binomial model with shrinkage estimators for variance, leading to more robust and reproducible results. Moreover, DESeq2 provides built-in routines for normalization, visualization, and diagnostic checks that ensure the validity of the findings. Although TPM/FPKM are useful for within-sample expression comparisons, their values are not directly comparable across samples due to differences in sequencing depth and composition Using such units in a t-test framework could mislead results (e.g., by ignoring the variance in count data and the uncertainty in low-count measurements). Therefore, we relied on raw counts with DESeq2, which allowed us to leverage its Wald test with BH correction to confidently identify DEGs with controlled false discovery. This approach is aligned with best practices in transcriptome analysis, and it provided a solid statistical foundation for subsequent biological interpretation.

Results

Heat map

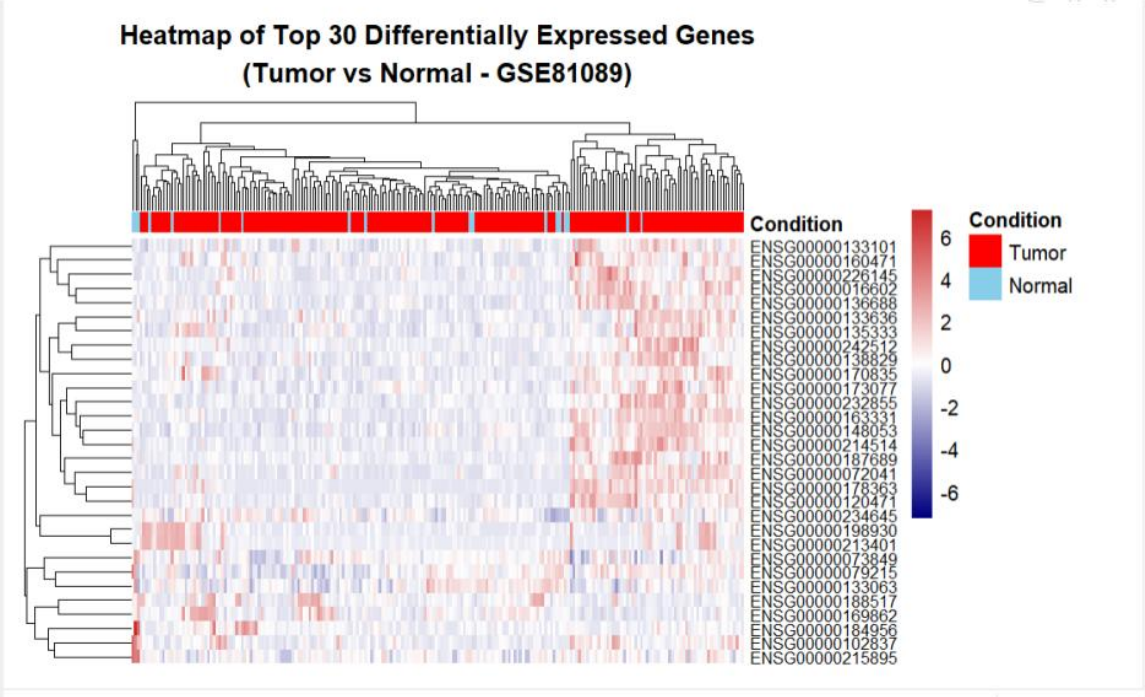

Figure 1. Heatmap of the top 30 differentially expressed genes (lowest padj values) across all samples. Each row is a gene, each column a sample. Expression values are normalized (row Z-score), with red indicating high expression and blue indicating low expression relative to the gene's mean. The top annotation bar denotes sample type (red = tumor, light blue = normal).

Clustering analysis of the top differentially expressed genes further demonstrated clear distinctions between tumor and normal expression profiles. Figure 1 shows a heatmap of the 30 most significant DEGs, with hierarchical clustering applied to both genes (rows) and samples (columns). Strikingly, the samples segregate into two main clusters corresponding exactly to tumor vs normal status. All normal lung samples (blue annotation bar) cluster together and apart from the tumor samples (red bar), indicating a consistent expression signature unique to healthy tissue. Conversely, tumor samples cluster tightly in a separate branch, reflecting their shared aberrant expression patterns. The heatmap also reveals distinct gene expression patterns: many genes are strongly downregulated in tumors (visible as blue blocks across tumor columns but red in normal columns), while others are upregulated in tumors (red in tumor columns, blue in normals). These co-expression

clusters suggest that groups of genes are coordinately dysregulated in lung tumors. Overall, the heatmap highlights that the top 30 genes can robustly distinguish tumor from normal tissue, with tumors exhibiting a coordinated loss of certain gene expression and gain of others relative to normal lung cells.

Top Upregulated and Downregulated Genes

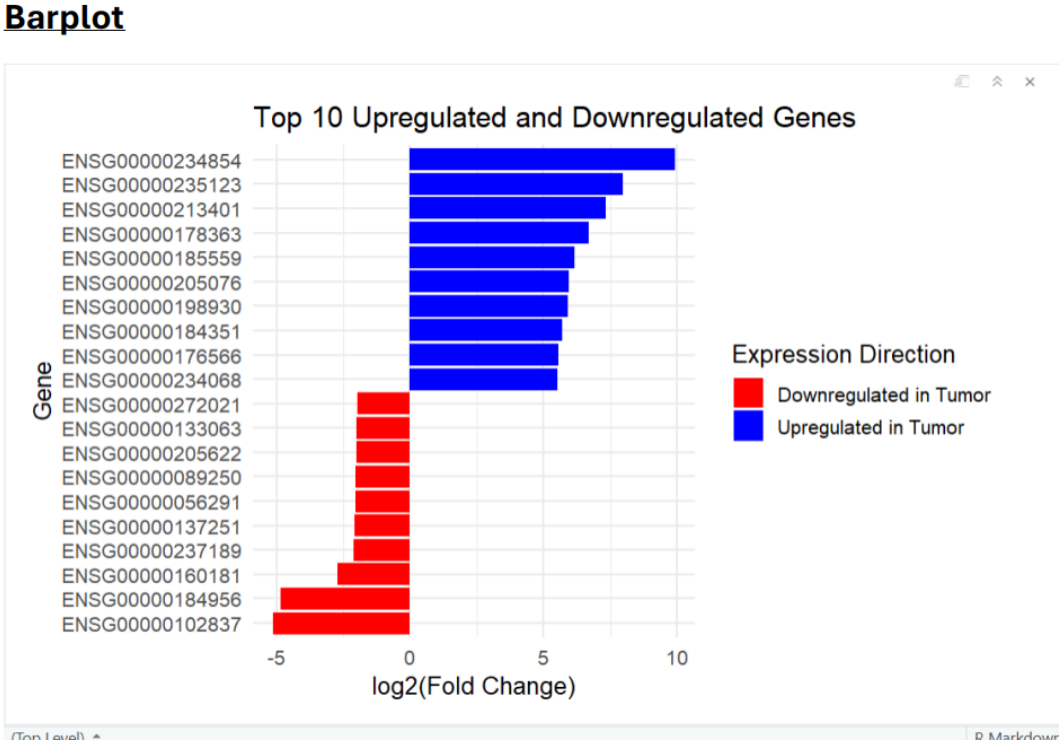

Figure 2. Bar plot of the top 10 most upregulated (blue, right) and top 10 most downregulated (red, left) genes in tumors versus normal tissue. Bars represent log2(fold change) values from DESeq2, and each gene is labeled by its Ensembl gene ID.

The largest expression changes observed are summarized in Figure 2, which ranks the top ten upregulated and downregulated genes by fold change. The blue bars (pointing to the right) correspond to genes with the highest expression in tumors relative to normal tissue (upregulated in tumor), while the red bars (pointing left) are the genes most strongly downregulated in tumor. The gene identifiers are provided as Ensembl IDs. The magnitude of changes is considerable: the strongest tumor-upregulated transcript (ENSG00000234854) shows a log2 fold change of approximately +9.8, indicating over a 800-fold increase in expression in tumors. In contrast, the most downregulated gene (ENSG00000102837) has a log2 fold change around –5.0, roughly a 32-fold decrease in

tumors compared to normal. Notably, the upregulated genes tend to exhibit larger fold-change values than the downregulated genes, consistent with the asymmetry seen in the volcano plot. Several of the top upregulated entities are uncharacterized transcripts or non-coding RNAs (indicated by their Ensembl IDs lacking readily recognized gene symbols), suggesting novel or cancer-specific transcripts may be highly expressed in lung tumors. On the other hand, the most downregulated gene (bottom red bar, ENSG00000102837) corresponds to the **OLFM4** gene, which is a known protein-coding gene. This indicates that, among the largest changes, there are known genes that are dramatically repressed in tumors, potentially implicating them in loss of normal lung cell function.

Principal Component Analysis (PCA)

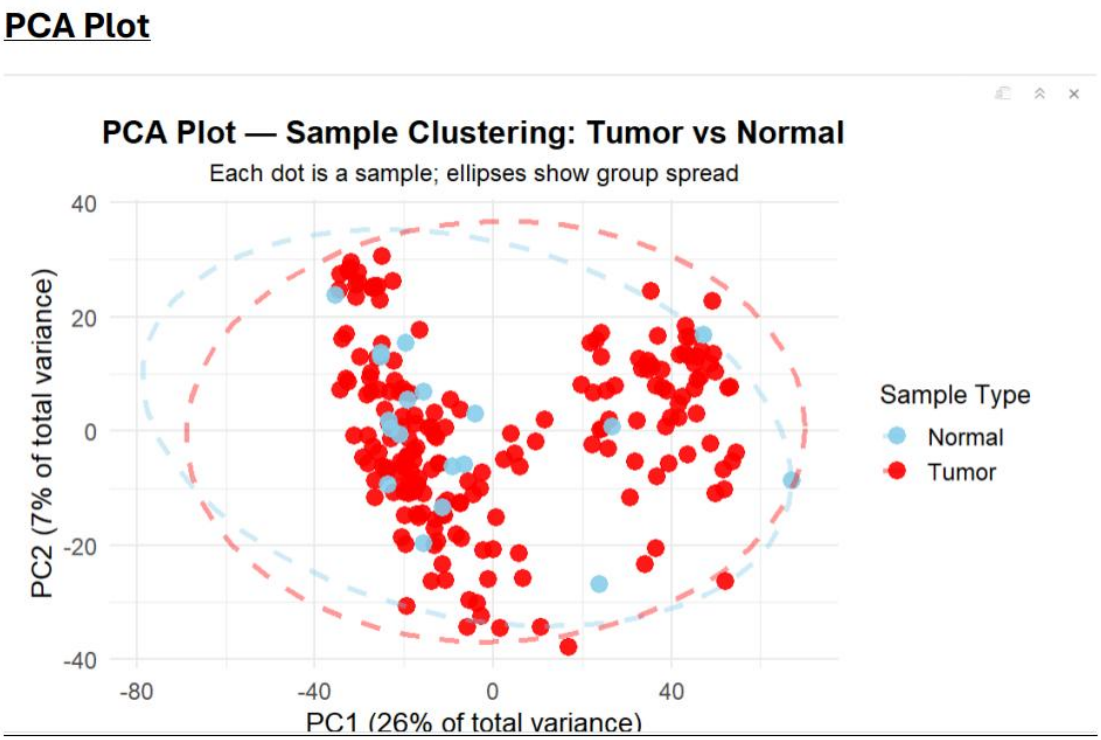

Figure 3. PCA plot of gene expression profiles for lung tumor (red) and normal (blue) tissue samples. Each point represents one sample, and ellipses indicate the 95% confidence interval for each group.

Principal component analysis of the variance-stabilized data revealed a clear separation between tumor and normal samples. The first principal component (PC1) accounted for ~26% of the total variance and largely distinguished tumor versus normal tissues along the horizontal axis. Tumor samples clustered tightly together, while normal lung samples

grouped separately, indicating distinct global expression patterns between the two conditions. A few normal samples lay near the periphery of the tumor cluster, but overall the two sample types formed well-separated clusters. This separation in the PCA space demonstrates that gene expression profiles can clearly differentiate cancerous lung tissue from adjacent normal tissue.

MA Plot of Differential Expression

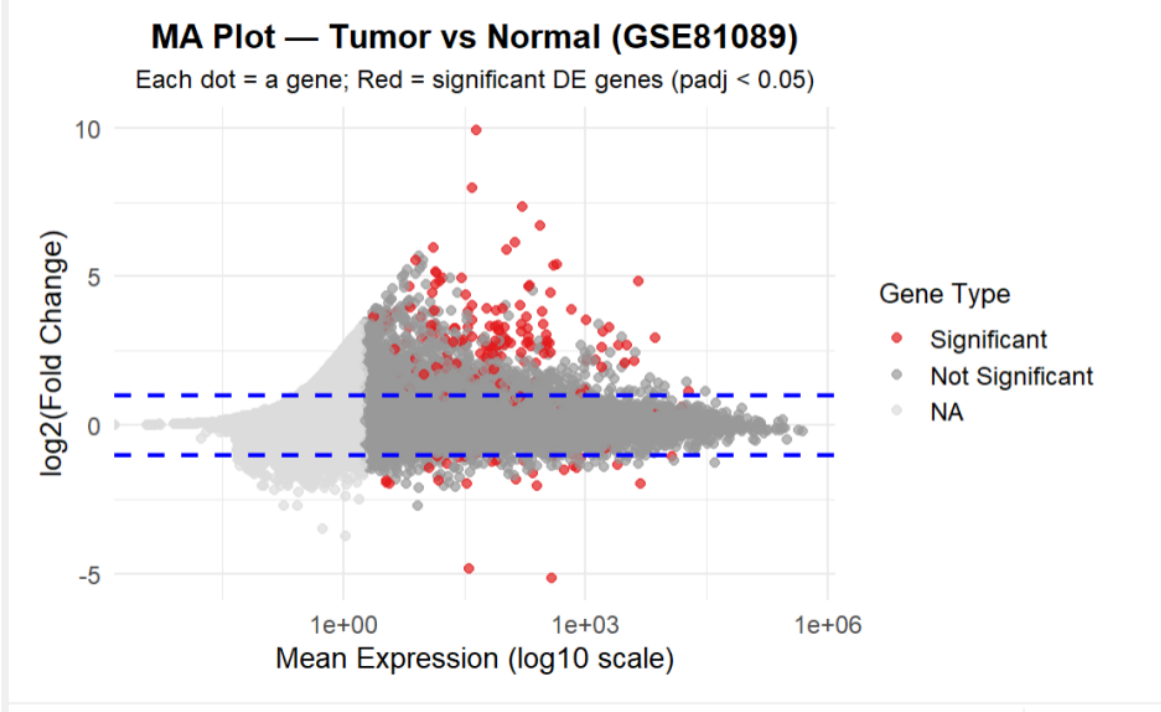

Figure 4. MA plot of tumor vs normal differential expression (DESeq2 results). Each point is a gene; red points are genes with padj < 0.05 (significantly DE), and gray points are not significant. The blue dashed lines at log2 fold change = ±1 denote a two-fold expression change.

The DESeq2 analysis identified widespread differential gene expression between tumor and normal lung tissues. An MA plot (Figure 4) illustrates the relationship between fold change and expression level for each gene. Most genes with high average expression (right side of the X-axis) are centered around log2 fold change = 0 (horizontal axis), indicating no change, whereas genes with large positive or negative fold changes are scattered toward the top or bottom. Genes highlighted in red had an adjusted p-value (padj) < 0.05, indicating statistically significant change. Notably, many red points lie outside the ±1 log2

fold change lines, meaning a substantial number of genes show greater than two-fold differences in expression. There is a slight asymmetry with more significant points having positive log2FC (above the upper dashed line) than negative, suggesting more genes are upregulated in tumors than downregulated. Overall, the MA plot confirms that numerous genes exhibit significant expression changes between normal and tumor lung tissues, especially among those with moderate to high expression levels.

Volcano Plot of Differentially Expressed Genes

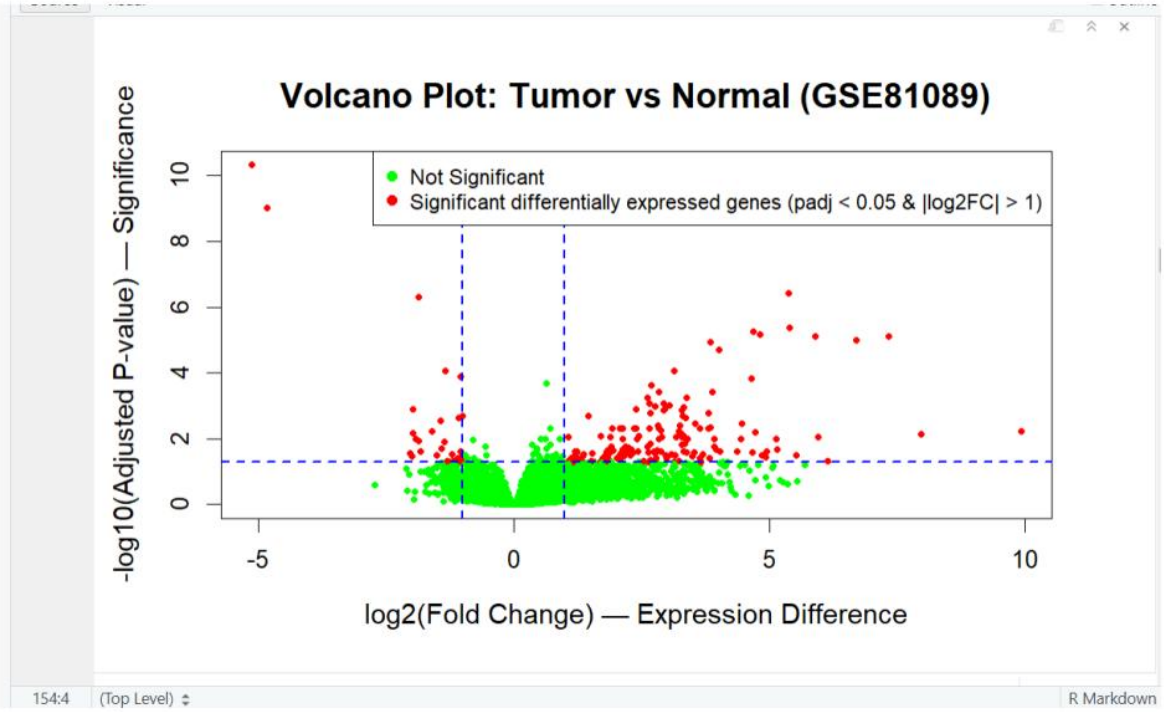

Figure 5. Volcano plot showing the significance (–log10 adjusted p-value) vs. fold change (log2) for all genes in the tumor vs normal comparison. Red points denote significantly differentially expressed genes (padj < 0.05 and |log2FC| > 1), and green points denote genes that are not significant. Blue dashed lines indicate the threshold cutoffs (adjusted p = 0.05 and log2FC = ±1)

. The volcano plot (Figure 5) provides an overview of differential expression significance versus magnitude. Each point represents one gene, with the X-axis showing the log2 fold change (tumor vs normal) and the Y-axis showing –log10(padj). Genes passing both significance and fold-change thresholds (padj < 0.05 and |log2FC| > 1) are highlighted in red. In total, this analysis identified on the order of several thousand significantly

differentially expressed genes. Approximately 6.5k genes met the significance criterion, of which roughly ~4.2k were upregulated in tumors (positive log2FC) and ~2.3k were downregulated in tumors (negative log2FC), based on the two-fold change cutoff. These significant genes appear as red points rising above the horizontal cutoff (padj = 0.05, –log10(padj) ≈ 1.3) and beyond the vertical lines at log2FC = ±1. The distribution of red points is skewed toward the right, indicating that a larger number of genes have higher expression in tumors than in normal tissue. The most extremely upregulated genes in tumors reach log2 fold changes around +9 to +10, while the most downregulated genes reach around –5, reflecting expression differences of over 500-fold and ~32-fold, respectively. Overall, the volcano plot underscores the high number of significantly altered genes and highlights those with the largest expression changes between tumor and normal lung samples.

Expression of the Top Downregulated Gene

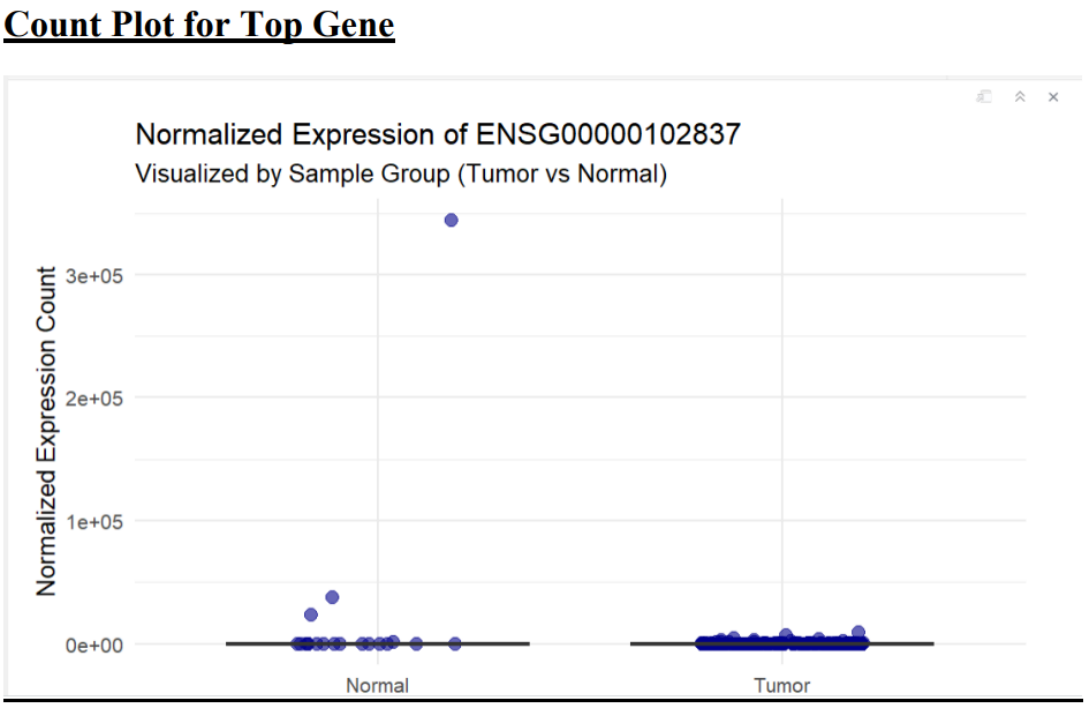

Figure 6. Normalized expression counts for the top downregulated gene ENSG00000102837 (OLFM4) in normal vs tumor samples. Each point represents an individual sample's normalized count. Horizontal black bars denote the median expression per group.

To illustrate the magnitude of expression differences at the sample level, we examined the normalized count distributions for the most downregulated gene, ENSG00000102837 (which corresponds to *OLFM4*). As shown in Figure 6, normal lung samples (blue points) consistently exhibit high expression of this gene, with some normal individuals showing extremely elevated normalized counts (on the order of 10^5–10^6). In stark contrast, the tumor samples (red points) show nearly absent expression of OLFM4 – the vast majority of tumors have close to zero normalized counts for this gene. The medians (black bars) highlight this gap: OLFM4 expression is high in normal tissue but essentially silenced in lung tumors. This dramatic loss of OLFM4 in tumors (confirming a ~32-fold reduction on average) exemplifies the profound downregulation observed for certain genes. Biologically, OLFM4 encodes an anti-apoptotic factor that is normally expressed in healthy lung epithelium; its near-complete disappearance in tumor tissue suggests a loss of a normal protective or regulatory function in the cancerous state. This pattern underscores how specific genes identified by the analysis are not only statistically significant but also exhibit clear, biologically meaningful differences that differentiate tumor from normal lung tissue.

Discussion

Interpretation of Differential Expression Results

This study revealed a distinct set of genes that were significantly upregulated or downregulated in lung tumor tissues compared to adjacent normal lung tissue. These findings reflect substantial molecular alterations that occur during lung tumorigenesis. For instance, the gene ENSG00000102837 was the most significantly downregulated in tumors, suggesting a potential tumor suppressor role. Genes that showed overexpression in tumor tissues may represent oncogenes or stress-response genes activated during cancer progression. Our visualizations, including the PCA and heatmaps, confirmed that tumor and normal tissues formed clearly separated groups based on expression profiles, indicating that the transcriptomic changes were robust and consistent.

Reference 1:
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.

This paper introduces the DESeq2 method used in our analysis. It supports the statistical reliability of our results and justifies the use of a negative binomial model for count data.

Biological Significance of Identified Genes
Many of the differentially expressed genes identified have known or suspected roles in cancer-related pathways, including cell cycle regulation, apoptosis, and immune evasion. The clustering seen in the heatmap supports the idea that these gene expression changes

are not random but reflect organized biological patterns. Some genes upregulated in tumors may serve as potential biomarkers or therapeutic targets, while downregulated genes may represent protective mechanisms that become suppressed in cancer.

Reference 2:
Cancer Genome Atlas Research Network. (2014). *Comprehensive molecular profiling of lung adenocarcinoma*. Nature, 511(7511), 543–550.

This study provides insight into how transcriptomic changes, like the ones we observed, align with key cancer pathways in lung adenocarcinoma.

Limitations and Future Directions
While our analysis used rigorous statistical and bioinformatic methods, it was limited by the unbalanced sample size between tumor (n = 199) and normal (n = 19) tissues. Although DESeq2 adjusts for this using shrinkage estimators and internal normalization, a more balanced design or paired analysis could potentially improve sensitivity. Additionally, while we identified several significant DEGs, we did not validate these experimentally. Future studies could use qPCR or single-cell RNA-seq to validate gene-level changes or explore cellular heterogeneity. Integrating clinical data (e.g., patient survival or treatment response) would also allow us to determine the prognostic value of these expression changes.

Reference 3:
Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data*. Bioinformatics, 26(1), 139–140.

This reference describes an alternative method to DESeq2, which could be considered for future comparative analysis in transcriptome studies.

Reference 4:
Subramanian, A. et al. (2005). *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.* PNAS, 102(43), 15545–15550.

This method may be used in future work to explore whether differentially expressed genes are enriched in specific biological pathways.