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

|  |  |
| --- | --- |
| **PROJECT ID** | **p18 Bwanya 2** |
| **Title:** | **Gene expression analysis** |
| **Version:** | **01** |
| **Author(s)**  **(+ initials)** | **Baars, S**  **Bayrak, O**  **Güzel, S**  **Lemmens, S**  **Ritzen, C**  **Staps, V** |
|  |  |

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

**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 remains one of the leading causes of cancer-related mortality, emphasizing the need for a better understanding of its molecular mechanisms [Leiter et al.]. Various techniques, such as microarrays and PCR, have been used to study gene expression. However, RNA sequencing (RNA-seq) has emerged as a more advanced method due to its high sensitivity and ability to provide a comprehensive view of transcriptome-wide changes. This capability is particularly valuable for identifying differentially expressed genes (DEGs) involved in disease progression [Han et al.][CD Genomics].

In our project, we focus on identifying significant DEGs using RNA-seq data from tumor and adjacent normal lung tissues, specifically utilizing the publicly available GEO dataset GSE81089. We aim to explore these DEGs and their potential roles in lung cancer-related pathways by applying bioinformatics approaches. To facilitate this analysis, we will establish a computational pipeline using R for differential expression analysis, Bash for automation, Conda for environment management, and Git/GitHub for version control.

1. Leiter, A., Veluswamy, R.R. & Wisnivesky, J.P. The global burden of lung cancer: current status and future trends. *Nat Rev Clin Oncol* **20**, 624–639 (2023). https://doi.org/10.1038/s41571-023-00798-3

2. Han Y, Gao S, Muegge K, Zhang W, Zhou B. Advanced Applications of RNA Sequencing and Challenges. Bioinform Biol Insights. 2015 Nov 15;9(Suppl 1):29-46. doi: 10.4137/BBI.S28991. PMID: 26609224; PMCID: PMC4648566.

3. [Difference between PCR and RNA-seq - CD Genomics](https://rna.cd-genomics.com/resource/difference-between-pcr-rna-seq.html)

|  |
| --- |
| 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) | Software 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 |

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

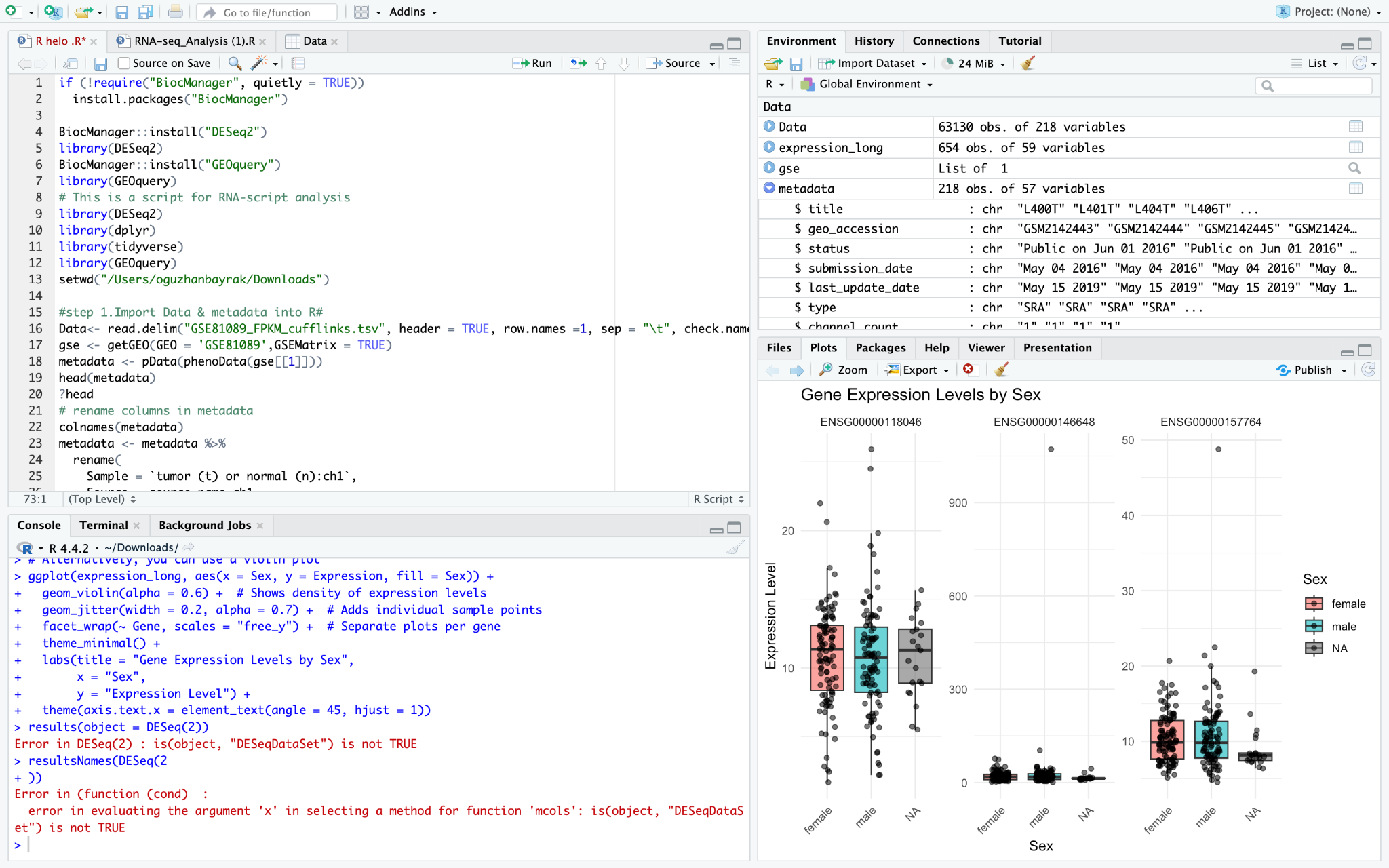
Not applicable

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

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

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

1. Install and download R
   1. ForMchips:<https://cran.r-project.org/bin/macosx/big-sur-arm64/base/R-4.4.2-arm64.pkg>
   2. For intel: <https://cran.r-project.org/bin/macosx/big-sur-x86_64/base/R-4.4.2-x86_64.pkg>
2. Download and install Rstudio
   1. For Mac: <https://download1.rstudio.org/electron/macos/RStudio-2024.12.0-467.dmg>
   2. For Windows: <https://download1.rstudio.org/electron/windows/RStudio-2024.12.0-467.exe>
3. Download the data from the GEO database

1. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81089>
2. Install and load the necessary libraries in R
3. DESeq2
4. ggplot2
5. dplyr

1. Set the working directory of the data analysis
2. *setwd(...)*
3. Import the earlier obtained data from GEO into R
4. *Data <- read.delim ("FPKM\_cufflinks.tsv", header=TRUE, row.names=1, sep="\t", check.names = FALSE)*
5. Obtain metadata from the obtained Data
   1. *gse <- getGEO(GEO = 'GSE81089', GSEMatrix = TRUE)*

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

* 1. Use *head(metadata)* to view what the metadata looks like

1. Rename the metadata columns
   1. *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`*

*)*

* 1. Use *print(colnames(metadata))* to check the updated column names

1. Select only the renamed columns from step 8 and overwrite the metadata
   1. metadata <- metadata %>%

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

* 1. Use *head(metadata)* to check if the changes were applied

1. Remove the last row from the data
2. *head(Data)*

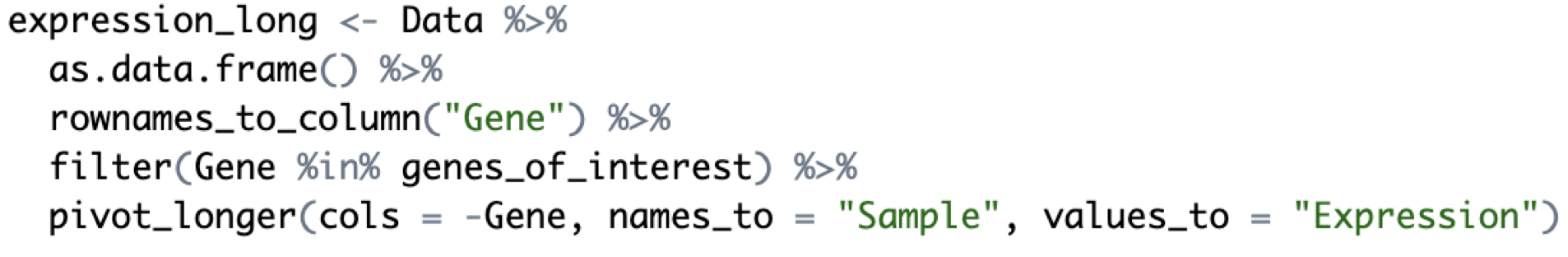
*dim(Data)*

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

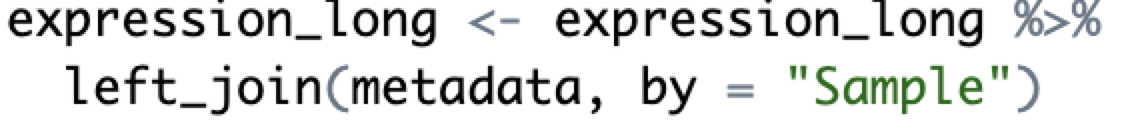
1. Check if the last row is removed

*dim(Data) # Check new dimensions*

1. Define genes of interest (EGFR, BRAF & LKB1)
   1. *genes\_of\_interest<-c("ENSG00000146648","ENSG00000118046","ENSG00000157764")*
2. Convert the data to a long format with the genes in rows and the samples in columns



1. Merge the expression data with metadata in order to include sample information



1. View the structure of the transformed data
   1. *print(head(expression\_long))*
2. Adjust the P-value to 0.05

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

For this study, the focus was on analyzing genomic expression in non-small cell lung cancer (NSCLC) tissues to understand how certain amplified genes contribute to cancer progression and how this will help in identifying potential treatment options. In the analysis distinctions were 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, we further specified the genes of interest: 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 roles in tumor biology and treatment response.

To analyze the data, we used RStudio, a powerful platform for statistical computing and visualization, along with the DESeq2 package for differential gene expression analysis. DESeq2 was specifically chosen for its ability to analyze RNA-seq count data using the Wald test as a default, which allowed us to assess whether variables such as tumor stage, smoking intensity, or gender influenced the expression of these genes. This approach also enabled us to account for confounding variables, ensuring that the statistical results were robust and reliable. Comparisons were made between categorical variables, such as cancerous versus non-cancerous tissues and gender differences, while correlations were assessed with continuous variables, such as age.

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

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

<https://pubmed.ncbi.nlm.nih.gov/29282718/>

<https://pubmed.ncbi.nlm.nih.gov/33576873/>

<https://www.youtube.com/watch?v=2RFYKTvCXHs&list=PLJefJsd1yfhbIhblS-85alaFsPdU00DaA&index=3>

|  |
| --- |
| 8. APPENDIX |