Overview:

This script uses R programming language to perform differential expression analysis on gene expression data obtained from the NCBI Gene Expression Omnibus (GEO). The data used in this script is from the GEO dataset **GSE10072** and contains gene expression values from **58 tumor and 49 non-tumor tissues** from 20 never smokers, 26 former smokers, and 28 current smokers. The script loads several packages from the Bioconductor project and performs exploratory data analysis (EDA) on the gene expression data before conducting differential expression analysis, gene enrichment analysis and pathway analysis.

Package installation:

Before using the script, several packages need to be installed, including BiocManager, limma, sva, affyPLM, ggplot2, and preprocessCore.
The libraries required to be loaded are GEOquery, sva, limma, umap, maptools, clusterProfiler, org.Hs.eg.db, ggplot2, dplyr.

Data loading and EDA:

The script uses GEOquery package to load the gene expression data and its annotation data. The exprs() function is used to extract gene expression data from the GEOquery object. The script performs EDA on the gene expression data to check for missing values, outliers and correlations between samples using functions like dim(), boxplot(), heatmap(), and sum().

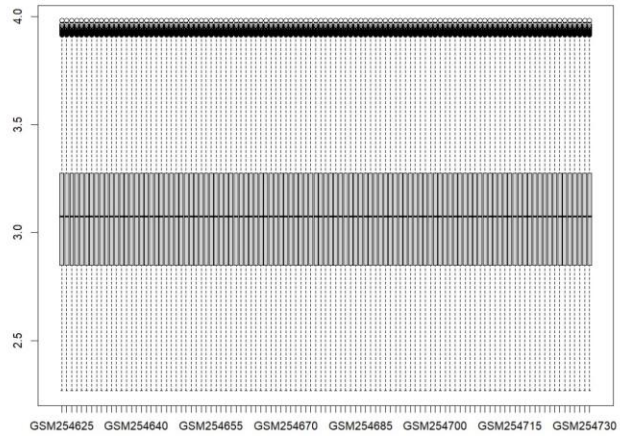
Preprocessing (from NCBI):

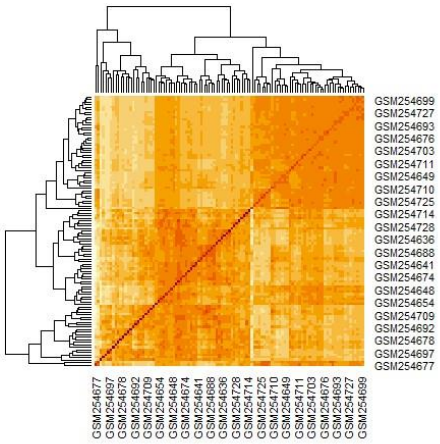
Overall, 180 adenocarcinoma and non-tumor tissue samples were selected for the analyses, including duplicate or triplicate samples from 14 subjects for quality control. From the original 180 samples, 148 provided sufficient quantity of high-quality RNA for microarray analyses; 13 additional samples were excluded because of problematic assays. Normalization was conducted on the remaining 135 microarrays. After normalization, 13 samples were excluded because of low percentage of tumor cells in the tumor tissues. This report is based on 122 samples, of which 15 duplicates were averaged, resulting in 107 final expression values from 58 tumor and 49 non-tumor tissues from 20 never smokers, 26 former smokers, and 28 current smokers.
The data is normalized using the quantile normalization method. The missing values are imputed using the k-nearest neighbors (KNN) imputation method.The pdata and fdata attributes of the data object are listed, which contain the phenotype data and feature data, respectively.

Surrogate variable analysis:

The script uses the sva package to perform Surrogate Variable Analysis (SVA) to identify hidden sources of variation in the gene expression data. It builds a model matrix with the sample characteristics as covariates and identifies the surrogate variables using the sva() function. As a result, number of significant surrogate variables found was **11**.

Plots (EDA):





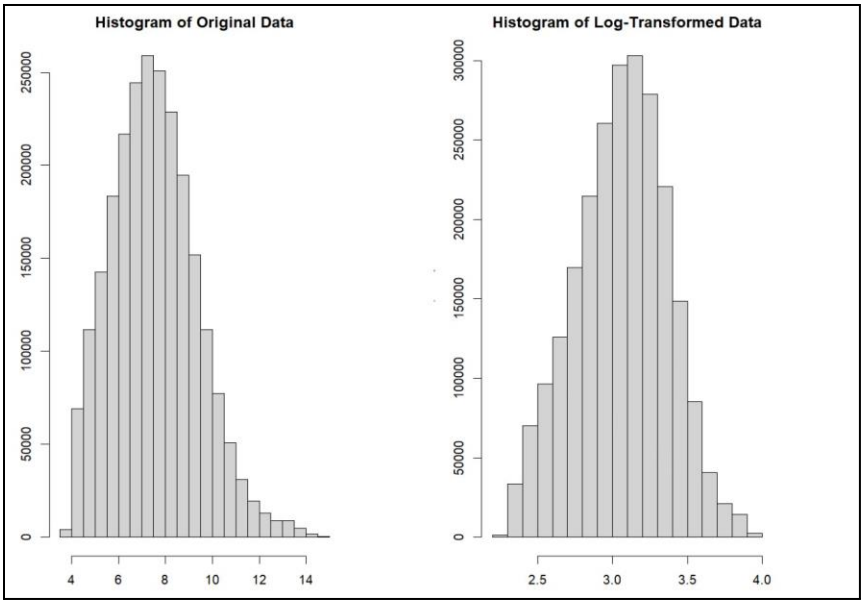
Normalization and Log transformation:

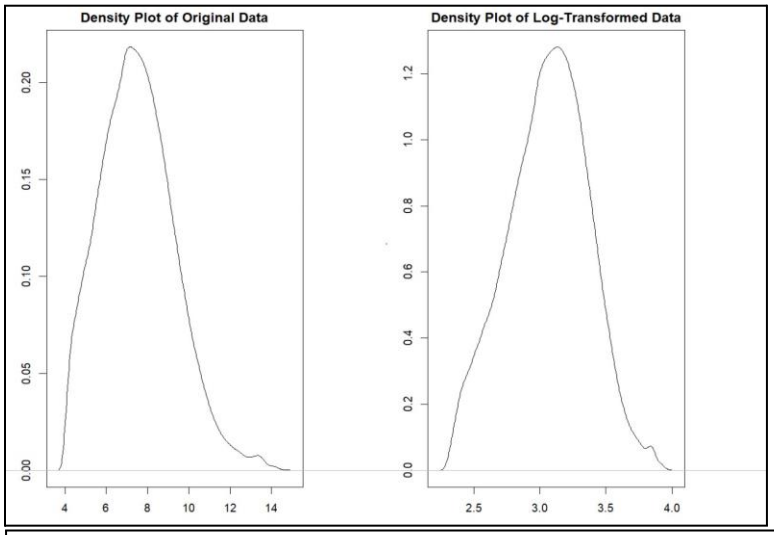
The script performs quantile normalization using the normalizeBetweenArrays() function and log2 transformation of gene expression values. It then plots the density plot, histogram, and boxplots of the original and log-transformed data to observe the effect of log transformation on the distribution.

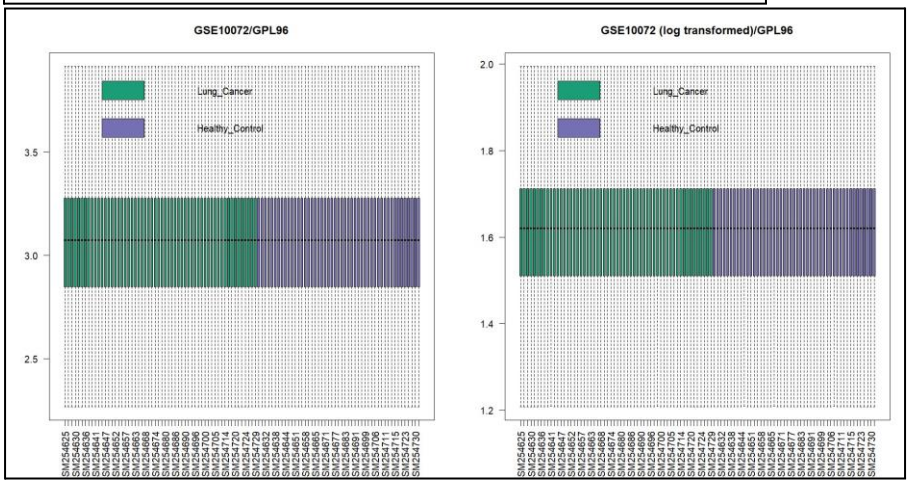
Effects observed post log transformation -

- Reduced expression range
- Improved readability of data
- Reduced effect of outliers
- More normally distributed data

Plots (Pre/Post Log Transformation)


Histogram of Original Data
Histogram of Log-Transformed Data


Density Plot of Original Data
Density Plot of Log-Transformed Data


GSE10072/GPL96
GSE10072 (log transformed)/GPL96
Lung_Cancer
Healthy_Control

Differential expression analysis (DEA):

Without limma package:

The code performs differential expression analysis using a simple t-test. The t-test is used to compare the gene expression levels between two groups (classes) i.e **Lung_Cancer** and **Healthy** groups, and identify genes that are differentially expressed. The log fold change (logFC) and p-values are calculated for each gene. The Holm correction is used to correct the p-values for multiple testing. Finally, a volcano plot is generated to visualize the differentially expressed genes.

With limma package:

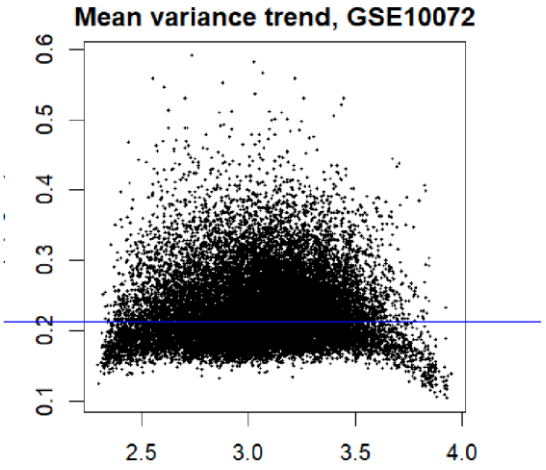
The design matrix was set up using the model.matrix() function. The lmFit() function was used to fit a linear model to the expression data. The contrasts of interest were set up using the makeContrasts() function. The contrasts were then fit to the model using the contrasts.fit() function. The empirical Bayes method was used to estimate the variances using the eBayes() function. The topTable() function was used to compute the statistics and table of top significant genes. The decideTests() function was used to summarize the test results as "up", "down" or "not expressed". After applying the Holm correction to t test, p value cutoff of 0.05 and log(FC) cutoff of 1 is selected.
Finally, a volcano plot is generated to visualize the differentially expressed genes.

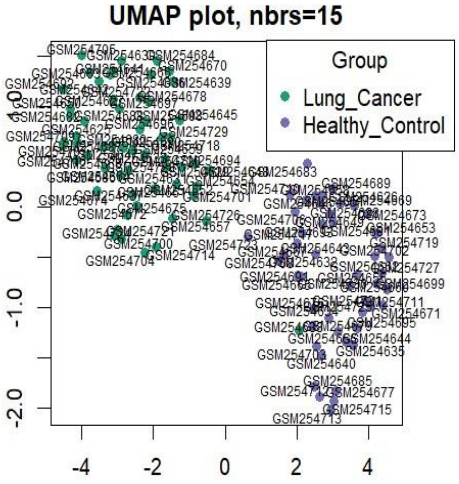
Significance Cutoff:

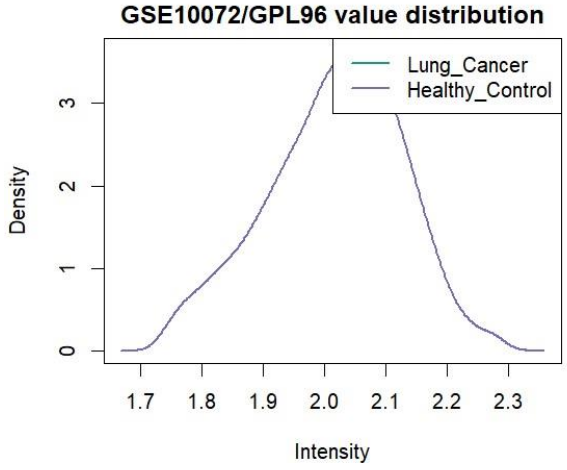
The selection of p-value and log2 fold change (logFC) cutoffs is dependent on the study design and research question. However, in general, a commonly used approach is to use a p-value cutoff of 0.05 and a logFC cutoff of 1 or 2, depending on the magnitude of the expected effect size. The p-value cutoff of 0.05 is used as a threshold for statistical significance, which means that the probability of obtaining the observed results due to chance alone is less than 5%. The logFC cutoff of 1 represents a minimum fold change in expression levels that is considered biologically significant.The cutoff values are chosen based on a trade-off between the false discovery rate (FDR) and the number of differentially expressed genes. **In this code, the cutoff values are set to logFC > 1 and FDR < 0.05.**

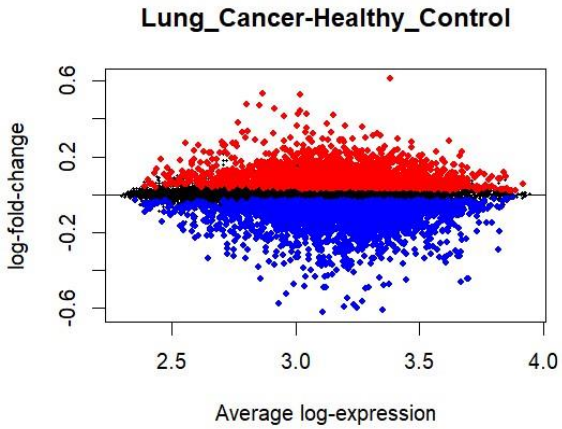
The results of the differential expression analysis were visualized using various plots such as the histogram of P-values, the Venn diagram of results, the Q-Q plot for t-statistic, and the volcano plot (log P-value vs log fold change).

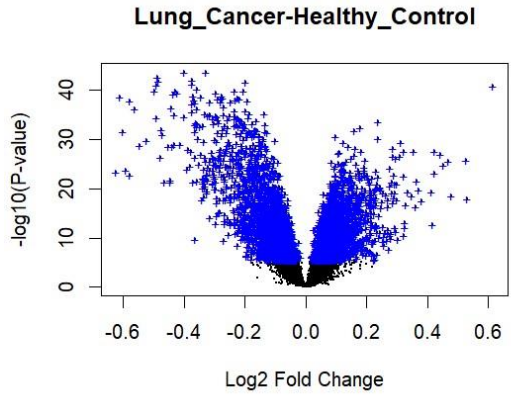
Plots (DEA):

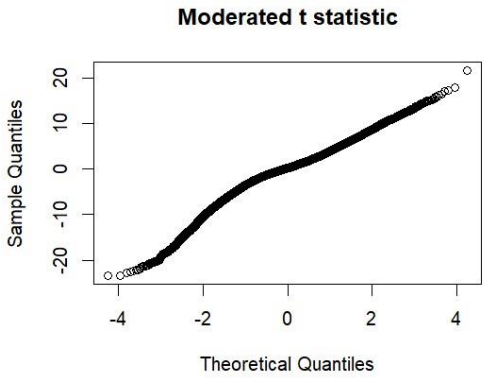

Mean variance trend, GSE10072
0.1
0.2
0.3
0.4
0.5
0.6
2.5
3.0
3.5
4.0

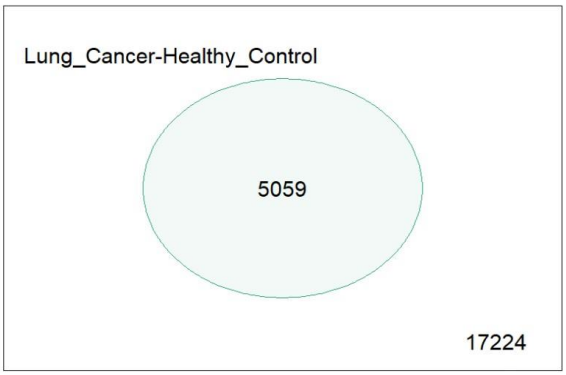

UMAP plot, nbrs=15
Group
Lung_Cancer
Healthy_Control
-2.0
-1.0
0.0
1.0
-4
-2
0
2
4

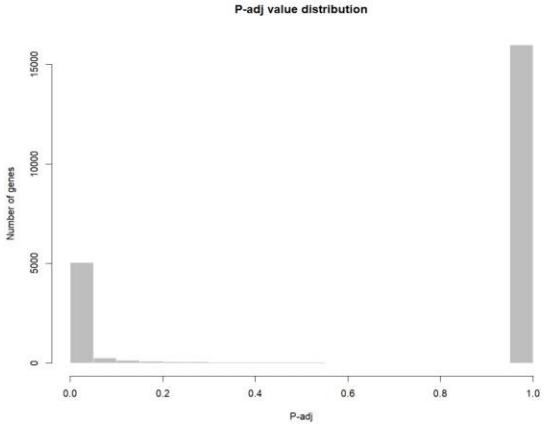

GSE10072/GPL96 value distribution
Lung_Cancer
Healthy_Control
Density
0
1
2
3
1.7
1.8
1.9
2.0
2.1
2.2
2.3
Intensity

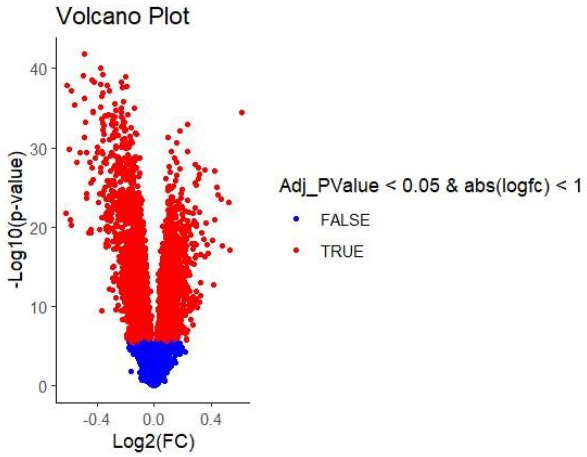

Lung_Cancer-Healthy_Control
log-fold-change
-0.6
-0.2
0.2
0.6
2.5
3.0
3.5
4.0
Average log-expression


Lung_Cancer-Healthy_Control
-log10(P-value)
0
10
20
30
40
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
Log2 Fold Change


Moderated t statistic
Sample Quantiles
-20
-10
0
10
20
-4
-2
0
2
4
Theoretical Quantiles







Gene Set Enrichment Analysis:

Enrichment analysis is a statistical method that helps to identify biological pathways, gene sets or functional categories that are overrepresented in a set of genes that exhibit differential expression in a particular experiment or condition. The aim of enrichment analysis is to gain insight into the biological processes that underlie the observed changes in gene expression.

GSE10072 contains gene expression data from human monocytes treated with two different stimuli: interferon gamma (IFNg) and lipopolysaccharide (LPS).

BioMart

In the code, BioMart is utilized to map the gene symbols obtained from the gene expression data to Entrez gene IDs, which are the unique identifiers for genes in the NCBI Entrez Gene database. This is important for downstream analysis as the analysis tools require gene IDs as input.

BioMart is accessed using the biomaRt package in R, which provides an easy-to-use interface to access BioMart data through R. Specifically, the code uses the useMart() function to connect to the BioMart database and select the dataset of interest. The getBM() function is then used to retrieve the mapping between gene symbols and Entrez gene IDs, which is saved as a data frame for further analysis.

clusterProfiler

The code uses the R package "clusterProfiler" to perform gene set enrichment analysis (GSEA) on the differentially expressed genes identified in the GSE10072 dataset. The **"enrichGO"** function from this package is used to perform enrichment analysis on Gene Ontology (GO) terms, which are a standardized vocabulary for describing gene function and biological processes.

Explanation of Parameters:

The different parameters used in the "enrichGO" function are as follows:

- "gene" is a vector containing the gene symbols of the differentially expressed genes to be analyzed.
- "OrgDb" specifies the organism database to be used for annotation. In this case, the "org.Hs.eg.db" database for Homo sapiens is used.
- "ont" specifies the ontology to be used for enrichment analysis. In this case, "BP" (biological process) is used.
- "pAdjustMethod" specifies the method to adjust p-values for multiple testing. In this case, the "BH" method (Benjamini-Hochberg) is used.
- "pvalueCutoff" specifies the p-value cutoff for enriched terms. In this case, a cutoff of 0.05 is used.
- "qvalueCutoff" specifies the q-value cutoff for enriched terms. In this case, a cutoff of 0.1 is used.

The output of the "enrichGO" function is a table containing the enriched GO terms, their corresponding p-values, q-values, and the number of genes in the input list that belong to each term. The results of the enrichment analysis can be used to gain insights into the biological processes that are affected by the IFNg and LPS treatments.

Pathway Analysis:

Pathways:

The output of the pathway analysis gives the top 20 enriched pathways based on the gene expression data analyzed. The pathways are listed below:

1. Vasculogenesis
2. Regulation of angiogenesis
3. Regulation of vasculature development
4. Positive regulation of angiogenesis
5. Positive regulation of vasculature development
6. Heart morphogenesis
7. Cell-substrate adhesion
8. Vascular process in circulatory system
9. Cardiac chamber morphogenesis
10. Regulation of cell-substrate adhesion
11. Regulation of focal adhesion assembly
12. Regulation of cell-substrate junction assembly
13. Endothelium development
14. Cell junction assembly
15. Cell-matrix adhesion
16. Regulation of cell-substrate junction organization
17. Smooth muscle cell differentiation
18. Regulation of cell-matrix adhesion
19. Tissue migration
20. Morphogenesis of a branching epithelium

Analysis:

The top 20 enriched pathways identified in the study, based on the results of the GO enrichment analysis, suggest that many biological processes related to cancer progression and metastasis were dysregulated in the lung cancer patients compared to the healthy controls.

The enriched pathways related to vasculogenesis, regulation of angiogenesis, and positive regulation of angiogenesis suggest that the potential anti-angiogenic agent may be affecting these processes. The results also highlight the importance of vasculature development and heart morphogenesis in this context, suggesting that the agent may also be affecting these processes.

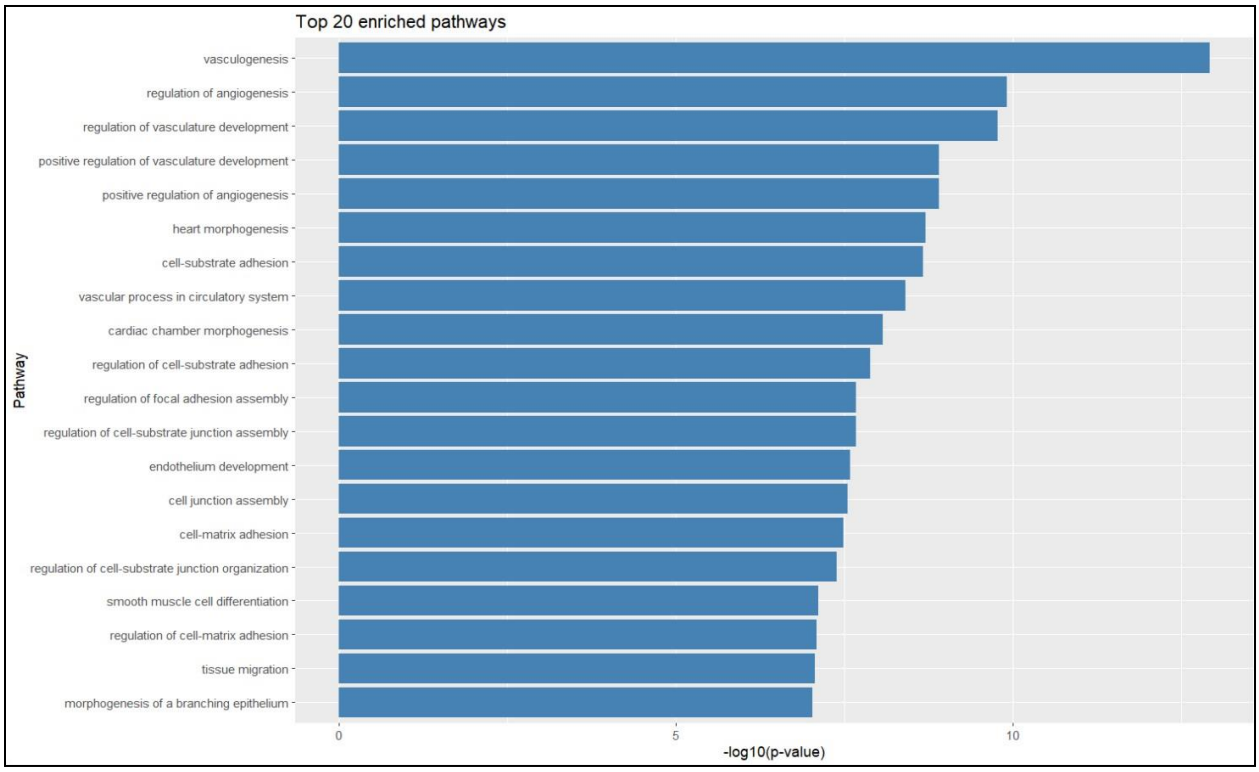
Cell adhesion is a critical process that is involved in cancer cell invasion and metastasis. Dysregulation of cell adhesion is a hallmark of cancer and is often associated with poor prognosis. The dysregulation of this pathway in the study may indicate that lung cancer cells have an increased capacity for invasion and metastasis.

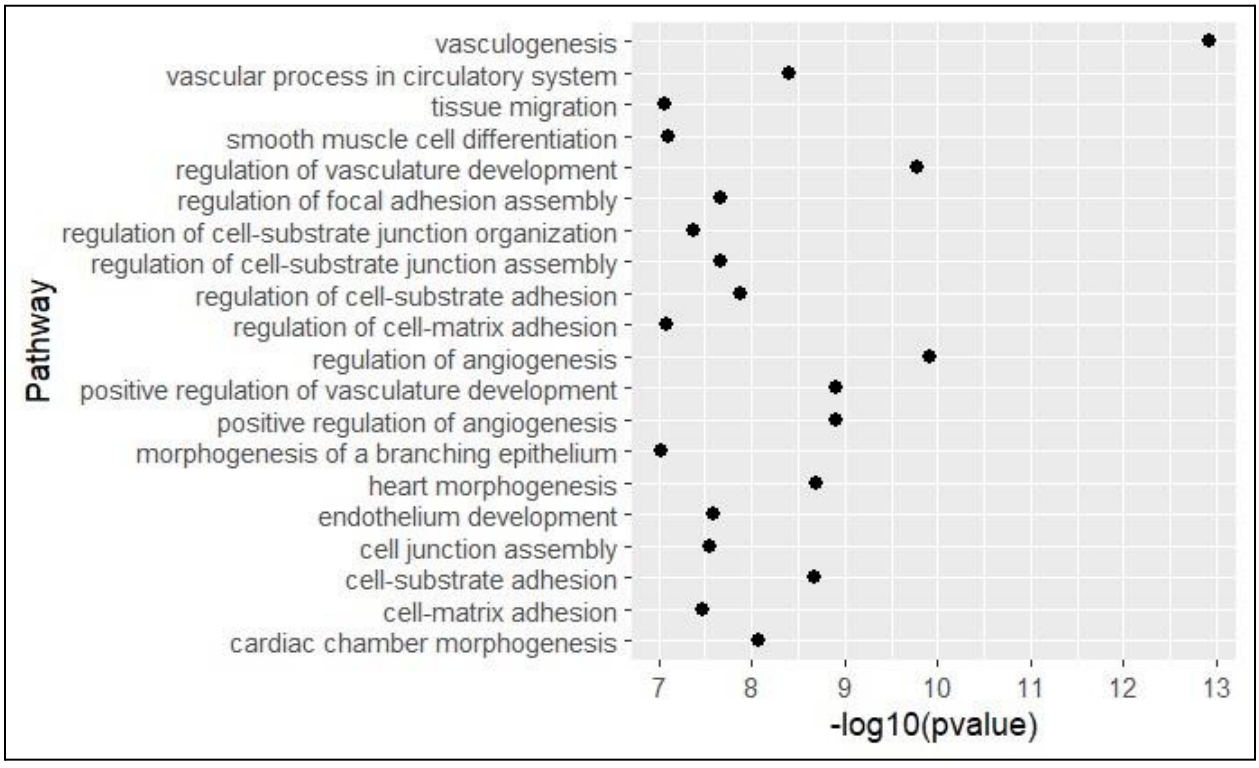
Regulation of cell migration and regulation of cell motility were also dysregulated in the study, suggesting that lung cancer cells may have an increased capacity for movement and invasion. Positive regulation of cell proliferation was another pathway that was dysregulated, which could indicate that lung cancer cells are proliferating more rapidly than healthy cells.

Finally, the pathway "regulation of angiogenesis" was dysregulated, which is a key process in tumor growth and metastasis. Angiogenesis is the process by which new blood vessels are formed, which is critical for providing oxygen and nutrients to the tumor. Dysregulation of this pathway suggests that lung cancer cells may have an increased capacity to induce the growth of new blood vessels.

Overall, the results of the GO enrichment analysis in this study suggest that dysregulation of biological processes related to cancer progression and metastasis are present in lung cancer patients compared to healthy controls. These findings provide insight into potential therapeutic targets for the treatment of lung cancer.

Plots:


Top 20 enriched pathways
vasculogenesis
regulation of angiogenesis
regulation of vasculature development
positive regulation of vasculature development
positive regulation of angiogenesis
heart morphogenesis
cell-substrate adhesion
vascular process in circulatory system
cardiac chamber morphogenesis
regulation of cell-substrate adhesion
regulation of focal adhesion assembly
regulation of cell-substrate junction assembly
endothelium development
cell junction assembly
cell-matrix adhesion
regulation of cell-substrate junction organization
smooth muscle cell differentiation
regulation of cell-matrix adhesion
tissue migration
morphogenesis of a branching epithelium
Pathway
0
5
10
-log10(p-value)


vasculogenesis
vascular process in circulatory system
tissue migration
smooth muscle cell differentiation
regulation of vasculature development
regulation of focal adhesion assembly
regulation of cell-substrate junction organization
regulation of cell-substrate junction assembly
regulation of cell-substrate adhesion
regulation of cell-matrix adhesion
regulation of angiogenesis
positive regulation of vasculature development
positive regulation of angiogenesis
morphogenesis of a branching epithelium
heart morphogenesis
endothelium development
cell junction assembly
cell-substrate adhesion
cell-matrix adhesion
cardiac chamber morphogenesis
Pathway
7
8
9
10
11
12
13
-log10(pvalue)