500 (randomized) genes

1-99

*Lung disease (under p<0.01)*

**ABCC2** [ "lung non-small cell carcinoma" ]

ALOX5 [ "asthma", "Radiation Pneumonitis", "pulmonary fibrosis", "pulmonary hypertension", "pulmonary tuberculosis", "pleuropneumonia", "exercise-induced bronchoconstriction", "Pneumococcal Pneumonia", "pulmonary edema", "**lung adenocarcinoma**" ]

ARHGAP44 [ "lung squamous cell carcinoma", "lung adenocarcinoma", "lung large cell carcinoma", "lung carcinoma" ]

**BRCA1** – lung non-small cell carcinoma, lung cancer

**BTK** [ "Pneumococcal Pneumonia", "**lung non-small cell carcinoma**", "lung adenocarcinoma", "Streptococcus pneumonia" ]

*CFTR* [ "asthma", "cystic fibrosis", "chronic obstructive pulmonary disease", "allergic bronchopulmonary aspergillosis", "pulmonary sarcoidosis", "Chronic Bronchitis", "**lung cancer**", "aspergillosis", "Staphylococcal Pneumonia" ]

**CYP24A1** [ "Lung Neoplasms", "lung non-small cell carcinoma", "lung adenocarcinoma", "lung cancer", "lung small cell carcinoma", "lung carcinoma" ]

**E2F2** [ "lung non-small cell carcinoma" ]

**GCLC** [ "cystic fibrosis", "pulmonary emphysema", "Lung Neoplasms", "lung non-small cell carcinoma" ]

*Lung neoplasms (p<0.05)*

HGF [ "malignant mesothelioma" ]

VIM [ "malignant mesothelioma" ]

100-199

*Lung disease (under p<0.01)*

**ADAM28** [ "asbestos-related lung carcinoma", "lung metastasis", "lung adenocarcinoma", "lung non-small cell carcinoma", "Lung Neoplasms" ]

ALDH18A1 [ "lung adenocarcinoma" ]

ALX4 [ "lung cancer", "Lung Neoplasms", "lung adenocarcinoma", "bronchiolo-alveolar adenocarcinoma" ]

CDH1 [ "pulmonary fibrosis", "asthma", "lung adenocarcinoma" ]

FOXP3 [ "asthma", "Lung Injury", "pulmonary tuberculosis", "bacterial pneumonia", "bronchiolitis obliterans", "Lung Neoplasms" ]

**LAMC2** [ "lung non-small cell carcinoma", "lung squamous cell carcinoma", "lung small cell carcinoma" ]

LIMCH1 [ "malignant mesothelioma" ]

**PREX2** [ "lung non-small cell carcinoma", "lung squamous cell carcinoma" ]

**ROS1** [ "lung non-small cell carcinoma", "lung sarcomatoid carcinoma", "lung adenocarcinoma" ]

**SOX30** [ "lung non-small cell carcinoma", "lung metastasis", "lung adenocarcinoma", "lung squamous cell carcinoma", "lung cancer", "Lung Neoplasms" ]

TNC [ "pulmonary hypertension", "asthma", "asbestosis", "Lung Injury", "pneumonia", "lung cancer", "pulmonary sarcoidosis", "extrinsic allergic alveolitis" ]

200-299

**CA12** [ "lung cancer", "lung metastasis", "Lung Neoplasms", "lung non-small cell carcinoma" ]

**ENO1** [ "lung adenocarcinoma", "lung non-small cell carcinoma" ]

**FAP** [ "lung non-small cell carcinoma", "lung squamous cell carcinoma" ]

FGF10 [ "interstitial lung disease", "viral pneumonia", "lung adenocarcinoma" ]

TP63 [ "lung cancer", "lung adenocarcinoma", "Lung Neoplasms" ]

TRAF4 [ "lung adenocarcinoma" ]

300-399

**ABCB1** [ "asthma", "lung non-small cell carcinoma", "pulmonary tuberculosis" ]

**BIRC5** [ "lung squamous cell carcinoma", "malignant pleural mesothelioma", "lung adenocarcinoma", "Lung Neoplasms", "lung non-small cell carcinoma", "pulmonary hypertension" ]

CEACAM1 [ "Lung Neoplasms" ]

DNMT3B [ "lung cancer", "lung small cell carcinoma", "lung adenocarcinoma" ]

**FAT1** [ "lung non-small cell carcinoma", "lung small cell carcinoma" ]

**NDC80** [ "lung non-small cell carcinoma" ]

**PTHLH** [ "lung non-small cell carcinoma" ]

**TP73** [ "lung small cell carcinoma", "lung cancer", "lung squamous cell carcinoma", "lung non-small cell carcinoma", "respiratory failure", "Lung Neoplasms" ]

400-500

CBX7 [ "Lung Neoplasms", "lung adenocarcinoma" ]

**CHGA** [ "lung non-small cell carcinoma" ]

GGT5 [ "Lung Neoplasms" ]

GNPNAT1 [ "lung adenocarcinoma" ]

**MMP11** [ "lung non-small cell carcinoma" ]

**MMP9** [ "pulmonary tuberculosis", "asthma", "Lung Neoplasms", "pulmonary emphysema", "bronchiolitis obliterans", "pneumonia", "pulmonary hypertension", "cystic fibrosis", "bacterial pneumonia", "chronic obstructive pulmonary disease", "pulmonary fibrosis", "pulmonary embolism", "interstitial lung disease", "bronchopulmonary dysplasia", "Lung Reperfusion Injury", "lymphangioleiomyomatosis", "paracoccidioidomycosis", "Hyperoxic Lung Injury", "Acute Lung Injury", "**lung non-small cell carcinoma**", "adult respiratory distress syndrome" ]

**PHGDH** [ "lung non-small cell carcinoma" ]

TF [ "lung adenocarcinoma", "malignant mesothelioma" ]

**TIMP3** [ "lung adenocarcinoma", "asthma", "lung non-small cell carcinoma" ]

SERPINA4 [ "DICER1 syndrome", "pleuropulmonary blastoma" ]

Genes of interest:

**ABCC2**

**BRCA1**

**BTK**

**CYP24A1**

**E2F2**

**GCLC**

**ADAM28**

**LAMC2**

**PREX2**

**ROS1**

**SOX30**

**CA12**

**ENO1**

**FAP**

**ABCB1**

**BIRC5**

**FAT1**

**NDC80**

**PTHLH**

**CHGA**

**MMP11**

**MMP9**

**PHGDH**

**TIMP3**

>> 24 genes of interest

Protocol

We exployed the MOET Tool in the RGD database to perform gene enrichment analysis on the first 500 (significant p<0.01) genes present in the dataset. Gene enrichment analysis focuses on the changes of expression in groups of predefined gene sets, giving insights into differentially expressed genes. To this end, we inserted 100 genes into the MOET Tool at a time, capping at 500 genes total. This analysis gave us a myriad of gene groupings, dependent on their disease associations and their (adjusted) significance. From these gene groupings, we chose the significant grouping (s) “Lung Disease” and “Lung Neoplasms” to search for genes associated with “lung non-small cell carcinoma”.

Through this approach, we found 24 genes which were associated with “lung non-small cell carcinoma”.

|  |  |  |
| --- | --- | --- |
| **Gene ID (GEO)** | **Gene** | **Function** |
|  | ABCC2 |  |
|  | BRCA1 |  |
|  | BTK |  |
|  | CYP24A1 |  |
|  | E2F2 |  |
|  | GCLC |  |
|  | ADAM28 |  |
|  | LAMC2 |  |
|  | PREX2 |  |
|  | ROS1 |  |
|  | SOX30 |  |
|  | CA12 |  |
|  | ENO1 |  |
|  | FAP |  |
|  | ABCB1 |  |
|  | BIRC5 |  |
|  | FAT1 |  |
|  | NDC80 |  |
|  | PTHLH |  |
|  | CHGA |  |
|  | MMP11 |  |
|  | MMP9 |  |
|  | PHGDH |  |
|  | TIMP3 |  |

Now we randomize:

# Read the file

df <- read.delim("Significant\_DEGs.tsv", header=TRUE, row.names=1, sep="\t", check.names=FALSE)

# Shuffle the rows of the dataframe

df <- df[sample(nrow(df)), ]

# Shuffle only the GeneExpression column (if needed)

if ("GeneExpression" %in% colnames(df)) {

df$GeneExpression <- sample(df$GeneExpression)

}

# Save the modified dataframe

write.table(df, "Randomized\_Significant\_DEGs.tsv", sep = "\t", row.names = TRUE, quote = FALSE)

# Print the first few rows to check

head(df)

Analyses randomized

1-99

**CA12** [ "lung non-small cell carcinoma", "lung cancer", "lung metastasis" ]

**FAT1** [ "lung non-small cell carcinoma", "lung small cell carcinoma" ]

**HS3ST2** [ "lung cancer", "lung non-small cell carcinoma" ]

RNASEH2A [ "lung adenocarcinoma" ]

SERPINA1 [ "DICER1 syndrome", "pleuropulmonary blastoma" ]

100-199

**DAPK1** [ "Lung Neoplasms", "lung non-small cell carcinoma" ]

ITGAV [ "lung metastasis" ]

MUC4 [ "lung adenocarcinoma", "lung carcinoma" ]

SFTPC [ "chronic obstructive pulmonary disease", "newborn respiratory distress syndrome", "interstitial lung disease", "pneumocystosis", "adult respiratory distress syndrome", "bacterial pneumonia", "asthma", "pneumonia", "cystic fibrosis", "pulmonary emphysema", "pulmonary fibrosis", "Hyaline Membrane Disease", "Pulmonary Surfactant Metabolism Dysfunction 2", "idiopathic pulmonary fibrosis", "desquamative interstitial pneumonia", "lung adenocarcinoma", "Pulmonary Surfactant Metabolism Dysfunction 1", "Surfactant Dysfunction", "pulmonary alveolar proteinosis", "interstitial lung disease 2" ]

CLDN8 [ "malignant mesothelioma" ]

**DAPK1** [ "lung non-small cell carcinoma" ]

IL2RA [ "malignant mesothelioma" ]

SELENOP [ "Lung Neoplasms" ]

200-299

**ABCC2** [ "lung non-small cell carcinoma" ]

**CDA** [ "lung non-small cell carcinoma" ]

FBL [ "lung adenocarcinoma" ]

SOX2 [ "Lung Neoplasms", "lung small cell carcinoma", "lung adenocarcinoma" ]

300-399

CHRNA9 [ "lung small cell carcinoma" ]

CNTNAP2 [ "lung small cell carcinoma" ]

PACRG [ "lung adenocarcinoma" ]

SELENBP1 [ "Lung Neoplasms" ]

**SFTPB** [ "adult respiratory distress syndrome", "newborn respiratory distress syndrome", "chronic obstructive pulmonary disease", "pneumonia", "interstitial lung disease", "respiratory failure", "asthma", "pneumocystosis", "bronchopulmonary dysplasia", "pulmonary alveolar proteinosis", "lung squamous cell carcinoma", "cystic fibrosis", "pulmonary fibrosis", "bacterial pneumonia", "Pulmonary Surfactant Metabolism Dysfunction 1", "Lung Neoplasms", "**lung non-small cell carcinoma**", "lung cancer", "Acute Lung Injury", "Congenital Deficiency of Pulmonary Surfactant Protein B", "Staphylococcal Pneumonia", "Surfactant Dysfunction" ]

400-500

ALDOA [ "lung adenocarcinoma" ]

EPHA7 [ "lung small cell carcinoma" ]

SATB2 [ "lung small cell carcinoma" ]

SULT1A1 [ "lung adenocarcinoma", "lung squamous cell carcinoma", "lung cancer" ]

**7 genes (lung non-small cell carcinoma):**

* **CA12**
* **FAT1**
* **HS3ST2**
* **DAPK1**
* **ABCC2**
* **CDA**
* **SFTPB**