# Final Assignment

Vui Doan

## 1 Introducción

The dataset includes RNA-seq blood samples from 128 donors, including 100 COVID-19-positive and 26 COVID-19-negative patients. The gene expression matrix derived from this dataset was shortened and includes data for 100 genes. The dataset includes information regarding 25 covariates associated with each sample and donor, including disease status, hospital-free days over a continuous 45-day period, and levels of ferritin, among others. To select the gene shown in the main plots, correlations between the expression levels of each gene and the selected continuous covariate: hospital-free days over a continuous 45-day period (HFD-45) were computed. The gene showing the top positive correlation with HFD-45 was selected as our main gene. Since lower values of HFD-45 are linked to worse prognosis and a higher number of hospitalization days, negative correlations between a specific gene and HFD-45 allow us to identify genes that are associated with better prognosis. Using this method, we identified and selected $ABCD4$ as our main gene of interest.$ABCD4$ ATP-binding cassette (ABC) transporter family. These proteins move molecules across membranes using ATP. In the blood the cell types presenting the highest expression of this gene are Natural Killer cells wich detect cells that are infected by viruses or have become cancerous and elease substances that make pores in the membranes of the infected celss and release enzymes that enter through the pores and trigger apoptosis of those cells [2].

## 2 Mathods

Gene expression derived from the blood samples of the128 donors was obtained from [1]. Correlation analyses were carried out between the selected continuous covariates HFD-45, Ferritin (ng/mL), Fibrinogen (mg/dL) and all genes in the dataset and those showing the top positive and negative correlations were selected for each cavariate. For categorical covariates (Sex, Mechanical Ventilation, Disease Status) t-tests were carried out for each gene and the top positive and negatve associated genes were selected for each covariate based on the values of the t-test statistics. Plots were generated using the ggplot 2 library [3] and the heatmap was produced with the help of the pheatmap library [4]. Sample nemes were removed from the heatmap for clarity. Clustering for rows and columns is carried out by default using hierarchical clustering with complete linkage and euclidean distance. Principal component analysis of the gene expression data was carried out using the prcomp function. All analyses were conducted using R version 4.5.1 [5].

## 3 Results

Table 1 shows the summary statistics for the three categoriacal and the three continous covariates selected. The dataset includes 51 females and 74 males . $ABCD4$ expression values ranged from 2.47 to 20. Figure 1 shows the histogram of the expression values of $ABCD4$ in our dataset. Figure 2 shows the expression values of $ABCD4$ compared to HFD-45. Higher expression levels of

*ABCD4* seem to be associated with heger values of HFD-45 suggesiting shorter hospital stays in idividuals presenting higher expression levels of this gene in their blood samples. Females seemed to present higher levels of *ABCD4* compared to males and no clear differences were observed in gene expression were observed based on disease status. Figure 3 depicts the expression levels of *ABCD4* in COVID-19-positive and COVID-19-negative patients stratifiying by Sex. To produce the heatmap we selected the top associations in our analyses. (The most positively and negatively correlated genes with our continous covariates and the genes showing the positive and negative effects in our t-tests for categorical variables). FIgure 4 shows a heatmap for the expression values of the selected genes. Higher expression of *ABCA13*, *ABHD5*, and *ABCB6* seem to be associated with COVID-19-posive disease status. Finally principal component analysis (PCA) was carried out using all genes in the dataset. Using the full expression data PCA creates synthetic variables that capture the top variability of the data. It is a technich that allows to reduce the dimensionality of our data and allows us to visualize the dataset in two dimensions. In addition we can plot information regarding specific covariates by using colors and shapes. Based on the two first principal components COVID-19 -possitive samples seem to group in the bottom left of the plot whereas the COVID-19-negatve ones do it in the upper-right sector. No clear grouping based on Sex is observed in this plot.

| | level | Female | Male |
|---:|---|---|---|
| n | | 51 | 74 |
| HFD-45 (mean (SD)) | | 26.37 (16.34) | 22.61 (17.02) |
| Ferritin (ng/mL) (mean (SD)) | | 619.28 (1054.33) | 993.35 (1013.05) |
| Fibrinogen (mg/dL) (mean (SD)) | | 469.15 (163.26) | 550.72 (219.68) |
| Sex (%) | Female | 51 (100.0) | 0 (0.0) |
| | Male | 0 (0.0) | 74 (100.0) |
| Mechanical Ventilation. (%) | no | 35 (68.6) | 39 (52.7) |
| | yes | 16 (31.4) | 35 (47.3) |
| Disease Status (%) | COVID-19 | 38 (74.5) | 62 (83.8) |
| | non-COVID-19 | 13 (25.5) | 12 (16.2) |

Table 1: Summary statistics for all the covariates

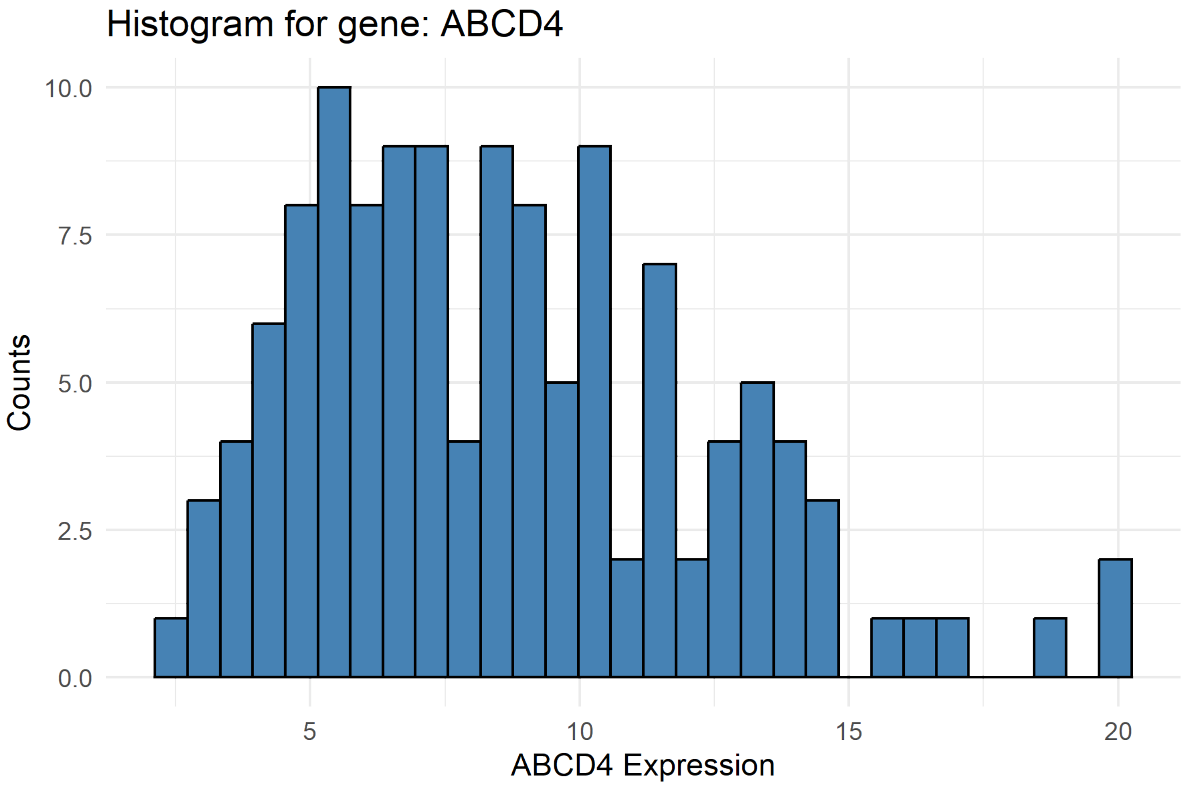

Figure 1: Histogram of *ABCD4* expression.

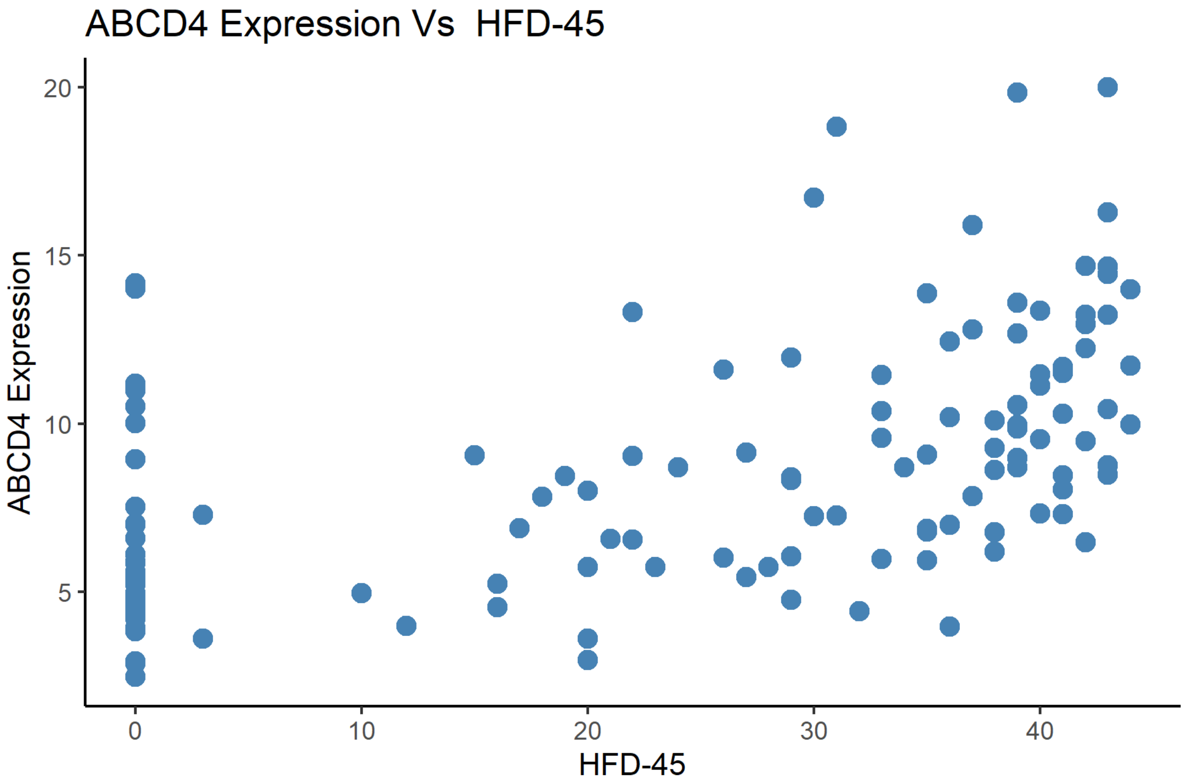

Figure 2: Scatter plot displaying the expression values of *ABCD4* and the HFD-45 scores.

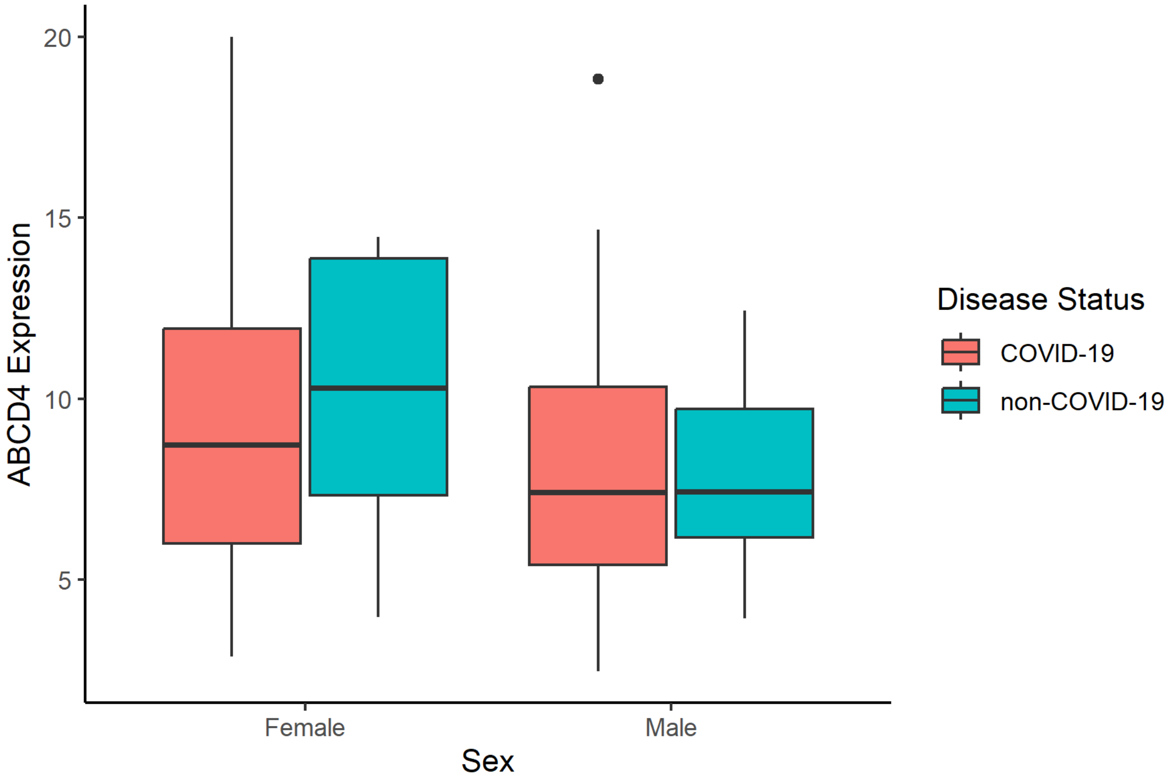

Figure 3: Box plot showing the expression values of *ABCD4* for COVID-19-positive and COVID-19-negative patients stratifiying by Sex.

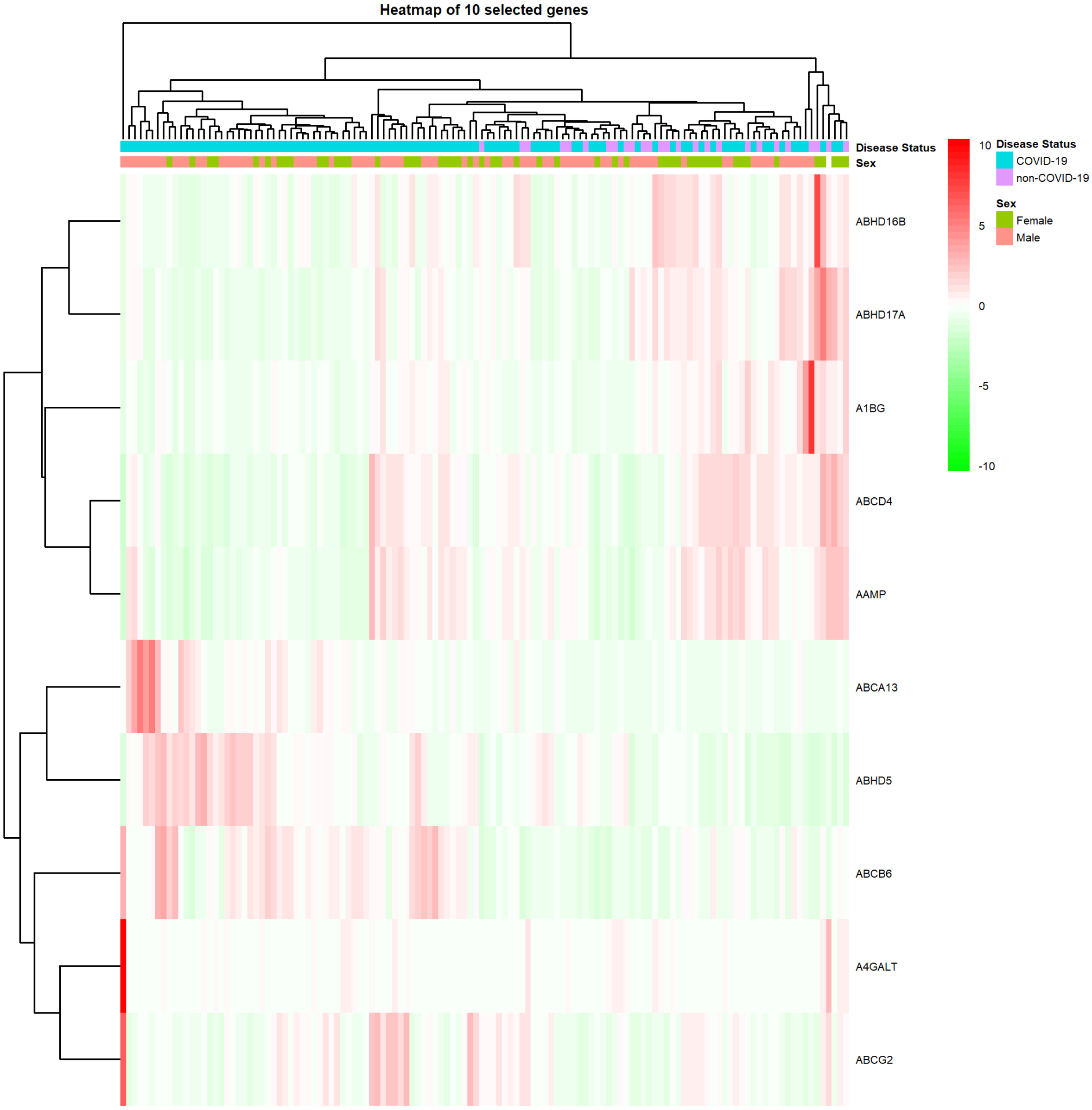

Figure 4: Heatmap displaying the expression values of the 10 selected genes and the annotations for Disease Status and Sex.

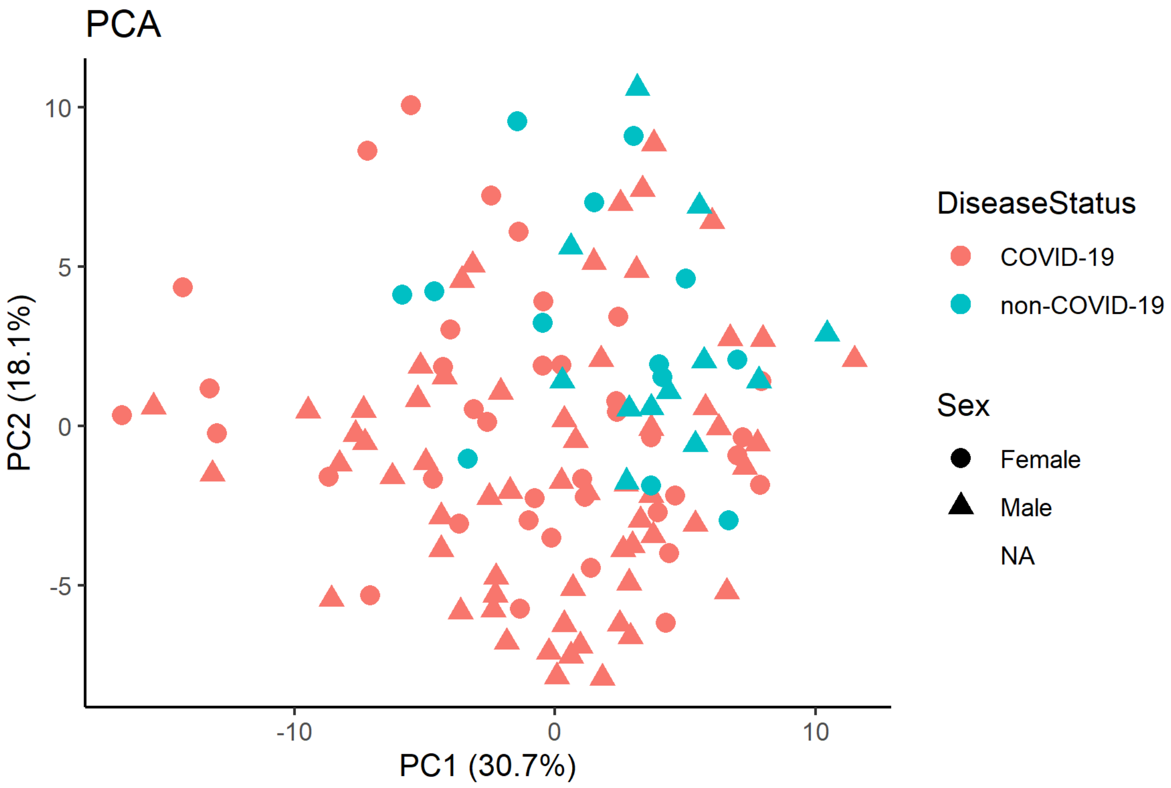

Figure 5: Principal component analsysis of the full gene expression data. COVID-19-positive patients are colored in red whereas COVID-19-negative patients are colored in blue. Circles represent females in the dataset whereas triangles are reserved for males.

## References

[1] Overmyer, K. A., *et al.* (2021). Large-scale multi-omic analysis of COVID-19 severity. *Cell Systems, 12*(1), 23–40.e7. doi:10.1016/j.cels.2020.10.003

[2] Coelho, D., *et al.* (2012). Mutations in ABCD4 cause a new inborn error of vitamin B12 metabolism. *Nature Genetics, 44*(10), 1152–1155. doi:10.1038/ng.2386

[3] Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* (2nd ed.). Springer. doi:10.1007/978-3-319-24277-4

[4] Kolde, R., *et al.* (2019). *pheatmap: Pretty Heatmaps.* R package version 1.0.12. Available at: `https://CRAN.R-project.org/package=pheatmap`

[5] R Core Team, *et al.* (2024). *R: A Language and Environment for Statistical Computing* (Version 4.5.1) [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. Available at: `https://www.R-project.org/`