Differential Expression and Functional Analysis of COVID-19 Patients based on Viral Load, Age and Sex

Wiktoria Brandys, Maria Chmielorz,
Julia Kwiecińska, Katarzyna Kuhny

January 28, 2026

Contents

1 Introduction

The COVID-19 pandemic, caused by SARS-CoV-2, has revealed profound heterogeneity in host responses. While many individuals remain asymptomatic, others progress to severe respiratory failure and multi-organ dysfunction. This difference in clinical outcomes suggests that this disease is is not determined just by the virus itself, but by how the host's organism's immune system responds to it [1]. It has been shown that factors such as age [3] and sex [4] can influence immune responses. Furthermore, the viral load, which is the amount of virus present at the site of infection, functions as the indicator of the disease's progression. In this project, we utilize the dataset from Lieberman et al. [2] to analyze how SARS-CoV-2 viral load drives gene expression changes, while taking into account the individual differences associated with patient's age and sex.

2 Materials and Methods

2.1 Dataset and Data Acquisition

We utilized the GSE152075 dataset from the Gene Expression Omnibus (GEO). The raw data consists of RNA-seq counts from nasopharyngeal swabs of patients tested for SARS-CoV-2. The full dataset includes 483 samples:

- 430 SARS-CoV-2 positive patients,
- 54 SARS-CoV-2 negative individuals (the control group).

2.2 Data filtering

Some samples were excluded from our final analysis. We applied specific filtering criteria:

- **SARS-CoV-2 positivity**
  Only samples marked as "positive" were included in the analysis to focus on differences in hosts' response to different states of infection.
- **Completeness**
  Samples with missing values (NA) for age or viral load were excluded.
- **Gender Selection**
  Analysis was restricted to samples with gender specified as male or female.

This resulted in 377 samples remaining for the analysis.

2.3 Data Categorization

We categorized samples into various groups to simplify the comparison. We maintained similar grouping rules as the original study to ensure that our analysis remains comparable.

- **Age Groups (AGE_GR)**
  Patients were divided into two groups based on age. One group consists of individuals younger than 60 years old ($n = 222$), while the other includes those over 60 ($n = 155$).

- **Viral Load (STATE_GR)**
  Using the thresholds defined in the publication, we used the N1 Cycle Threshold (Ct) values to group patients into three levels of viral intensity:
  - **High:** $Ct < 19$ ($n = 94$)
  - **Medium:** $19 \leq Ct \leq 24$ ($n = 190$)
  - **Low:** $Ct > 24$ ($n = 93$)

- **Gender (GENDER)**
  Female ($n = 201$) or Male ($n = 176$).

2.4 Statistical Framework (DESeq2)

Differential Gene Expression (DGE) analysis was performed using the DESeq2 package in R. Our central research question asks: *How does gene expression change as SARS-CoV-2 infection progresses, considering the age and sex of the patients?*

- **The Additive Model**

To isolate the effect of the viral load, we utilized a multi-factorial design formula:
Design = $\sim$ GENDER + AGE_GR + STATE_GR
In this model, gender and age are treated as individual traits that influence baseline expression, while STATE_GR is the primary factor of interest. By using this design, DESeq2 tests whether a gene has a different mean expression in "High" vs "Low" viral load groups while comparing individuals of the same sex and in the same age group.

2.5 Quality Control

The first step in our quality control was to examine how the gene expression varies across samples. We used the plotDispEsts(dds) function to visualize the relationship between the average expression of a gene and its dispersion.

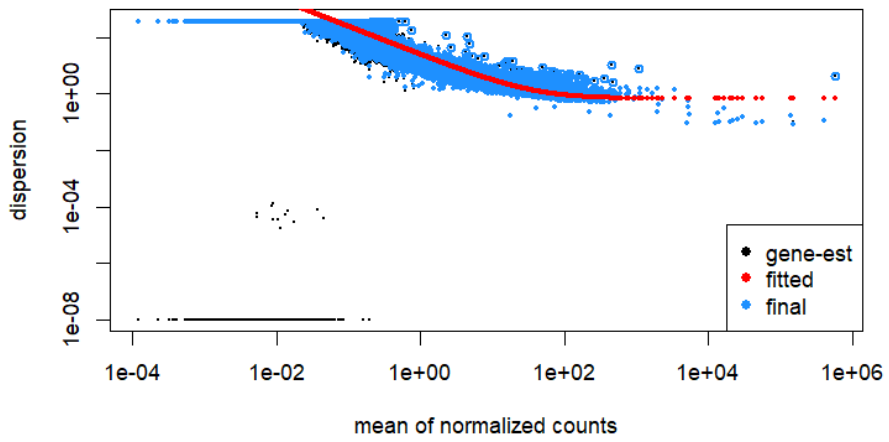

Figure 1: **Dispersion plot for the differential expression analysis using the design = ~ GENDER + AGE_GR + STATE_GR**. The successful shrinkage of individual gene estimates (black dots) toward the fitted trend line (red line) indicates that the DESeq2 model effectively stabilized the data variance. This ensures that the differential expression results will not be biased by high variance in genes with low count numbers.

3 Results

3.1 Principal Component Analysis (PCA)

We performed Principal Component Analysis (PCA) to visualize the global transcriptomic relationships between samples and assess sample clustering across groups.

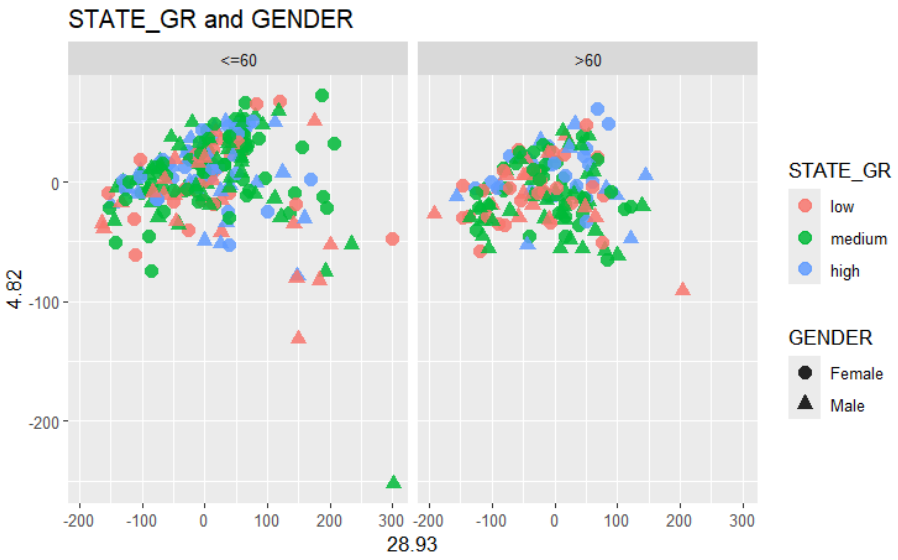

Figure 2: **PCA of Gene Expression Patterns faceted by age**. The data is projected onto the first two principal components, with PC1 (x-axis) explaining 28.93% of the total variance and PC2 (y-axis) explaining 4.82%. Samples are colored by viral intensity (red: low, green: medium, blue: high). Gender is represented by shapes (circle: Female, triangle: Male). The plot is faceted into two panels: individuals ≤ 60 (left) and > 60 (right). The results show no visible grouping.

3.2 Differential Gene Expression (DGE) Analysis

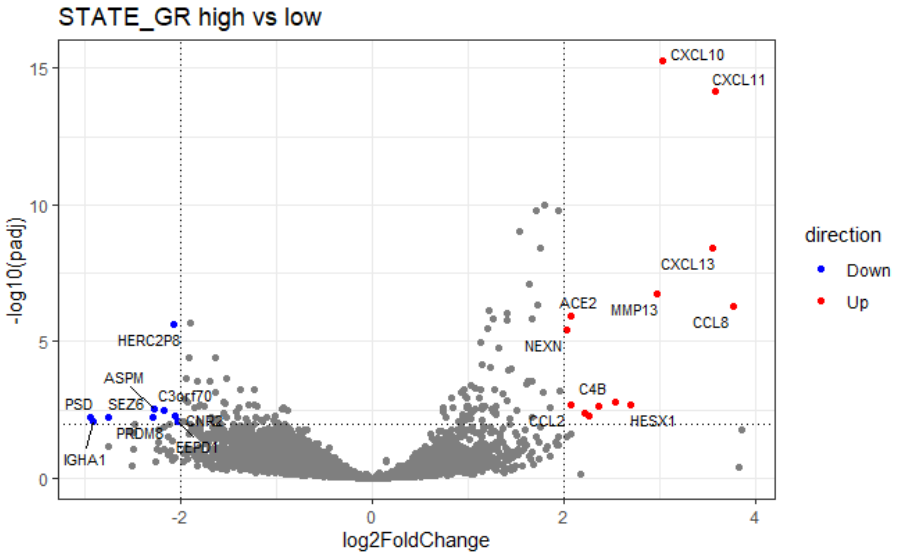

Figure 3: **Volcano Plot of Differently Expressed Genes**. The figure shows a volcano plot summarizing differential expression analysis for the comparison "STATE_GR high vs low" (viral load based on N1 Ct values). The x-axis represents $\log_2$ fold change, and the y-axis represents $-\log_{10}$ adjusted p-value (padj). Vertical dashed lines mark fold change thresholds at $\log_2 \mathrm{FC} = -2$ and $+2$, while a horizontal dashed line denotes the padj significance cutoff at $-\log_{10}(0.01) \approx 2$. Points are colored by direction: red for genes meeting criteria "Up" (padj < 0.01 and $\log_2 \mathrm{FC} \geq 2$), blue for "Down" (padj < 0.01 and $\log_2 \mathrm{FC} \leq -2$), and gray for non-significant genes.

15 significantly up-regulated genes (padj < 0.01, $\log_2 \mathrm{FC} \geq 2$):
CXCL10, CXCL11, CXCL13, CCL8, MMP13, ACE2, C4B, NEXN, HESX1, CCL2, CXCL8, CXCL9, IFIT3, ISG15, MX1.

9 significantly down-regulated genes (padj < 0.01, $\log_2 \mathrm{FC} \leq -2$):
HERC2P6, ASPM, SEZ6, CNTN2, IGHA1, POU6F2, NEK2, KIF4A, TOP2A.

3.3 Gene Annotation

Significant genes were annotated using the biomaRt interface to the Ensembl database. Retrieved annotations included:

- Ensembl gene identifiers,
- HGNC gene symbols,

- chromosomal locations,
- UniProt identifiers,
- functional gene descriptions.

HGNC	**Ensembl ID**	**Chr**	**UniProt**	**Description**
IGHA1	ENSG00000282633	HSCHR14_3_CTG1	P01876	immunoglobulin heavy constant alpha 1 [HGNC:5478]
IGHA1	ENSG00000211895	14	P01876	immunoglobulin heavy constant alpha 1 [HGNC:5478]
C3orf70	ENSG00000187068	3	A6NLC5	chr 3 open reading frame 70 [HGNC:33731]
CNR2	ENSG00000188822	1	P34972	cannabinoid receptor 2 [HGNC:2160]
PSD	ENSG00000059915	10	A5PKW4	pleckstrin and Sec7 domain [HGNC:9507]
SEZ6	ENSG00000063015	17	Q53EL9	seizure related 6 homolog [HGNC:15955]
ASPM	ENSG00000066279	1	Q8IZT6	spindle microtubules assembly factor [HGNC:19048]
PRDM8	ENSG00000152784	4	Q9NQV8	PR/SET domain 8 [HGNC:13993]
EEPD1	ENSG00000122547	7	Q7L9B9	endo/exo/phosphatase family domain 1 [HGNC:22223]

Table 1: Downregulated Genes Annotation (STATEGR high vs low)

HGNC	Ensembl ID	Chr	UniProt	Description
CXCL10	ENSG00000169245	4	P02778	C-X-C motif chemokine ligand 10 [HGNC:10637]
CXCL11	ENSG00000169248	4	O14625	C-X-C motif chemokine ligand 11 [HGNC:10638]
CXCL13	ENSG00000156234	4	O43927	C-X-C motif chemokine ligand 13 [HGNC:10639]
C4B	ENSG00000236625	HSCHR6_MHC_DBB_CTG1	P0C0L5	complement C4B (Chido/Rodgers) [HGNC:1324]
C4B	ENSG00000228454	HSCHR6_MHC_SSTO_CTG1	P0C0L5	complement C4B (Chido/Rodgers) [HGNC:1324]
C4B	ENSG00000228267	HSCHR6_MHC_COX_CTG1	P0C0L5	complement C4B (Chido/Rodgers) [HGNC:1324]
MMP13	ENSG00000137745	11	P45452	matrix metallopeptidase 13 [HGNC:7159]
CCL8	ENSG00000108700	17	P80075	C-C motif chemokine ligand 8 [HGNC:10635]
CCL2	ENSG00000108691	17	P13500	C-C motif chemokine ligand 2 [HGNC:10618]
HESX1	ENSG00000163666	3	Q9UBX0	HESX homeobox 1 [HGNC:4877]
ACE2	ENSG00000130234	X	Q9BYF1	angiotensin converting enzyme 2 [HGNC:13557]
SLC38A5	ENSG00000017483	X	Q8WUX1	solute carrier family 38 member 5 [HGNC:18070]
C4B	ENSG00000224389	6	P0C0L5	complement C4B (Chido/Rodgers) [HGNC:1324]
PARP16	ENSG00000138617	15	Q8N5Y8	poly(ADP-ribose) polymerase 16 [HGNC:26040]
NEXN	ENSG00000162614	1	Q0ZGT2	nexilin F-actin binding protein [HGNC:29557]

Table 2: Upregulated Genes Annotation (STATEGR high vs low)

3.4 Age–Severity Interaction Analysis

To investigate whether the effect of clinical severity differs between age groups, an extended model including an interaction term was fitted:

$$\sim \text{GENDER} + \text{AGE\_GR} + \text{STATE\_GR}$$

A Likelihood Ratio Test (LRT) was applied to compare the full model against a reduced model lacking the interaction term. This analysis tests whether age modifies the transcriptional response to disease severity, revealing genes with age-dependent severity effects.
Genes were ranked according to their adjusted p-values from the LRT. The top-ranked genes (lowest padj values) were selected for visualization to facilitate interpretation of potential age-specific severity effects.
For the top candidate genes, normalized expression counts were extracted using plotCounts and visualized as gene-wise interaction plots.

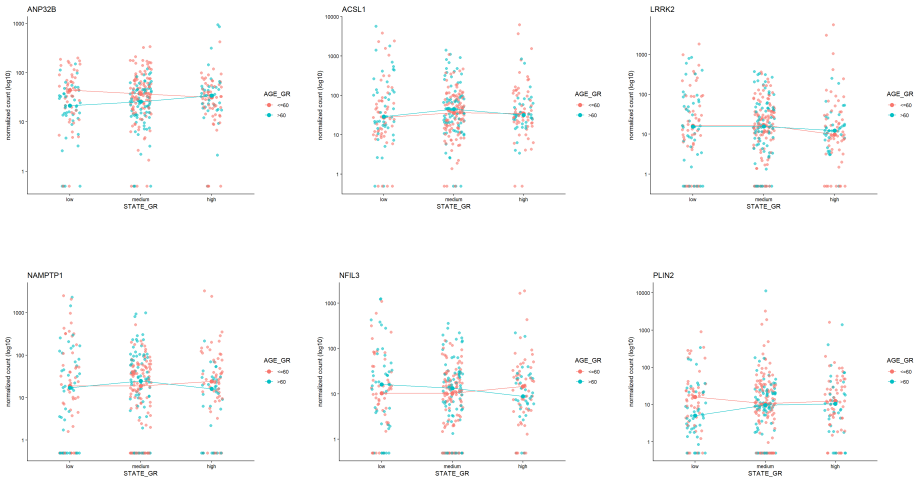

Figure 4: Normalized expression counts (log10 scale) for the top 6 genes (ANP32B, LRRK2, NAMPTP1, NFIL3, PLIN2, ASCL1) with age-dependent severity effects (selected by lowest padj from LRT analysis), presented in 6 panels. Each panel shows jittered individual samples by disease state (STATE_GR: x-axis) and age group (AGE_GR: color; pink: $\leq$60, blue: >60), with median lines linking age groups within states and highlighted median points.

3.5 Gender–Severity Interaction Analysis

In addition to the age–severity interaction, a second interaction analysis was performed to assess whether biological sex modifies the transcriptional response to disease severity. Specifically, this analysis tested whether the effect of clinical severity (STATE_GR) on gene expression differs between male and female

patients.

To investigate whether the effect of clinical severity differs between biological sexes, an extended model including a sex–severity interaction term was fitted:

$$\sim \text{GENDER} + \text{AGE\_GR} + \text{STATE\_GR} + \text{GENDER:STATE\_GR}$$

A Likelihood Ratio Test (LRT) was applied to compare the full interaction model against a reduced model without the interaction term. This analysis tests whether biological sex modifies the transcriptional response to disease severity, thereby identifying genes whose severity-associated expression patterns differ between male and female patients.

Genes were ranked according to their adjusted p-values from the LRT. The top-ranked genes (lowest padj values) were selected for visualization to facilitate interpretation of potential sex-specific severity effects.
For the top candidate genes, normalized expression counts were extracted using plotCounts and visualized as gene-wise interaction plots.

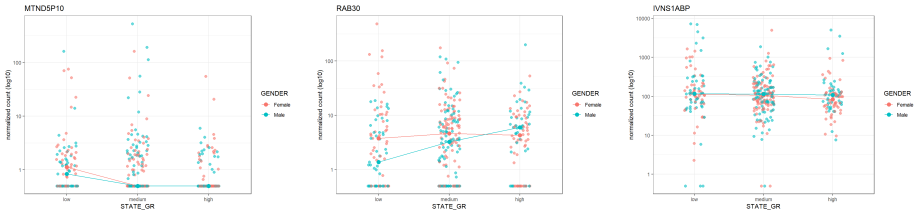

Figure 5: Normalized expression counts (log10 scale) for the top 3 genes (MTND5P10, RAB30, IVNS1ABP) with sex-dependent severity effects (selected by lowest padj from LRT analysis), presented in 3 panels. Each panel shows jittered individual samples by disease state (STATE_GR: x-axis) and biological sex (GENDER: color; pink: Female, blue: Male), with median lines linking sexes within states and highlighted median points.

3.6 Functional Enrichment Analysis (GO)

We conducted a Functional Enrichment Analysis (GO) to identify the biological processes in which the differentially expressed genes are involved 6.

1. **Viral Intensity (STATE_GR - upregulated)**: The enrichment analysis for viral load reveals activation of the **chemokine-mediated signaling pathway** and **leukocyte chemotaxis**. The term **humoral immune response** achieved the highest Gene Ratio (0.50), indicating that half of the significantly upregulated genes are involved in antibody-mediated immunity.

2. **Age Groups (AGE_GR - upregulated and downregulated)**: The impact of age shows two distinct biological characteristics:

    - Upregulation: There are differences in expression of genes involved in processes related to the **adaptive immune response** and **humoral immunity**. The most enriched terms include **lymphocyte mediated immunity**, **leukocyte mediated immunity**, and **B cell mediated immunity**, highlighting the role of B lymphocytes in age-related expression shifts. Significant enrichment was also observed for the **B cell receptor signaling pathway** and the **adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains**. The presence of terms such as **immunoglobulin mediated immune response** and **humoral immune response** confirms the changes of engagement of antibody production mechanisms. The highly significant adjusted p-values (ranging from 10^{-7} to 10^{-5}) underscore that these age-related differences are intrinsically linked to adaptive immunity.
    - Downregulation: Significant differences were also observed in the cell's protein-building machinery, specifically in processes like **cytoplasmic translation** and **ribonucleoprotein complex biogenesis**. With over 200 genes affected by these changes, the results indicate a shift in the capacity to synthesize proteins necessary for repair and defense. This "translational shutdown" may provide a molecular basis for the varying levels of disease severity and vulnerability observed across different ages.

3. **Gender (GENDER)**: Functional Enrichment for gender-specific differences resulted in very few statistically significant results. Only one gene is associated with **Rab protein signal transduction** for the p-value cutoff set on 0.01. The results suggest that gender is not a primary determinant of transcriptomic response in this cohort.

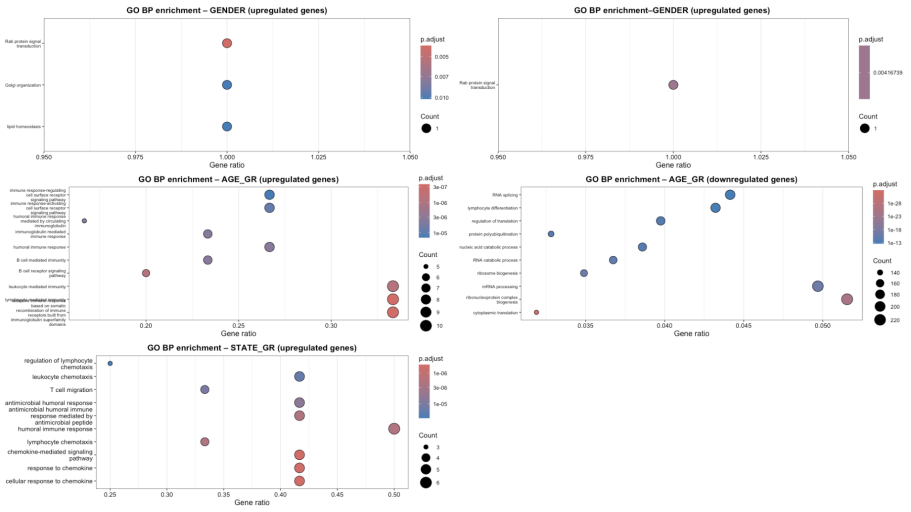

Figure 6: **Functional Gene Ontology (GO) Enrichment Analysis of Differentially Expressed Genes.** Dot plots represent significantly enriched Biological Processes (BP) for three variables: Viral Load (STATE_GR), Age (AGE_GR), and Gender (GENDER). The x-axis represents the Gene Ratio (the proportion of genes from the input list assigned to a specific GO term). Dot size indicates the total gene count for each term and the color gradient reflects the adjusted p-value. The first row shows enriched terms for upregulated genes in the Gender-Severity Model, comparing an adjusted p-value cutoff of set on 0.05 (left panel) and on 0.01 (right panel); no significant results were identified for downregulated genes in this group. For all other analyses in the second and third rows, an adjusted p-value cutoff 0.01 was applied. The second row displays results for the Age-Severity Model, with upregulated genes (left) and downregulated genes (right). The third row presents enriched terms for upregulated genes in the Additive Model; no significant results were identified for downregulated genes in this group.

4 Discussion

Our analysis identified several key gene sets that match the findings of Lieberman et al.[2]. Regarding Viral Load (STATE_GR), both the study and our analysis conclude that host responses are primarily driven by viral intensity. Similarly to the publication, our list of upregulated genes for high viral load includes ISG15, MX1, and IFIT3, which are markers of an interferon-mediated antiviral response, and ACE2 gene, which, as the publication finds, increases as a function of viral load. We also found, similarly to the publication, pro-inflammatory chemokines such as CXCL9, CXCL10, and CCL8, which are responsible for recruiting immune cells to the infection site. Additionally, both studies observed a downregulation of cell-cycle regulators like TOP2A and ASPM, suggesting a

viral-induced pause in normal cell growth.

However, our approach differs by focusing exclusively on the SARS-CoV-2 positive cohort and utilizing a sex- and age-specific interaction models. While the original publication noted a general reduction in ribosomal proteins, our results demonstrate that this "translational shutdown" (the suppression of over 200 genes related to protein synthesis) differs based on age group. Regarding biological sex, Lieberman et al. utilized cell-type deconvolution (CIBERSORTx) to identify a reduction in B-cell and NK-cell transcripts in males. This approach differs from ours, which might explain why in our analysis these differences did not reach statistical significance, with both our PCA [2] and GO plots [6] showing no distinct clustering by gender.

5 Resources and Software

5.1 Bioinformatics and Statistical Tools

Analysis was conducted using the R programming language (4.5.2) and code from Assignments 7 and 8 from the course. We used the following packages:

1. **GEOquery**: used to download the dataset from the Gene Expression Omnibus database,
2. **biomaRt**: annotation of genes,
3. **dplyr, tibble**: filtering, sorting and cleaning up the table,
4. **stringr**: fixing messy text within data,
5. **DESeq2**: Differential Expression Analysis,
6. **ggplo2, enrichplot, ggrepel, cowplot**: used for plots,
7. **clusterProfiler**: functional enrichment analysis,
8. **GO.db**: Gene Ontology term database,
9. **org.Hs.eg.db**: annotation database for the human genome,
10. **enrichplot**: visualization of enrichment results.

A list of all R packages used in the project, including versions and citations, is available in the project repository: `https://github.com/WBrandys/HTS_project/packages.txt`

5.2 Artificial Intelligence

We used generative AI in specific parts of the project:

1. **Perplexity AI**: helped with the search of suitable datasets from GEO to perform DEA and functional analysis on.

2. **ChatGPT (OpenAI)**: aided with plot styling and encoureged the usage of ggrepel package. Helped with code commenting.

3. **Gemini (Google)**: used for grammar correction and improving stylistic clarity in the Introduction and Discussion section.

6 Division of Responsibilities

- Wiktoria Brandys - developed R code for RNA-seq analysis using DESeq2, including differential expression analysis, PCA-based exploratory analysis, interaction modeling for age- and sex-dependent effects and visualization of gene expression patterns.
- Maria Chmielorz - conducted the functional enrichment analysis of Gene Ontology, prepared and formatted the HTML report, presented visual presentation
- Julia Kwiecińska - prepared a visual presentation and presented it in class; described the Differential Expression Analysis (DEA), DGE analysis, gene annotation and interaction analysis in the report; wrote the README on GitHub; added detailed comments throughout the code to explain each step.
- Katarzyna Kuhny - prepared the majority of the final LaTeX report, excluding specific sections drafted by other team members that are mentioned above.

References

[1] Daniel Blanco-Melo, Benjamin E Nilsson-Payant, Wen-Chun Liu, Skyler Uhl, Daisy Hoagland, Rasmus Møller, Tristan X Jordan, Kohei Oishi, Maryline Panis, David Sachs, et al. Imbalanced host response to sars-cov-2 drives development of covid-19. *Cell*, 181(5):1036–1045, 2020.

[2] Nicole AP Lieberman, Vikas Peddu, Hong Xie, Lasata Shrestha, Meei-Li Huang, Megan C Mears, Maria N Cajimat, Dennis A Bente, Pei-Yong Shi, Francesca Bovier, et al. In vivo antiviral host transcriptional response to sars-cov-2 by viral load, sex, and age. *PLoS biology*, 18(9):e3000849, 2020.

[3] Amber L Mueller, Maeve S McNamara, and David A Sinclair. Why does covid-19 disproportionately affect older people? *Aging (albany NY)*, 12(10):9959, 2020.

[4] Takehiro Takahashi, Mallory K Ellingson, Patrick Wong, Benjamin Israelow, Carolina Lucas, Jon Klein, Julio Silva, Tianyang Mao, Ji Eun Oh, Maria Tokuyama, et al. Sex differences in immune responses that underlie covid-19 disease outcomes. *Nature*, 588(7837):315–320, 2020.