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

| **PROJECT ID** | **p18 Bwanya 2** |
| --- | --- |
| **Title:** | **The role of differentially expressed genes in NSCLC progression** |
| **Version:** | **02** |
| **Author(s)**  **(+ initials)** | **Baars, S.S.**  **Bayrak, O**  **Güzel, S**  **Lemmens, S**  **Ritzen, C.J.T.**  **Staps, V** |
|  |  |

| **Revision / History** | | **Review** | |
| --- | --- | --- | --- |
| **Version** | **Change description** | **Date** | **Initials** |
| 01 | First issue | 13/02/2024 |  |
| 02 | Feedback was incorporated | 20/02/2024 |  |
|  |  |  |  |

**Contents**

[1. INTRODUCTION 2](#_gjdgxs)

[2. EQUIPEMENT / MATERIAL / SOFTWARE / DATA / SAMPLES 2](#_30j0zll)

[3. HEALTH AND SAFETY 3](#_1fob9te)

[4. SPECIFIC RECOMMENDATIONS / WARNING 3](#_3znysh7)

[5. PROCEDURE TO FOLLOW 3](#_2et92p0)

[6. DATA ANALYSIS AND STATISTICS 4](#_3dy6vkm)

[7. LITERATURE 5](#_1t3h5sf)

[8. APPENDIX 5](#_4d34og8)

| 1. INTRODUCTION |
| --- |

Lung cancer is one of the leading causes of cancer-related mortality, with non-small cell lung cancer (NSCLC) being the most prevalent subtype (1,2). Investigating gene expression in NSCLC can provide insights into the molecular changes associated with tumor development and progression. This research protocol describes the bioinformatics approach used to analyze RNA sequencing (RNA-seq) data from the publicly available GEO dataset GSE81089.

RNA-seq has emerged as an advanced method for studying gene expression, particularly by identifying differentially expressed genes (DEGs). Due to its high sensitivity and comprehensive view of transcriptome-wide changes, RNA-seq offers a more detailed analysis of gene expression profiles compared to traditional methods like microarrays or PCR. This enables the identification of DEGs that may play significant roles in disease progression. For example, in cancer, DEGs can help uncover mechanisms involved in tumorigenesis (3,4).

Our study focuses on three genes of interest - EGFR (Epidermal Growth Factor Receptor), BRAF (B-Raf Proto-Oncogene, Serine/Threonine Kinase), and LKB1 (Liver Kinase B1) - which have been implicated in lung cancer. The analysis will be conducted using R (version 4.2.2) and RStudio (version 2024.12.1+563), with DESeq2 for differential gene expression analysis. Additional R packages, such as ggplot2 and dplyr, will be employed for data visualization and processing.

In our project, we aim to explore how DEGs, particularly EGFR, BRAF, and LKB1, contribute to the molecular mechanisms underlying NSCLC progression by analyzing RNA-seq data from tumor and adjacent normal tissues, specifically utilizing the publicly available GEO dataset GSE81089. By applying bioinformatics approaches, we will identify the key DEGs and their potential roles in lung cancer progression, providing insights into the molecular mechanisms involved. The main question driving this research is how these genes, through their differential expression, contribute to the progression and molecular mechanisms of NSCLC.

| 2. EQUIPEMENT / MATERIAL / SOFTWARE / DATA / SAMPLES |
| --- |

| **Name** | **Description** | **Supplier / Reference** |
| --- | --- | --- |
| R (version 4.2.2) | Software for data analysis and statistical computing | Windows:<https://cran.r-project.org/bin/windows/base/>  Mac:<https://cran.r-project.org/bin/macosx/> |
| RStudio (version 2024.12.1+563) | Integrated development environment for R | Windows:<https://download1.rstudio.org/electron/windows/RStudio-2024.12.0-467.exe>  Mac:<https://download1.rstudio.org/electron/macos/RStudio-2024.12.0-467.dmg> |
| GSE81089 | RNAseq dataset comparing Non-small cell lung cancer (NSCLC) tissue with healthy control, originally published by Mezheyeuski et al. 2018 and available on GEO database | Mezheyeuski et al. 2018  GSE81089 |
| DESeq2 (version 1.46.0) | Statistical software package for analysis of RNA sequencing data | Windows and Mac: <https://bioconductor.org/packages/release/bioc/html/DESeq2.html> |
| Ggplot2 (version 3.5.1) | Data visualization package for R | Windows and Mac: <https://ggplot2.tidyverse.org/> |
| Dplyr (version 1.1.4) | R-package to manipulate and summarize data | Windows and Mac: <https://cran.r-project.org/web/packages/dplyr/readme/README.html> |
| GEOquery (version 2.74.0) | R-package to facilitate the retrieval and processing of data | Windows and Mac: <https://www.bioconductor.org/packages/release/bioc/html/GEOquery.html> |
| BiocManager (version 1.30.25) | R-package to facilitate the installation and management of Bioconductor packages | Windows and Mac: <https://www.bioconductor.org/install/> |

| 3. HEALTH AND SAFETY |
| --- |

Since this project is entirely computational, the main health and safety considerations are related to prolonged screen time and desk work. Spending long hours at a computer can lead to eye strain, muscle tension, and fatigue, especially without proper posture or breaks. To reduce the risk of discomfort, an ergonomic workstation setup is important, with a good chair and screen positioning to prevent back and neck strain. Blue light exposure from screens can also contribute to eye fatigue, which can be managed using blue light filters or taking regular breaks. Following the 20-20-20 rule - looking at something 20 feet away for 20 seconds every 20 minutes - can help prevent digital eye strain. Taking short walks or stretching throughout the day can also improve circulation and maintain focus. While this project doesn’t involve hazardous materials or lab work, maintaining a comfortable and healthy work environment is essential for staying productive.

| 4. SPECIFIC RECOMMENDATIONS / WARNING |
| --- |

To run the software used in the procedure, a ram memory of 16gb or higher is recommended.

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

To begin the analysis, R and Rstudio must be installed. The appropriate version of R should be downloaded based on the type of the system, either for Mchips or Intel. Afterwards, The GEO dataset should be downloaded for Rstudio. After opening Rstudio, firstly the necessary libraries are downloaded and the working directory is set. Next, the RNA-seq expression data from the GEO dataset is imported. Metadata is then retrieved and organised by renaming columns and selecting specific variables such as tumor type, age, sex, smoking status, and disease stage.

Moreover, to prepare the dataset, the last row is removed and dimensions are confirmed. The genes EGFR, BRAS and LKB1 are chosen and the data is changed into a long format. Additionally, the data is also merged with metadata to integrate sample information. Lastly, the p-value is set to 0.05 for the statistical significance for the expression analysis.

| 6. DATA ANALYSIS AND STATISTICS |
| --- |

For this study, the focus is on analyzing transcriptomic expression in NSCLC tissues to understand how certain expressed genes contribute to cancer progression and how this will help in identifying potential treatment options. In the analysis, distinctions are made between individuals with cancerous and non-cancerous tissues, smoking status (never, former/current or missing), gender, tumor stage, life status (deceased or not), and age. Furthermore, the genes of interest are specified: EGFR (Epidermal Growth Factor Receptor), BRAF (B-Raf Proto-Oncogene, Serine/Threonine Kinase), and LKB1 (Liver Kinase B1). These genes exhibit diverse functions, providing insights into their function in cell division, cell growth and energy metabolism within cancerous tissues (5, 6, 7). These insights give a deeper understanding of tumour biology and its treatment options.

To analyze the data, RStudio is used, a powerful platform for statistical computing and visualization, along with the DESeq2 package for differential gene expression analysis (8). DESeq2 is specifically chosen for its ability to analyze RNA-seq count data using the Wald test as a default, which allows assessment of variables such as tumor stage, smoking intensity, or sex influence the expression of these genes. This approach also makes it possible to account for confounding variables, ensuring that the statistical results are robust and reliable. Comparisons are made between categorical variables, such as cancerous versus non-cancerous tissues and gender differences, while correlations are assessed with continuous variables, such as age.

The results are presented using a variety of visualizations generated in R. Boxplots and violin plots will be used to illustrate gene expression differences across groups, such as between male and female participants or across tumor stages, while bar graphs display the expression levels of specific genes, including EGFR, BRAF, and LKB1, across different samples. These visualizations help to effectively communicate the findings and highlight key trends in the data.

| 7. LITERATURE |
| --- |

1. Leiter A, Veluswamy RR, Wisnivesky JP. The global burden of lung cancer: current status and future trends. Nat Rev Clin Oncol [Internet]. 2023;20(9):624–39. Available from: http://dx.doi.org/10.1038/s41571-023-00798-3
2. Non-small cell lung cancer (NSCLC) [Internet]. Yale Medicine. [cited 2025 Feb 19]. Available from: <https://www.yalemedicine.org/clinical-keywords/non-small-cell-lung-cancer>
3. Difference between PCR and RNA-seq [Internet]. Cd-genomics.com. [cited 2025 Feb 19]. Available from: https://rna.cd-genomics.com/resource/difference-between-pcr-rna-seq.html
4. Han Y, Gao S, Muegge K, Zhang W, Zhou B. Advanced applications of RNA sequencing and challenges. Bioinform Biol Insights [Internet]. 2015;9s1:BBI.S28991. Available from: http://dx.doi.org/10.4137/bbi.s28991
5. Voldborg BR, Damstrup L, Spang-Thomsen M, Poulsen HS. Epidermal growth factor receptor (EGFR) and EGFR mutations, function and possible role in clinical trials. Annals Of Oncology. 1 december 1997;8(12):1197–206. Available on: <https://doi.org/10.1023/a:1008209720526>
6. Ottaviano M, Giunta E, Tortora M, Curvietto M, Attademo L, Bosso D, e.a. BRAF Gene and Melanoma: Back to the Future. International Journal Of Molecular Sciences [Internet]. 27 maart 2021;22(7):3474. Available on:<https://doi.org/10.3390/ijms22073474>
7. Shackelford DB, Shaw RJ. The LKB1–AMPK pathway: metabolism and growth control in tumour suppression. Nature Reviews Cancer [Internet]. 24 juli 2009;9(8):563–75. Available on:<https://doi.org/10.1038/nrc2676>
8. Love MI, Huber W, Anders S. Differential analysis of count data – the DESeq2 package. Genome Biology [Internet]. 5 december 2014;15(12). Available on: https://doi.org/10.1186/s13059-014-0550-8

| 8. APPENDIX |
| --- |

### Appendix 1 - R script for data analysis

# This is a script for RNA-seq analysis

# Load libraries for RNA-seq Data Analysis

*if (!require("BiocManager", quietly = TRUE))*

*install.packages("BiocManager")*

*BiocManager::install("DESeq2")*

*library(DESeq2)*

*BiocManager::install("GEOquery")*

*library(GEOquery)*

*library(dplyr)*

*library(tidyverse)*

# Set your working directory

*setwd("")*

# Step 1. import data and metadata in R

# Load your Data

*Data <- read.delim("FPKM\_cufflinks.tsv", header=TRUE, row.names=1, sep="\t", check.names = FALSE)*

# Get metadata

*gse <- getGEO(GEO = 'GSE81089', GSEMatrix = TRUE)*

*metadata <- pData(phenoData(gse[[1]]))*

*head(metadata) #have a glimpse of what metadata looks like*

# Rename metadata columns

*metadata <- metadata %>%*

*rename(*

*Sample = `tumor (t) or normal (n):ch1`,*

*Source = source\_name\_ch1,*

*Tumor\_stage = `stage tnm:ch1`,*

*Age = `age:ch1`,*

*Sex = `gender:ch1`,*

*Life\_Status = `dead:ch1`,*

*Smoking\_Status = `smoking:ch1`*

*)*

# Check updated column names

*print(colnames(metadata))*

# Select only the renamed columns and overwrite metadata

*metadata <- metadata %>%*

*select(Sample, Source, Tumor\_stage, Age, Sex, Life\_Status, Smoking\_Status)*

# Check if the changes were applied

*head(metadata)*

# Remove the last row from data

*head(Data)*

*dim(Data)*

*Data <- Data[-nrow(Data), ]*

# Check if the last row is removed

*dim(Data) # Check new dimensions*

# Explore visualization of the data

# Perform quick visualization of known lung cancer genes:

# EGFR, BRAF & LKB1

# Define the genes of interest

*genes\_of\_interest <- c("ENSG00000146648", "ENSG00000118046", "ENSG00000157764")*

# Convert Data to a long format (genes in rows, samples in columns)

*expression\_long <- Data %>%*

*as.data.frame() %>%*

*rownames\_to\_column("Gene") %>%*

*filter(Gene %in% genes\_of\_interest) %>%*

*pivot\_longer(cols = -Gene, names\_to = "Sample", values\_to = "Expression")*

# Merge expression data with metadata to include sample information

*expression\_long <- expression\_long %>%*

*left\_join(metadata, by = "Sample")*

# View the structure of the transformed data

*print(head(expression\_long))*

# Plot expression levels of selected genes

*ggplot(expression\_long, aes(x = Sample, y = Expression, fill = Gene)) +*

*geom\_bar(stat = "identity", position = "dodge") +*

*facet\_wrap(~ Gene, scales = "free\_y") + # Separate plots per gene*

*theme\_minimal() +*

*labs(title = "Gene Expression Levels Across Samples",*

*x = "Sample",*

*y = "Expression Level") +*

*theme(axis.text.x = element\_text(angle = 90, hjust = 1)) # Rotate sample labels*

# Boxplot of gene expression by sex

*ggplot(expression\_long, aes(x = Sex, y = Expression, fill = Sex)) +*

*geom\_boxplot(alpha = 0.7, outlier.shape = NA) + # Transparent boxplot*

*geom\_jitter(width = 0.2, alpha = 0.6) + # Adds individual points for visibility*

*facet\_wrap(~ Gene, scales = "free\_y") + # Separate plots for each gene*

*theme\_minimal() +*

*labs(title = "Gene Expression Levels by Sex",*

*x = "Sex",*

*y = "Expression Level") +*

*theme(axis.text.x = element\_text(angle = 45, hjust = 1)) # Rotate labels for readability*

# Alternatively, you can use a violin plot

*ggplot(expression\_long, aes(x = Smoking\_Status, y = Expression, fill = Smoking\_Status)) +*

*geom\_violin(alpha = 0.6) + # Shows density of expression levels*

*geom\_jitter(width = 0.2, alpha = 0.7) + # Adds individual sample points*

*facet\_wrap(~ Gene, scales = "free\_y") + # Separate plots per gene*

*theme\_minimal() +*

*labs(title = "Gene Expression Levels by Smoking\_Status",*

*x = "Smoking\_Status",*

*y = "Expression Level") +*

*theme(axis.text.x = element\_text(angle = 45, hjust = 1))*