PROJECT 2 – SARS MICE

Guillaume CHYZAK

Alexandre MARTINEZ

Hosnia SHALABI

Minh-Anh HUYNH

Sorbonne Université

UE 5BM578 “Workshop in immunological data analysis”

Table of Contents

Introduction

In 2002, an outbreak caused by the SARS-CoV (Severe Acute Respiratory Syndrome Coronavirus) occurred. This virus infected over 8,000 people and among them 800 died across the globe.(Institut Pasteur, 2021)

Studies have shown that certain species of bats were the natural host of coronaviruses and due to some genetic variations, strains like the SARS-CoV get developed and could infect humans too.

It is important to understand what the similarities and differences in the host responses are, in order to develop medical strategies.

To highlight the consequences of coronavirus infection, many techniques can be used, such as comparative analysis of the proteome. However, this information is not enough to explain the mechanisms activated for the different strains. Therefore, a complementary study on transcriptomic can be interesting.

Transcriptomes give information about the level of expression of genes in various types of cells and tissues, through a quantification of transcripts.

To better understand the differences between three different of SARS virus (SARS-WT, and two mutants SARS-BatSRBD and SARS-icSARS), we focused on the transcriptomic responses in a mouse model of infection with these SARS viruses. Four conditions were tested: infection with either one of the three strains or a mock infection. SARS-WT is the wild-type strain SARS-WT. SARS-BatSRBD is a reconstruction from a Bat-SCoV consensus genome with the murine SARS-CoV receptor-binding domain. The strains SARS-icSARS is another modified specific infectious clone of SARS virus.(Aevermann et al., 2014)

In order to determine the differences in the host responses to these SARS strains, we represented with various MDS to observe rough clusters, detected the differentially expressed genes between the conditions, applied enrichment analyses to the resulting gene lists and represented gene and sample clusters on heatmaps.

Here, we report the results of the comparison of the host responses to SARS-WT, SARS-BatSRBD, icSARS, and mock at 4 time points (D1, D2, D4, and D7). We found that only the SARS-WT strain was infectious. Then we determined how the response to this strain evolved over the different timepoints and predict two major phenotypes: early response (D1 and D2) and late response (D4 and D7).

Materials and methods

3 strains of viruses have been injected to mice (SARS-icSARS, SARS-WT, SARS-BatSRBD) and a control injection, mock, that is used to see if the injection has any effect on the mice.

Tissue samples and data collection :

The samples that are used for this project come from lung mice tissue. They were harvested and briefly rinsed in cold (4ºC) PBS. Following the RNALater (Ambion) protocol, tissue was cut into small chunks (<0.5 cm in any single dimension) and placed immediately into 10-20 volumes (w/v) (e.g., 100mg/ml) RNALater. After a 4ºC incubation overnight, samples were stored at -80ºC until processing. Lung tissue was removed from RNALater, washed in a small volume of Trizol, homogenized in 10-20 volumes (w/v) Trizol and stored at -80°C until RNA isolation.

RNA extraction protocol: For the RNA extraction, all Trizol lysates were processed simultaneously: they were phase-separated, and RNA was isolated from the aqueous phase (diluted 2-fold with RLT buffer) using Qiagen RNeasy Mini columns and the manufacturer's recommended protocol (Qiagen Inc., Valencia, CA). RNA quality was assessed on an Agilent 2100 Bioanalyzer using the nanochip format, and only intact RNA was used for microarray analyses.

Microarray profiling: Transcriptomics profiling was performed using Agilent-014868 Whole Mouse Genome Microarray 4x44K G4122F (Probe Name version).

Data analysis:

Comparison of all strains:
The aim of this part was to spot the differences between the 4 conditions and to see the up- and down-regulated genes and their function.

We used the MDS visualization which is a method highlighting similarities and dissimilarities between our 4 conditions (SARS-icSARS, SARS-WT, SARS-BatSRBD, mock) on a two-dimensional representation.

In order to determine down- and up-regulated genes, we compared the different conditions with a Wilcoxon test. We selected for further analyses the genes with a significant corrected p-value (FDR < 0.01) and with a fold-change higher than 20%.

The gene lists obtained through differential gene expression analysis were then analyzed with the EnrichR package to identify their functional profile. The biological functions for which genes from our differential analysis were statistically over-represented compared to the Gene Ontology databases (GO_Biological_Process_2021, GO_Molecular_Function_2021, GO_Cellular_Component_2021) were selected and plotted.

Comparison of the response to SARS-WT strain at various time points:

As we had observed that only SARS-WT strain induced a noticeable change in gene expression, we only analyzed this strain for this part of the analysis. We compared the host's response to infection with this strain at different time points.

As for the first part, we used an MDS to see that the samples were clustering according to days, and identified the genes responsible for the host response through a combination of differential gene analysis and functional enrichment.

With a setmap, we identified that the timepoints clustered into two major phenotypes: "early" (D1 and D2) and "late" (D4 and D7). A setmap is some sort of heatmap used to compare different sets. Each row represents an item, each column represents a set, each cell is colored red or green depending on if it's present in the corresponding set or not. A dendrogram can be added to represent the similarity among different sets.

To assess the expression of genes throughout the timepoints, we represented them on a ratio heatmap. On this heatmap representation, we represent with the colors the ratio of the gene expression in the sample compared to the mean expression in mean samples. It is coupled to a hierarchical clustering on both rows and columns. This allows to visualize and to cluster genes and samples that behave similarly.

Comparison of all strains

The global analysis of the data allows to discriminate the SARS-WT strain from the mutant strains (Bat-SRBD, icSARS) as well as discovering samples that seemed to be outliers. The use of MDS on the global transcriptomic data creates a proximity map of the samples according to their gene expression. The MDS separates the mouse samples by the virus strains, with which they were infected **(Figure 1)**. The SARS-BatSRBD and SARS-icSARS infected mice have a very similar gene expression profile compared to the mock mice, whereas the gene expression profile of the SARS-WT infected mice is really distinct from the mock mice and divided in 3 groups, all equally distant from each other.

However, 3 samples seem to be outliers: 2 SARS-WT infected samples have a gene expression profile close to SARS-BatSRBD, SARS-icSARS and mock samples, 1 mock sample has a gene expression profile close of the upper-right SARS-WT infected group **(Figure 1**). These samples will be excluded from the rest of the analyses. An analysis of differentially expressed genes of infected mice according to mock mice shows that no gene expression differs between mock mice and mice infected by SARS-BatSRBD strain and that only 2 genes (CXCL9 and CXCL10) were differentially expressed in SARS-icSARS infected mice.

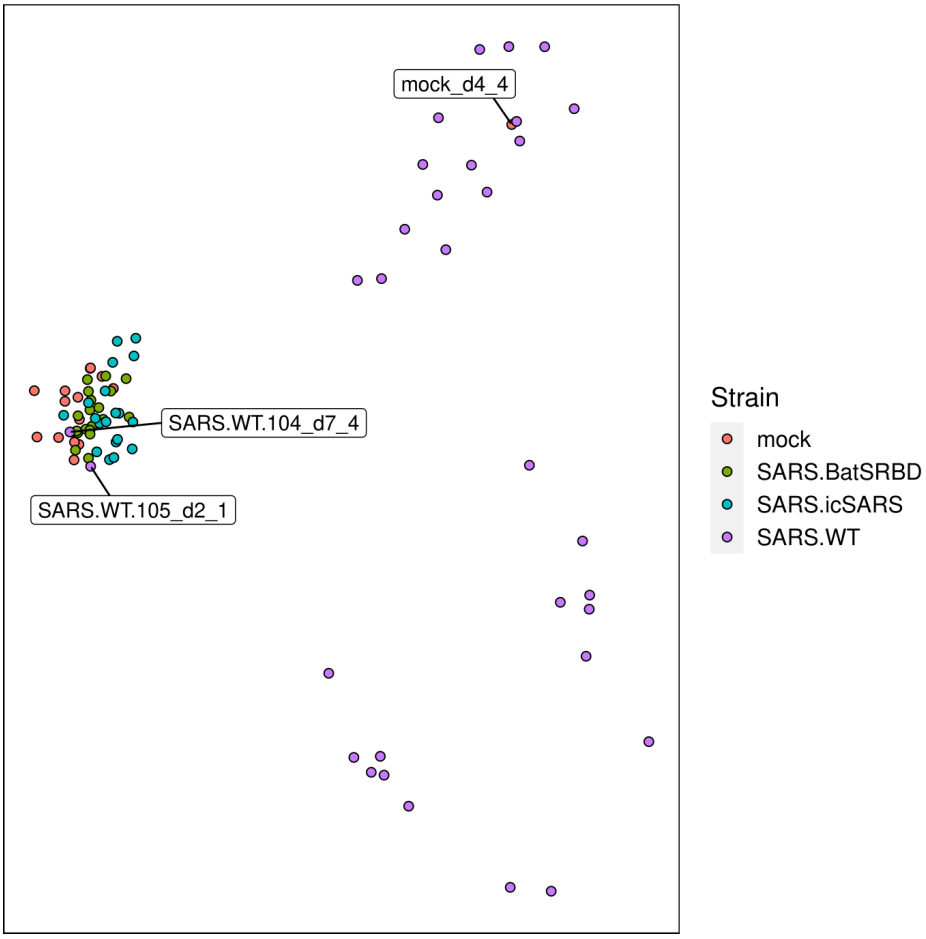


Figure 1:

MDS representation of every strain, separating strains into two different groups, SARS-WT and mock, SARS-BatSRBD, SARS-icSARS. Outliers have been identified and marked accordingly. Notably, SARS.WT is significatively positioned far away from the cluster containing SARS-BatSRBD, SARS-icSARS and mock, which seem to have a similar phenotype.

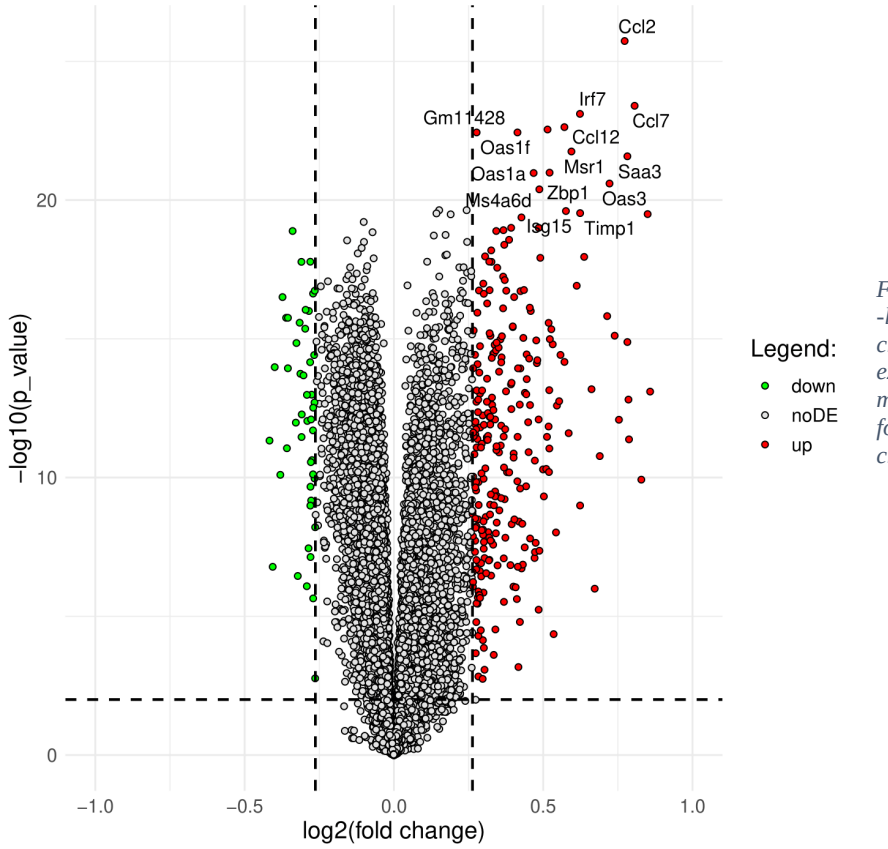


Figure 2: Volcano plot showing the -log10(p-value) and the log2(fold change) when comparing gene expression between SARS-WT and mock samples. Thresholds are 0.01 for p-values and 20% for fold changes.

Figure 3: Representation of the enriched terms from the genes that are response elements to SARS-WT infection. The names of terms that are enriched in up (down) regulated genes are colored in red (green).

On the other hand, SARS-WT infected mice showed 317 differentially expressed genes according to mock mice **(Figure 2)**. That gene list is annotated by the Gene Ontology database (Biological process 2021, Molecular functions 2021 and Cellular Content 2021). We can see in the functional analysis **(Figure 3)** that most of the significative pathways found by enrichment analysis are up-regulated ones (red names) mostly from immune system activation in response to a viral infection and chemotaxis of NK cells. The only pathway showed that seemed to get down-regulated, "Dendritic spine membrane", is from a Cellular Component database and that pathway is constituted of only 8 genes. It happens that we have 2 of those genes present in our dataset (GRIA1, ATP2B2). The low number of genes defining this pathway make it easy to obtain a high combined score in an enrichment analysis, even if you have just a few of those genes differentially expressed.

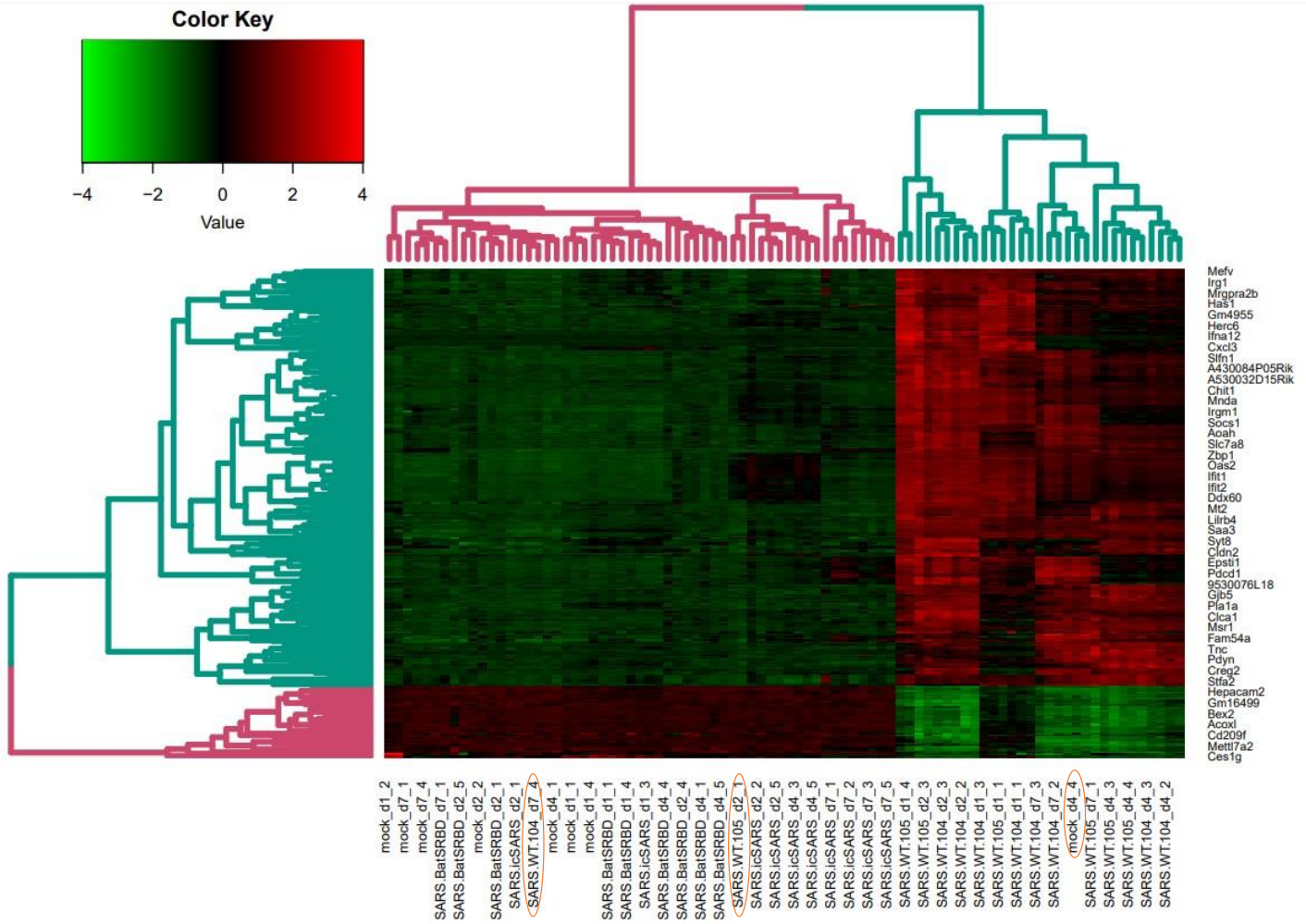


Figure 4: Heatmap of the 317 differentially expressed genes between the WT and the mock conditions. Up-regulated genes are represented in red and down-regulated genes are represented in green. A hierarchical clustering was applied both on rows and columns and determined two clusters of genes and two clusters of samples (pink and blue). Outliers are circled in orange.

To compare the gene expression of the different samples, we made a heatmap that shows the down-regulated genes (green) and the up-regulated genes (red). A hierarchical clustering forms two clusters of samples: the SARS-WT (blue) and the others (pink), as well as two clusters of genes **(Figure 4)**.

For the blue gene cluster, the genes are up-regulated for the SARS-WT strain and down-regulated for the other samples. For the pink cluster, the genes are down-regulated for the SARS-WT strain and are up-regulated for the other samples.

This heatmap was generated only with 317 DEGs from the SARS-WT vs mock. It allows to reproduce the clustering seen previously on the MDS representation (Figure 1), so it validates our first observation. These 317 genes can be considered as a biological signature.

Moreover, the outliers present in the MDS and that we excluded are also outliers in the heatmap.

Comparison of the response to SARS-WT strain at various time points

As only the SARS-WT strain induced a noticeable change in gene expression, we focused on this strain for this part of the analysis. To assess the heterogeneity of the response and the determinants of its variation, we first represented it with an MDS.

In order to see if the inoculation dose had an impact on the host response, we represented the inoculum on the MDS plot (10^4 or 10^5 PFU of SARS-WT) and didn't observe any segregation **(Figure 5)**. So, the dose of the inoculation doesn't have any effect on the response of the host. Further, a differential gene expression analysis showed no significant result. Thus, we concluded that inoculum had no effect on gene expression.

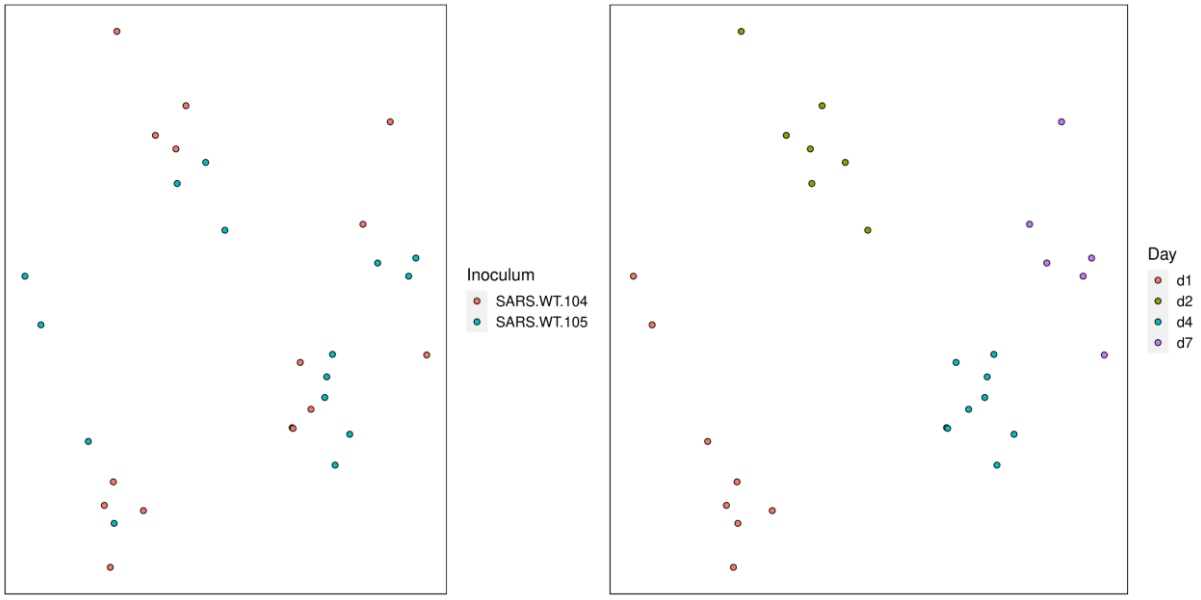


Figure 5 5(left): MDS representation of the effect of different inoculum doses on gene expression in SARS-WT samples. MDS was performed on SARS-WT samples only, but considering all 21200 measured transcripts.

Figure 6 (right): MDS representation of the effect of different timepoints on gene expression in SARS-WT samples. MDS was performed on SARS-WT samples only, but considering all 21200 measured transcripts.

We knew by the results that we had before that SARS-WT and mock have differentially expressed genes, and we wanted to know how the host response varied over the different timepoints **(Figure 6)**. We can see that the samples cluster according to timepoints (D1 to D7). Therefore, we will for the rest of the analysis focus on the variation of the response according to the different timepoints.

As we want to compare the host response to SARS-WT infection over the different time points by comparing it to the mock samples, we wanted to see the effect of the timepoint on the host

response for the mock condition, so we did an MDS representation using only the genes that are differentially expressed between WT and mock, and colored according to the time points. We see that mock samples are also separated accorded to days when we consider the genes that are the response elements to SARS-WT infection. Therefore, we will consider in the following that mock samples from different time points are not equivalent.

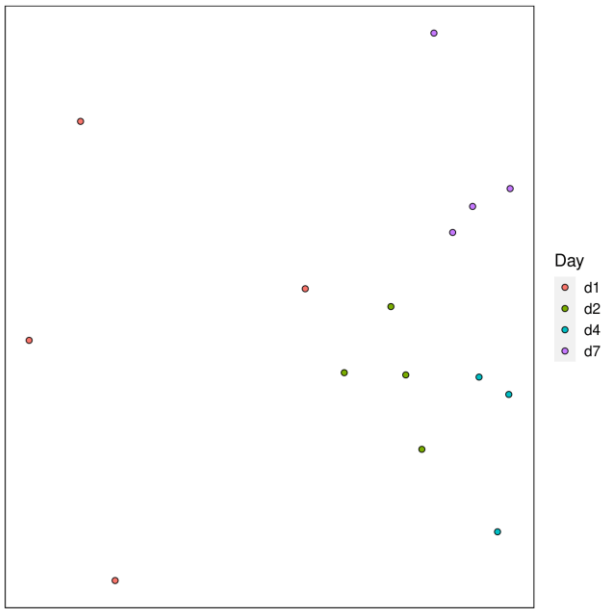


Figure 7: MDS representation of the effect of different days on gene expression in mock samples. MDS was performed on mock samples only, but considering only the 317 genes that are differentially expressed between SARS-WT and mock.

To compare the responses to the WT SARS strain between the 4 different timepoints (day 1, day 2, day 4, and day 7), we first performed differential gene expression analyses. For each timepoint, we compared gene expression in the WT SARS-stimulated samples to the mock samples of the same day and selected differentially expressed genes (p-value < 0.01 and fold-change > 20%). The mock samples from different days were not considered equivalent because they segregated on a MDS **(Figure 7)**.

Thus, we obtained 4 lists of 332 to 606 genes corresponding to each timepoint of the study. We compared these 4 gene sets with a setmap, indicating for each gene whether it is present or absent in the gene set of each day **(Figure 8)**. This representation showed only a partial overlapping of the gene sets among the different timepoints. Further, a clustering showed two major phenotypes: early response (days 1 and 2) and late response (days 4 and 7). Thus, new differential gene expression analyses were performed for each of these phenotypes rather than comparing each day, as fewer lists was easier to manage and interpret.

The WT SARS-stimulated samples from days 1 and 2 were merged and compared to mock samples from days 1 and 2 to obtain a list of 404 “early” differentially expressed genes. Similarly, we obtained a list of 446 “late” genes from the samples of days 4 and 7. To compare these 2 phenotypes, we represented both “early” and “late” gene sets on a Venn diagram **(**

Figure 9**).** This representation showed that some genes were response elements in only one of the two phenotypes (“unique early” and “unique late”), whereas others were differentially expressed at both timepoints (“common”).

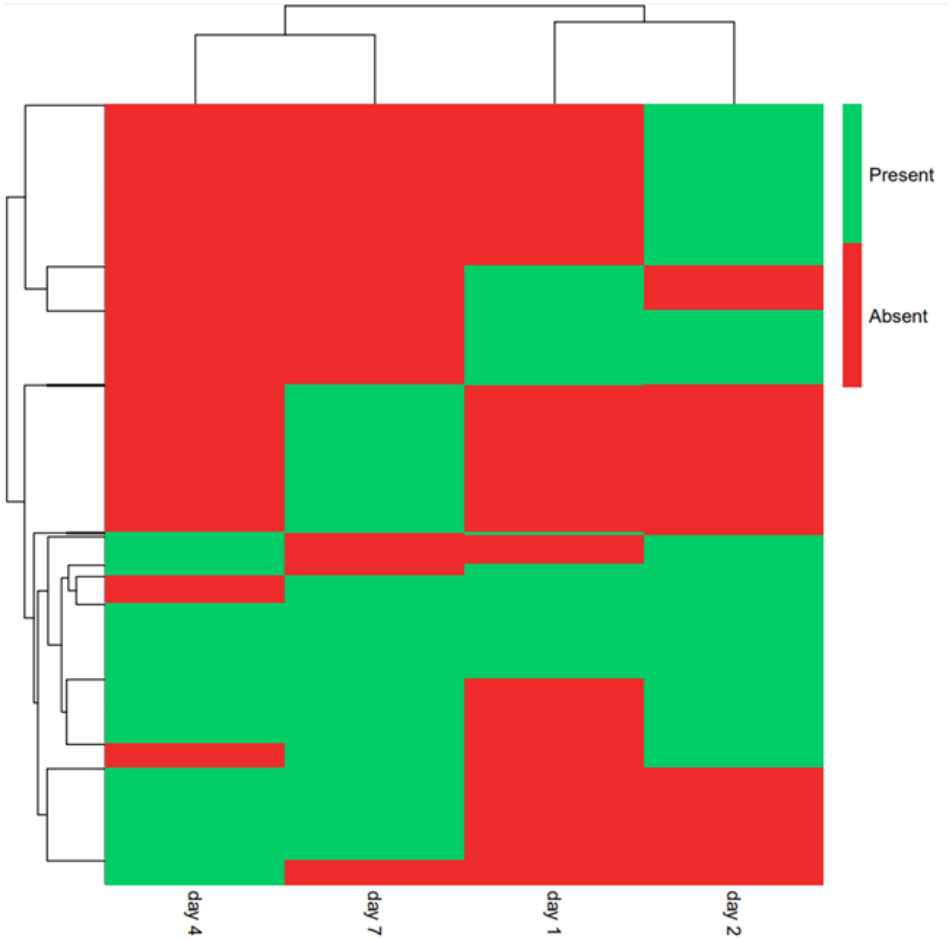


Figure 8: A setmap representation of the differentially expressed genes (SARS WT vs mock) at each timepoint (days 1, 2, 4, and 7) allows to distinguish two major phenotypes "early" (days 1 and 2) and "late" (days 4 and 7). On this setmap, all genes that are differentially expressed in at least one time point are represented in rows. These rows are colored to represent if they are present or not in the gene set of a given timepoint. Dendrograms represent the result of a clustering and identify the 2 major phenotypes.

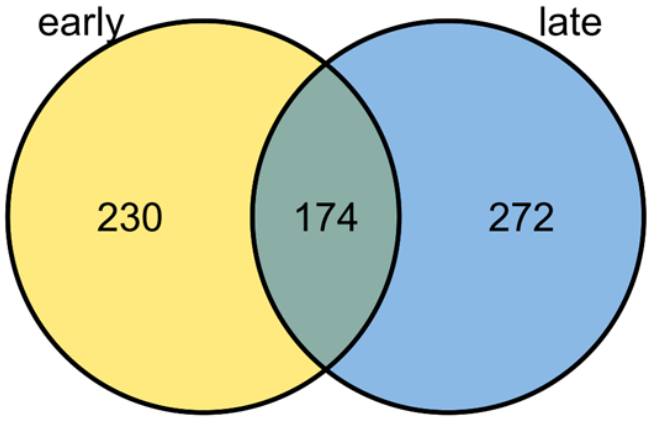
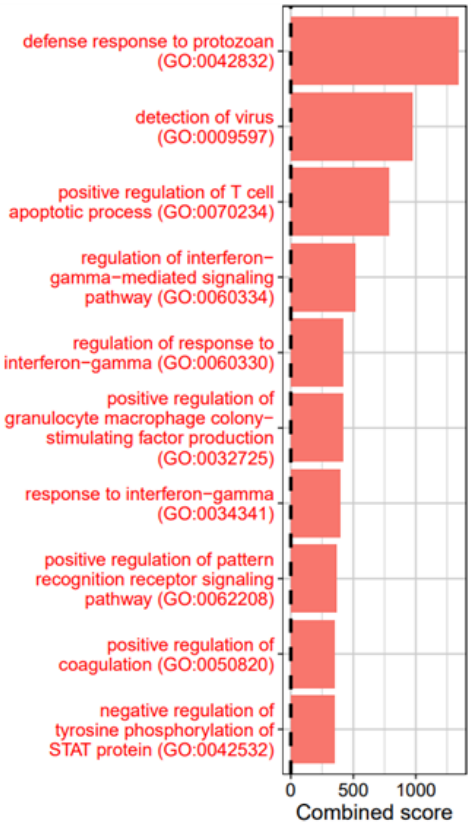


Figure 9: Venn diagram allowing to represent how many genes are uniquely differentially expressed in either "early" (1 and 2) or "late" (4 and 7) days, or differentially expressed at all timepoints. All three gene sets are about equally populated.

To assess the functional profile of these gene sets, we performed an enrichment analysis by comparing the gene sets to the Gene Ontology databases. Regarding the genes that are only differentially expressed in early days (

Figure 10) the upregulated genes correspond mostly to broad innate mechanisms (detection of virus, response to protozoans, pattern recognition receptor signaling) and to a pro-inflammatory cytokine profile (IFN-γ and GM-CSF). Surprisingly, there are also 4 genes related to T cell apoptosis (CD274, CCL5, LGALS9, and IDO1). To the contrary, no particular biological

function is enriched in the down-regulated genes, suggesting that all sorts of pathways are inactivated to save resources in order to activate immune mechanisms.



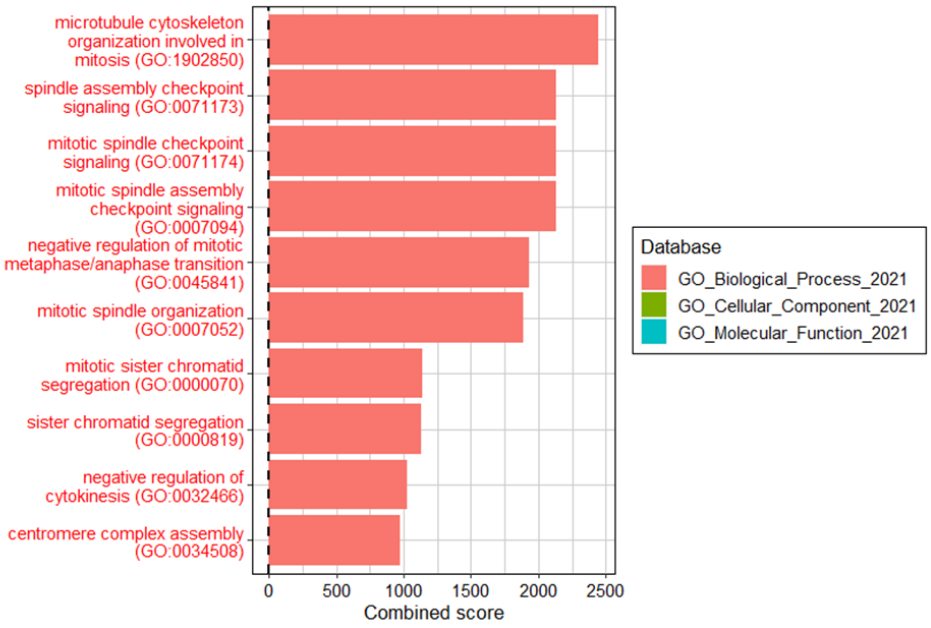


Figure 10: (left) An enrichment analysis of the genes that are differentially expressed only in "early" days shows a broad innate activation. The up-regulated and down-regulated genes were analyzed separately, and a term name is colored red (green) if it is enriched in up (down) regulated genes. Only significant terms were selected (FDR < 0.01) and the 10 terms with the highest combined score are represented.

Figure 11: (right) An enrichment analysis of the genes that are differentially expressed only in "late" days shows a specific activation of genes related to the inhibition of metaphase/anaphase transition. The up-regulated and down-regulated genes were analyzed separately, and a term name is colored red (green) if it is enriched in up (down) regulated genes. Only significant terms were selected (FDR < 0.01) and the 10 terms with the highest combined score are represented.

Regarding the genes that are only differentially expressed in late days **(Figure 11)**, there also is no enriched biological function in down-regulated genes. In the up-regulated genes, there is an enrichment of genes related to mitosis. More specifically, these genes are related to a "negative regulation of mitotic metaphase/anaphase transition", suggesting that the viral infection is inducing a cell cycle blockage. Although viral inhibition of the anaphase promoting complex is a well-known feat (Fehr and Yu, 2013), it usually depends on protein phosphorylation or degradation and not on the activation of the host transcription.

Overall, the functional analysis of enriched terms in the genes that are specific to either early or late days indicates the upregulation of genes related to immune pathways. In early days, a broad innate and pro-inflammatory response is established, but late days are characterized by a specific inhibition of mitosis.

Next, an analysis of the ratio of the expression (SARS-WT *versus* mock) of genes, which are differentially expressed both in early and late days, is represented on a heatmap **(Figure 12)** for all SARS.WT samples. This allows us to identify common upregulated and downregulated genes, that can be generally clustered into four different groups by their time points because of

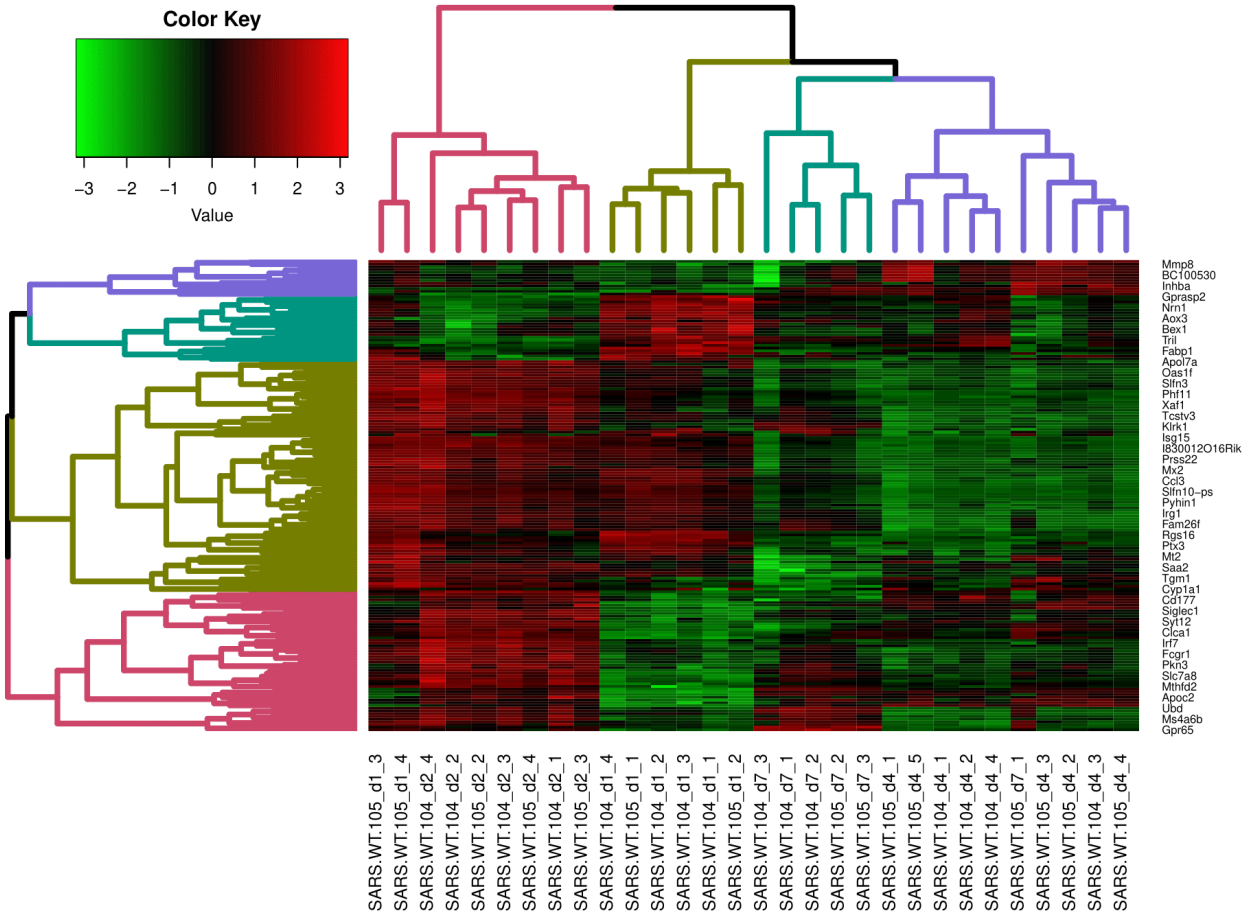


Figure 12: Inter-set heatmap of the 174 common genes determined in the previous Venn diagram. The heatmap allow us to see two different clustering based on two different time points, "early" (D1+D2) and "late" (D1+D2). The ratio is calculated as SARS.WT vs mock. Significantly upregulated genes are shown in red, while downregulated genes are shown in green. Four different clusters can be identified thanks to the column dendrogram which corresponds to the four different colors, pink, khaki green, turquoise and purple. Significantly upregulated genes are shown in red, while downregulated genes are shown in green. Four different groups of genes are differentially expressed, as shown in the row dendrogram, going from bottom to up, which are pink, khaki green, turquoise and purple respectively.

their similar gene expression as indicated by the column dendrogram. Those groups are "D1", "D2" and "D4" and "D7", and globally two sets of early (D1 and D2) and late D4 and D7) groups can be identified where the "early" group differs strongly from the "late" group.

EnrichR analysis has been performed on each of those four different time points **(Figure 13)** using 3 2021 Gene Ontology (GO) databases: Biological process, Cellular Component and Molecular Function, and shows the 10 most significant results.

Analysis of the pink cluster **(Figure 13, a)** shows 52 genes that are mostly upregulated in D2, involved in T cell chemotaxis. Among these 52 genes, the presence of CXCR3 binding suggests an increase in T cell trafficking and function (Groom and Luster, 2011) whilst CCR5 is involved in the selective recruitment of monocytes, T_H1-like T-cell clones, and peripheral T cells enriched for CD45RO+ "memory" (Weber et al., 2001). CXCL10 (or IP-10) is a ligand, binding to CXCR3 and favors T cell differentiation into Th17 and Th1 (Karin and Wildbaum, 2015). CXCL11 is the most potent CXCR3 agonist (Murphy, 2003) and favors Th2 subset along Tr1 producing IL-10 cells (Karin and Wildbaum, 2015).

EnrichR analysis of the khaki green cluster **(Figure 13, b)** reveals that 85 genes are upregulated in D1 and D2, majorly relating to NK chemotaxis and lymphocyte chemotaxis. T cell chemotaxis seems to be especially relevant along with CXCR3 chemokine receptor binding and production of cytokines related to Th2 cell subset.

In the third turquoise cluster **(Figure 13, c)**, 24 genes are expressed, significantly upregulated in D1, related to lymphocyte chemotaxis again, mostly due to CH25H; CCL20; SAA1.

In the fourth purple cluster, there was 13 significant genes significantly differentiated, however no enriched biological processes were returned by EnrichR. However, inputting the list of 13 in Reactome shows that they concern IL-6, IL-4, IL-10, and IL-13 along with IFN cytokines pathways.

In a nutshell, this analysis per time points suggests that two main groups, “early” and “late”, can be identified, where NK and lymphocyte chemotaxis seem to be especially prominent in the “early” group.

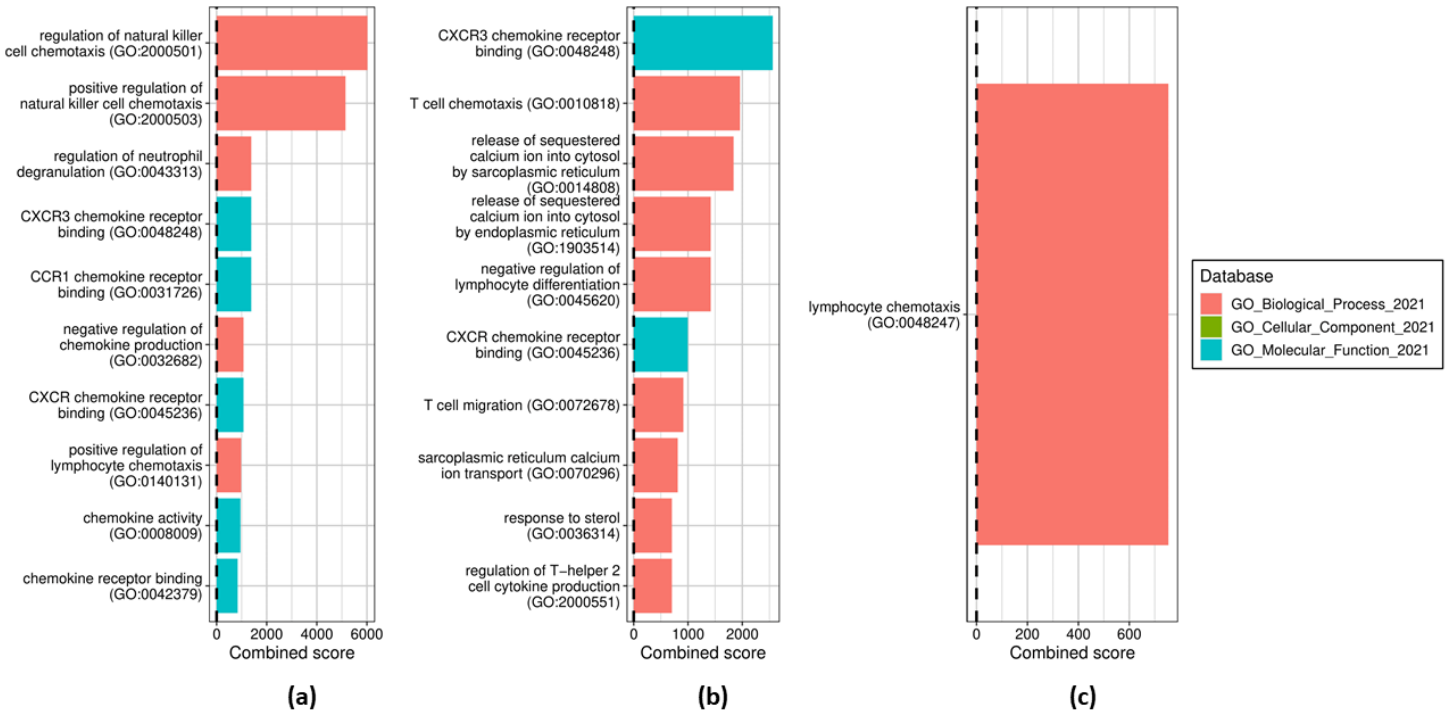


*Figure 13 : Functional enrichment analysis done by EnrichR. It was done on 3 different 2021 Gene Ontology (GO) databases: Biological process, Cellular Component and Molecular Function, and shows the 10 most significant results at a p-value threshold of 0.01. Most notably, the pink cluster **(a)** shows an overexpression of natural killer genes along with positive regulation of lymphocyte chemotaxis. Similarly, the khaki green **(b)** cluster shows multiple genes relating to T cell chemotaxis, while the small turquoise **(c)** cluster exclusively relates to lymphocyte chemotaxis.*

References

Aevermann, B.D., Pickett, B.E., Kumar, S., Klem, E.B., Agnihothram, S., Askovich, P.S., Bankhead, A., Bolles, M., Carter, V., Chang, J., et al. (2014). A comprehensive collection of systems biology data characterizing the host response to viral infection. Sci Data *1*, 140033.

Fehr, A.R., and Yu, D. (2013). Control the Host Cell Cycle: Viral Regulation of the Anaphase-Promoting Complex. J Virol *87*, 8818–8825.

Groom, J.R., and Luster, A.D. (2011). CXCR3 in T cell function. Exp Cell Res *317*, 620–631.

Institut Pasteur (2021). SRAS : informations et traitements - Institut Pasteur.

Karin, N., and Wildbaum, G. (2015). The Role of Chemokines in Shaping the Balance Between CD4+ T Cell Subsets and Its Therapeutic Implications in Autoimmune and Cancer Diseases. Front Immunol *6*, 609.

Murphy, P.M. (2003). Encyclopedia of Hormones. Cytokines Article Titles C 351–362.

Weber, C., Weber, K.S.C., Klier, C., Gu, S., Wank, R., Horuk, R., and Nelson, P.J. (2001). Specialized roles of the chemokine receptors CCR1 and CCR5 in the recruitment of monocytes and TH1-like/CD45RO+T cells. Blood *97*, 1144–1146.