|  |
| --- |
| **PROTOCOL TEMPLATE** |

LOGO

|  |  |
| --- | --- |
| **PROJECT ID** | **18 Bwanya** |
| **Title:** | **Gene Expression Analysis and Visualization of Lung Cancer Pathways Using R and Bioinformatics Pipelines** |
| **Version:** | **01** |
| **Author(s)**  **(+ initials)** | **Silke Römkens i6325117**  **Bart Bruijnen i6330097**  **Jarne Paar i6318995**  **Cecilia Mantovani i6323828**  **Sabya Sherpa i6326426**  **Anne Fleur van Groningen i6315029** |
|  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Revision / History** | | **Review** | |
| **Version** | **Change description** | **Date** | **Initials** |
| 01 | First issue |  |  |
|  |  |  |  |
|  |  |  |  |

Contents

[1. INTRODUCTION 2](#_Toc119389311)

[2. EQUIPEMENT / MATERIAL / SOFTWARE / DATA / SAMPLES (select what is applicable) 2](#_Toc119389312)

[3. HEALTH AND SAFETY (if applicable) 2](#_Toc119389313)

[4. SPECIFIC RECOMMENDATIONS / WARNING (if applicable) 2](#_Toc119389314)

[5. PROCEDURE TO FOLLOW 2](#_Toc119389315)

[6. DATA ANALYSIS AND STATISTICS (if applicable) 3](#_Toc119389316)

[7. LITERATURE 4](#_Toc119389317)

[8. APPENDIX (if required) 4](#_Toc119389318)

|  |
| --- |
| 1. INTRODUCTION |

This protocol describes the steps involved in analyzing differentially expressed genes (DEGs) in lung cancer using RNA-seq data. In the project, we focus on analyzing the publicly available dataset GSE1089 to identify significant DEGs between tumor and healthy tissues. We aim to gain more insight into biological pathways and genes relevant to lung cancer.

* **Write some objectives we have**

We try to look at risk factors and characteristics of non-small lung cancer to identify differential gene expression and consequently guide future interventions.

|  |
| --- |
| 2. EQUIPEMENT / MATERIAL / SOFTWARE / DATA / SAMPLES (select what is applicable) |

Detail here specific instrumentation / material / software / data or biomedical samples you obtained or used to perform the experiments.

*Here are some examples…*

|  |  |  |
| --- | --- | --- |
| **Name** | **Description** | **Supplier / Reference** |
| Leica 3001 Fluorescence microscope | Microscope equipped for fluorescence microscopy | Leica Ltd., Tours, France |
| Lung tissue samples | Lung tissue samples from patients with severe COVID19 infection and healthy controls. Samples were obtained within a clinical trial, permission of METC granted (permission number 2022-8299) | MUMC, Department of pulmonology |
| Cytoscape (version 9.2) | Software for creation and analysis of networks. | <https://cytoscape.org/>  Shannon et al. 2013 |
| GEO-E-32998 | Transcriptomics dataset comparing SARS-CoV2 infected lung epithelial cells with healthy control, originally published by Doe et al. 2021 and available on GEO database. | Doe et al. 2021  GEO-E-32998 |
| MES buffer solution | PH buffered MES (2-(N-morpholino)ethanesulfonic acid) solution. | BASF, Ludwigsburg, Germany |

|  |
| --- |
| 3. HEALTH AND SAFETY (if applicable) |

We will be working entirely with computers thus this section is not applicable

-> **posture, right computer brightness, taking breaks**

|  |
| --- |
| 4. SPECIFIC RECOMMENDATIONS / WARNING (if applicable) |

We are using GitHub, so we have to take caution with working on our code together. If we work individually, we have push our code not together otherwise the GitHub will not accept multiple edits at once and some of the code will be lost.

|  |
| --- |
| 5. PROCEDURE TO FOLLOW |

These are the procedure to follow to get results

1. Download the lung cancer data from GEO database​ to perform the analysis of the data
2. Install and require libraries in R, the libraries we will be using are DESeq2, ggplot2, dplyr, pheatmap, clusterProfiler, org.HS.eg.db and GEOquery.
3. Load the count matrix, it has the data from step 1
4. Load metadata, this can be required from the R library GEOquery
5. Create a DESeq2 dataset in R to create a statistical model
6. Perform Quality control
7. Get results and sort by pValue to get the significant genes.
8. Filter significant genes with adjusted p-value < 0.05 to account for false positives.
9. Visualize results through plots in R. for example the volcano plot, heatmap, or MA plot.
10. Lastly, functional enrichment analysis using GO, KEGG to analyze high-throughput experimental results.

|  |
| --- |
| 6. DATA ANALYSIS AND STATISTICS (if applicable) |

For the analysis of the normalized gene expression data (GSE81089\_FPKM\_cufflinks.tsv), we used R and the DESeq2 package. DESeq2 applies a statistical model to assess differences in gene expression between two or more sample groups. In our case, it assessed the difference between the people with NSCLC and normal lung tissue. It first estimates the variance of gene expression levels and fits a negative binomial distribution to each gene, accounting for the over-dispersion inherent in sequencing data. This approach improves the accuracy of p-values and false discovery rate (FDR) estimates.

To identify differentially expressed genes (DEGs), DESeq2 calculated a p-value for each gene, representing the likelihood that the observed differences in expression occur by random chance. A threshold is then applied (typically adjusted p-value < 0.05) to determine statistically significant DEGs.

|  |
| --- |
| 7. LITERATURE |

* Introduction
* Packages references

|  |
| --- |
| 8. APPENDIX (if required) |

* Put the code we have thus far