***Data acquisition and Computational environment setup***

To investigate differentially expressed genes (DEGs) in lung cancer, RNA sequencing data was obtained from the Gene Expression Omnibus (GEO) database, specifically dataset GSE81089. This dataset consists of transcriptomic profiles of tumor and adjacent normal lung tissues from patients with non-small cell lung cancer (NSCLC). The dataset was accessed and retrieved using the GEOquery package in R, which allows for the direct downloading and processing of GEO datasets.

To ensure reproducibility and streamline the computational workflow, a dedicated environment was created with all group members. This environment included dependencies such as R (version 4.2.2). Whitin R, several packages were installed to facilitate data processing and analysis, including *DESeq2* for differential expression analysis, *ggplot2* and *pheatmap* for visualization, and *clusterProfiler* for functional enrichment analysis. The entire project was managed through GitHub, ensuring systematic version control, allowing multiple contributors to work collaboratively while maintaining an organized storage of scripts and results.

…

***Data preprocessing and quality control***

Before conducting differential expression analysis, preprocessing steps were carried out to clean and structure the dataset.

Exploratory analyses were performed using count data corrected by Fragments Per Kilobase of transcript per Million reads mapped (FPKM). Normalisation methods are used to account for technical variabilities such as sequencing depth, transcript length, sample-to-sample variability and batch-to-batch variability (Conesa et. Al, 2016). FPKM corrects for variations in both gene length and sequencing depth. It is calculated by … Because it normalises reads by the total number of fragments mapped, it is indicated to compare gene expression levels between different samples. This is not the case with TPM which is useful when comparing gene expression levels within a sample. -> RPKM?.. The statistical tool DESeq2 automatically uses the median-of-ratios normalisation method thus, to avoid distortion of the normalisation process, raw counts data was provided as input.

The raw count matrix and associated metadata were loaded into R, ensuring that sample identifiers were correctly matched between the two files. Genes with very low expression levels, defined as having an average count below 10 raw reads, were removed to reduce noise and improve statistical power.

…

***Differential expression analysis***

To identify genes that were significantly differentially expressed between tumor and normal lung tissues, differential expression analysis was performed using the DESeq2 package in R. The first step in this analysis involved converting the raw count matrix into a DESeq dataset, incorporating sample metadata to specify experimental conditions. After that we performed quality control, and we excluded genes with a mean raw count of <10. We set Human non-malignant tissue as reference and ran DESeq2. To obtain a list of DEGs, the results of the differential gene expression analysis were filtered by selecting genes with an adjusted p-value < 0.01, ensuring statistical significance, and an absolute Log 2Fold change > 1, why??

…

***Visualization of Differentially Expressed Genes (DEGs)***

Several visualization techniques were employed to facilitate the interpretation of the results. Volcano plots were generated using *ggplot2* to provide an overview of significantly upregulated and downregulated genes, highlighting their statistical significance. Heatmaps, created using the pheatmap package, were used to visualize hierarchical clustering patterns of DEGs across tumor and normal samples, helping to identify potential gene expression identification. MA plots were constructed to display log2 fold-change values against mean expression levels, assisting in the identification of expression trends.

…

***Functional enrichment analysis***

To better understand the biological relevance of the differentially expressed genes, functional enrichment analysis was conducted using the *clusterProfiler* package in R. This analysis focused on identifying overrepresented biological processes, molecular functions, and cellular components through Gene Ontology (GO) analysis.

…

In addition to GO analysis, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis was performed to identify signaling pathways that were significantly enriched among the DEGs.

…

***Version control and collaboration***

Throughout the project, all scripts, analyses, and results were managed using GitHub to ensure efficient version control and collaboration among our group members and supervisor. Frequent commits were made to track the progress, and GitHub fetch and pull requests were used to document everything, review code, and discuss the results. By maintaining a well-organized repository, the workflow remained transparent and reproducible, allowing for seamless integration of contributions from all group members.

***Statistical analysis***

…