Project Proposal

The research

During the current pandemic a lot of research has been done about COVID-19, which is caused by SARS-CoV-2 infection. Yet it is still unclear what causes the wide range of disease severity between individuals. The research from Chu et al. (2021) is about identifying the gene expression differences between COVID-19 patients that show mild or severe symptoms. Importantly and in contrast to related researches, these patients did not have any comorbidities so that any factors that may afflict the disease severity are excluded [1].

The severity was rated according to the answers of a patient to a set of questions about 15 different symptoms. The gene expressions were determined by carrying-out RNA-seq analysis on memory T-cells. Memory T-cells were obtained through blood sampling two weeks after a positive SARS-CoV-2-RNA test result. The RNA-seq data was then processed by use of the Drop-seq pipeline where the GRCh38 human reference genome was used for the alignment of the reads [1][2].

After statistical analysis the research found 31 differential expressed genes (DEG) with all of them being downregulated in the patients with a mild disease severity. Most of these DEG's were associated with the cell-cycle processes. The names of these genes are not given. In contrast, one upregulated gene (OASL) was called as significant even though this was not the case in the volcano plot. Furthermore, the IFI27 gene showed some upregulation but this was not significant [1].

Experimental design

The experiment was carried out using 5 different patients, subdivided into 3 patients that show mild symptoms and 2 patients that show severe symptoms. The severe afflicted patients have been hospitalized due to COVID-19. Both groups consists patients of both genders and different ages. In addition, all of the patients did not have any comorbidities, meaning that they did not suffer from any other conditions or diseases that may influence the disease severity. For all of the 5 patients, 3 technical repeats were performed. So all together 15 different analysis were performed [1].

Project

I have chosen to redo the RNA-seq data analysis in its entirety, this includes performing a principal component analysis, making a volcano plot and making a heatmap of the genes of interest. The most important part is finding the DEG's. A gene will be considered a DEG if they have an adjusted p-value < 0.05 and a fold-change ≥ 2 or ≤ 0.5.

The reason to redo the research is that the outcomes of the research were not very interesting (only some downregulated genes of the mild patients) and different results may be obtained if some other methods or choices are made. Moreover, the research did not give the specific names of the DEG's and I am still curious which genes were considered as a DEG. The research also contained some mistakes in their results, such as falsely calling an upregulated DEG, and it may be interesting to see if this gene is significantly upregulated or not.

Software

RStudio will be used to analyse the data and perform the statistical analysis [3]. In R the DESeq2 package (version 1.28.1) will be used to perform differential expressed gene analysis [4]. In addition, the pheatmap package (version 1.0.12) will be used to obtain the heatmap [5]. Furthermore, a functional enrichment analysis will be carried out using the clusterProfiler package (version 3.16.1) [6]. Lastly, the data will be visualised using the treemap package (version 2.4.2) [7].

Data

The used data is shown in the figure below. The rows consists out of all the different analysed genes and the columns are the different analysis (5 patients with each 3 repeats). The actual data are the raw count data which are obtained from the RNA-seq analysis.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	GENE	AAGATT	ATTACT	TAATGA	TTATTG	TCTGCA	AATACA	ATTCTA	TACTAT	TTGAAA	CCAACC	AATCTT	ATTTCA	TAGTAA	TTTACA	GTACCG
2	A1BG	0	0	2	0	1	0	0	0	0	0	0	0	1	1	0
3	A2M	1	2	1	1	0	0	1	0	1	1	3	1	1	0	0
4	A2ML1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
5	A4GALT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	AAAS	8	6	4	9	8	3	12	10	10	19	16	15	15	9	12
7	AACS	2	1	0	1	1	3	3	1	1	2	0	1	1	3	0
8	AADAC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	AADACL3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	AAGAB	21	12	9	13	18	18	14	17	19	30	20	19	17	12	23
11	AAK1	163	177	159	163	133	145	180	151	179	238	159	173	142	171	176

Literature

[1] C.-F. Chu, F. Sabath, S. Fibi-Smetana, S. Sun, R. Öllinger, E. Noeßner, Y.-Y. Chao, L. Rinke, E. Winheim, R. Rad, A. B. Krug, L. Taher en C. E. Zielinski, „Convalescent COVID-19 Patients Without Comorbidities Display Similar Immunophenotypes Over Time Despite Divergent Disease Severities," *frontiers in Immunologie,* nr. 12, 2021.

[2] Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell (2015) 161(5):1202–14. doi: 10.1016/j.cell.2015.05.002.

[3] RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

[4] Love MI, Huber W, Anders S. Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data With Deseq2. Genome Biol (2014) 15(12):550. doi: 10.1186/s13059-014-0550-8.

[5] Kolde R. Pheatmap: Pretty Heatmaps. (2012)..

[6] Yu G, Wang LG, Han Y, He QY. Clusterprofiler: An R Package for Comparing Biological Themes Among Gene Clusters. OMICS (2012) 16(5):284–7. doi: 10.1089/omi.2011.0118.

[7] Supek F, Bosnjak M, Skunca N, Smuc T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. PLoS One (2011) 6(7):e21800. doi: 10.1371/journal.pone.0021800.