**Relationship between gene expression in COPD and lung cancer**

Don Chen, Gavin Sidhu, and Anand Mathi (Team 1)

CGS4144 - Introduction to Bioinformatics Algorithms

1. **Abstract**Chronic obstructive pulmonary disease (COPD) is a group of diseases which may block airflow and cause breathing problems within patients. COPD is a recognized risk factor for the development of lung cancer, with several studies finding that COPD can increase the chances of developing lung cancer by over six times in smokers, and nearly three times in non-smokers (Park, et al.). This suggests that there may be unique gene expression patterns within COPD stroma that contribute to the development of cancer cells. Here we show that there are significant gene expression patterns within COPD lung stroma containing cancer cells. We analyzed gene expression in 60 samples of lung stroma gathered from COPD patients with and without lung cancer, specifically examining tumor, tumor-adjacent, and non-tumor control tissues. After conducting statistical analyses including clustering, differential expression, and gene set enrichment analysis, we found that samples gathered from tumor samples contained several differentially expressed genes from control samples. These results demonstrate how COPD contributes as a risk factor to lung cancer by affecting gene expression in lung tissue. This research not only enhances our understanding of the interplay between COPD and lung cancer, but also lays groundwork for the development of potential treatments targeting lung cancer in COPD patients.
2. **Introduction**The dataset comprises 60 samples of lung tissue, categorized into three distinct groups, each consisting of 20 samples. The first group comprises tumor samples, the second consists of control samples, and the third encompasses tumor-adjacent samples. Each of the 60 samples in the dataset is characterized by the gene expression data of tens of thousands of genes, providing a comprehensive molecular profile for analysis and exploration. This dataset enables the investigation of gene expression patterns across different sample types, offering valuable insights into potential expression patterns distinct to each sample group.Our goal was to determine if there are specific gene expression patterns in COPD lung stroma associated with the presence of lung cancer. The first step in this process was to determine which genes were significantly differentially expressed between the tumor and control groups. To do this, we used differential expression analysis to find the log two fold change for each gene and found the genes with the greatest absolute log fold change compared to the control group.  
     
   Next, we used clustering techniques to identify structures and relationships within the data. We used K-Means, PAM, and Hierarchical Clustering techniques to find clusters within the gene expression data and test if the cluster grouping was similar to the original control/near-tumor/tumor groups. By finding a clustering model that accurately separates the tissue groups, we can accurately figure out the specific genes that influence   
     
   Finally, we employed modeling algorithms to test if the dataset could be used to accurately predict tissue types based on the gene expression data. The predictive models also allow us to pinpoint genes that significantly influence modeling outcomes. These genes could serve as biomarkers for distinguishing lung cancer and COPD, which could offer insight into diagnostic and treatment strategies.
3. **Methods**

We used several different systematic methods to conduct a comprehensive analysis of the relationship between gene expression in COPD and lung cancer. Our approach involved five different categories of algorithms: clustering, differential expression, gene set enrichment analysis, predictive modeling, and statistics. This section will only give a brief overview of the algorithms; the resulting plots and their implications are discussed under Section IV (“Results & Discussion”). All code and analysis was performed in R and is available at our project’s GitHub repository: <https://github.com/gavins53/cgs4144-proj>.

* 1. **Data Preprocessing**

The gene expression dataset we used to perform our analysis was curated for use in a similar investigation and was presented on refine.bio (Sandri, et al.). The dataset includes 60 samples of lung tissue categorized into one of three groups: non-cancerous (control), tumor-adjacent, and cancerous (tumor). All observed patients were diagnosed with COPD.

The original data associated genes with their Ensembl IDs, which were replaced with their corresponding HUGO gene symbols before being used in our analysis. Additionally, we log-scaled each entry in the expression matrix to obtain more meaningful visualizations by compressing the range of values. For several of the data analysis methods described below, the tumor-adjacent and cancerous samples were combined into one describing cancerous patients. For some special cases (clearly denoted), we used the original three categories to allow for a more complete analysis.

* 1. **Dimensionality Reduction**
     1. **Principal Component Analysis (PCA)**

PCA is a dimensionality reduction algorithm that can help to visualize clusters in data as a by-product. By reducing the dimensions but preserving variance, it can help patterns and clusters become more apparent when plotted.

The data underwent a variable stabilizing transformation (VST) before usage. We used the DESeq2 package’s “plotPCA” function, using the two divisions of cancerous and non-cancerous patients as the group-by variable.

* + 1. **t-distributed Stochastic Neighbor Embedding (t-SNE)**

t-SNE is another dimensionality reduction algorithm, however unlike PCA, it emphasizes the preservation of local relationships between samples that originally existed in high-dimensional space.

Similarly to PCA, we used the VST version of the data before feeding it to the t-SNE algorithm. We used the Rtsne package to apply the algorithm. The parameters included the seed (set to 123 for reproducibility) and the perplexity (set to 10). The perplexity determines the balance between the local and global relationships within the data. A perplexity value of 10 is relatively moderate.

* 1. **Clustering**
     1. **Partitioning Around Medoids (PAM)**

PAM clustering is a clustering approach that iteratively selects medoids as representative points, assigning samples to clusters based on their Euclidean distances.

The PAM clustering approach was performed in several different environments, each using the “pam” function from the cluster package. We applied the algorithm on the top 10, 100, 1,000, 5,000, and 10,000 most variable genes. For the 10, 100, 1,000, and 10,000 gene counts, we used five clusters. For the 5,000 gene count, we re-ran it with three, five, and 10 clusters to see how the number of clusters might impact the clustering of data points.

* + 1. **Hierarchical Clustering (h-clust)**

H-clust is another clustering algorithm that creates a dendrogram based off of the similarities in the data points, splitting and merging clusters based on the identified similarities and dissimilarities in the samples. It allows for easy visualization of the existence of relationships between certain samples in the expression matrix.

Like PAM clustering, h-clust was performed on the top 10, 100, 1,000, 5,000, and 10,000 most variable genes from the expression matrix. H-clust is unique from the other clustering algorithms in that it does not take the number of clusters as a parameter. Instead, the reader can choose the number of clusters we want by imagining a horizontal cut across any part of the dendrogram. Programmatically, we can also use the “cutree” function to exactly cut the number of clusters we want for use in comparative visualizations like heatmaps.

* + 1. **K-Means**

The final clustering approach used, k-means, is similar to PAM in that it finds cluster centers and assigns samples based on their Euclidean distances. However, k-means uses centroids (which are calculated as the means of data points) while PAM uses medoids (which are actual data points). This means that k-means may be more susceptible to outliers.

Again, the k-means algorithm was applied to the top 10, 100, 1,000, 5,000, and 10,000 most variable genes. In the 10, 100, 1,000, and 10,000 gene environments, we used five clusters. For the 5,000 gene run, we ran it three times: with two, three, and four clusters.

* 1. **Differential Expression**

Differential expression is a technique of comparing the levels of gene expression between different samples to identify which genes are significantly upregulated or downregulated in either cancerous or non-cancerous patients.

The expression matrix was filtered to only identify samples with sufficient data; the gene expression levels were filtered out if they were below 10. We used the “DESeq” function from DESeq2 to perform differential expression. There were no special parameters used.

* 1. **Gene Set Enrichment Analysis (GSEA)**
     1. **gProfiler2**

The package gProfiler2 is a package used for gene ontology enrichment analysis, which can help identify which annotations, terms, and pathways are overrepresented after performing differential expression.

We use the “gost” function with the significance parameter set to true, so as to restrict the analysis to only significantly enriched terms. Our analysis does not measure the underrepresented terms. Of course, the specimen parameter is set to look at humans.

* + 1. **topGO**

topGO is a gene ontology enrichment analysis algorithm that can identify overrepresented terms in the differential expression results.

The topGO enrichment analysis was performed by reading the differential expression results, selecting a subset of genes that were statistically significant (within a p-value of 0.05), and then using the topGO package to generate data. For the ontology parameter, we used “BP”. This limits the results to only looking at the biological processes, rather than the molecular function or cellular components of the identified genes. We used the “classic” algorithm alongside the “Fisher” test.

* + 1. **clusterProfiler**

Additionally, we use clusterProfiler for gene ontology enrichment analysis.

It looks at the log2FoldChange values, sourced from the differentially expressed gene results. We sorted the data in descending order and used the “gseGO” function from the clusterProfiler package with epsilon 1e-300. This is very small, and is helpful because it allows us to estimate the p-values very accurately.

* 1. **Predictive Modeling**
     1. **Random Forest**

RandomForest is the first of the three predictive modeling algorithms we ran on the data. For this algorithm, we divided the sample data into training and testing data. The training data consisted of the first 45 samples (in the order that it was arranged in the metadata), and the testing data consisted of the remaining 15.

We used the “randomForest” function from randomForest with an ntree value of 2001. This is the number of decision trees that the random forest will be generated with. Like with the clustering methods, we performed this algorithm on the subset of the top 10, 100, 1,000, 5,000, and 10,000 most variable genes. It should be noted that for this algorithm, we used the original three categories rather than the combined categories.

* + 1. **Logistic Regression**

We also used logistic regression, a binary classifier, to see if models could accurately predict samples as coming from cancerous or non-cancerous patients. The models were similarly divided into a training set of 45 samples and a testing set of 15.

Several functions, including “logistic\_reg”, “workflow”, and “recipe” (all from the tidymodels package) were used to apply the algorithm. It used the “glm” engine and “classification” mode. Again, this technique was run on five different setups for the top 10, 100, 1,000, 5,000, and 10,000 most variable genes.

* + 1. **Support Vector Machine (SVM)**

The final predictive modeling algorithm used was SVM. It used the same 45 samples for training and 15 samples for testing.

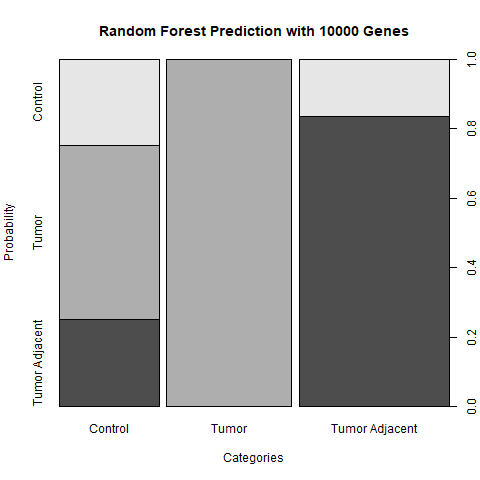
The data used “svm” from e1071 and looked specifically at the “response” type, which allowed us to view the probabilities of each class. We ran the technique for the top 10, 100, 1,000, 5,000, and 10,000 most variable genes. Like RandomForest, this method used the original three categories rather than the dual-category setup created in the preprocessing step.

* 1. **Statistics**

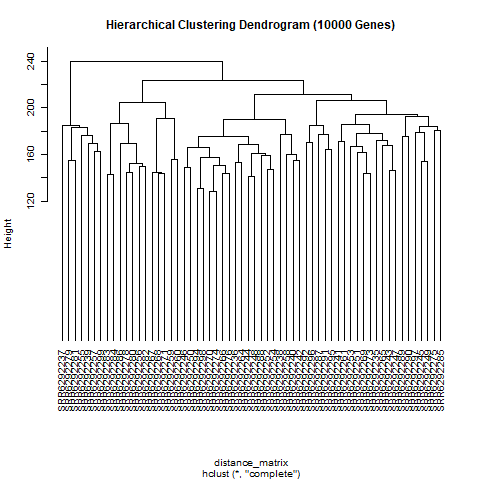
For each of the 21 previously mentioned clustering techniques, we performed a chi-squared test with the algorithm’s groupings of the samples and the real groupings obtained via the metadata. Then, we adjusted the values based on the Bonferroni test.

1. **Results & Discussion**

With the aid of the aforementioned methods, we were able to identify several patterns in the data across the 60 samples. These patterns then allowed us to draw conclusions about how COPD may impact gene expression in lung tissue such that it increases the incidence of tumorigenesis.

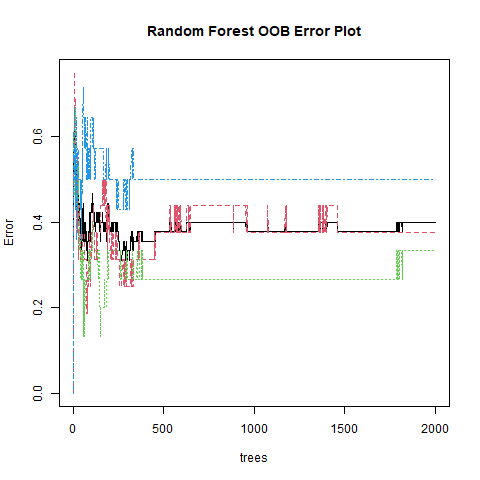


Following a comprehensive examination of the data, we identified distinctive gene expression patterns linked to the occurrence of lung cancer in individuals with COPD. While our predictive modeling exhibited suboptimal accuracy for non-tumorous samples, it demonstrated remarkable precision in predicting tumorous samples.

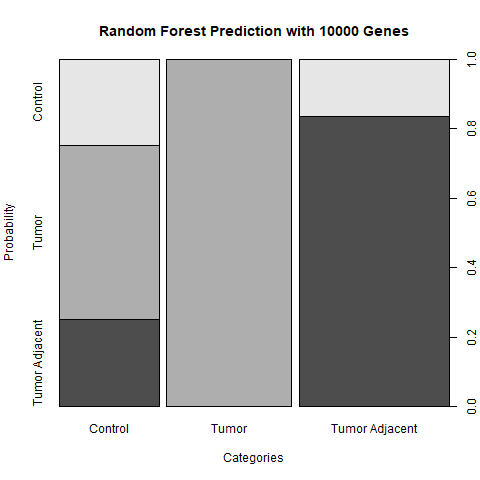
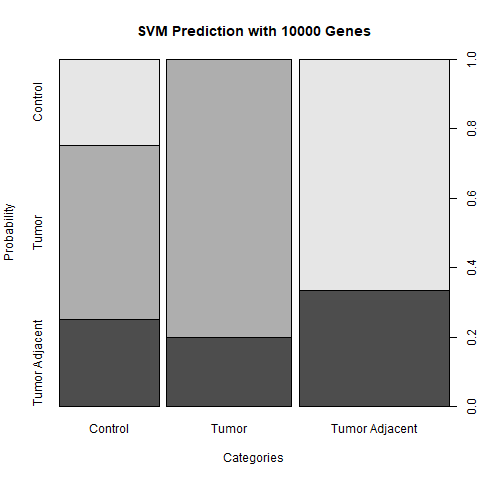


The hierarchical clustering of significantly expressed genes, visualized through dendrograms, unveiled a notable trend. A majority of these genes exhibited striking similarity, as evidenced by their concentration on a single side of the dendrogram when it was divided into two distinct trees.

Despite the similarity of the genes, it is very important to note that due to our small sample size, our results should not be used to draw any important conclusions.



An interesting example of the effects of the small sample is size in large leaps in error rate as the Random Forest algorithm runs on the data set. The shape of this graph would also change drastically with each run.



Another interesting effect of the small sample size is the large variation in prediction rates among the different predictive models, especially for the tumor adjacent group. This detail could be mitigated by combining the control and tumor adjacent group, which could make for an interesting alternative analysis to this one.

1. **Conclusion**

The purpose of this analysis was to investigate how COPD manipulates gene expression in such a way that makes patients more susceptible to lung cancer. Though our research, due to limitations in the database’s samples, was unable to explicitly identify which genes are manipulated, we were able to uncover some interesting high-level patterns and resolve several specific questions that fall under the umbrella of our hypothesis.

As demonstrated by the results, the variation in genetic expression caused by lung cancer does not seem to permeate into cells beyond the affected tumor. As such, one way the project could be improved is by either combining the tumor adjacent and non-cancerous samples into one category, while looking at the tumor samples independently. Alternatively, we could omit the algorithms that require binary data and perform a more rigorous analysis of the variation in genetic expression between the original three categories. As has been indicated in the predictive modeling results, this would make for a much more insightful investigation into the relationship between COPD and lung cancer.

For future work, one additional analysis that could be performed is network analysis. This would allow us to better understand the interplay of COPD and lung cancer by uncovering key gene pathways between our samples, providing a systems-level view of the expression matrix. A database like BioGRID would be especially useful for performing this analysis.

1. **References**

Sandri BJ, Kaplan A, Hodgson SW, Peterson M, Avdulov S, Higgins L, et al. (2018). Multi-Omic Molecular Profiling of Lung Cancer Risk in Chronic Obstructive Pulmonary Disease. <https://www.refine.bio/experiments/SRP125001/multi-omic-molecular-profiling-of-lung-cancer-risk-in-chronic-obstructive-pulmonary-disease>

Park HY, Kang D, Shin SH, et al. (2020) Chronic obstructive pulmonary disease and lung cancer incidence in never smokers: a cohort study.

<https://thorax.bmj.com/content/75/6/506>