Gabriela Ochoa
RNA-Seq Final Project

SARS-CoV-2 primarily impacts the lungs, infecting the cells of the respiratory system and creating inflammation and damage. Overall, older populations are deemed more susceptible to complications and at higher risk of mortality. There are still ongoing investigations of the long term impacts and health implications. I am curious about the immunosenescence, or the changes we see in the immune system associated with aging, that may impact the ability to form effective immune responses and considerations for development of long term impacts. *How does the gene expression profile of lung endothelial cells differ between uninfected young individuals and uninfected mid-aged individuals, and what implications does this have for age-related susceptibility to respiratory infections?* A follow up could then analyze the status of those cells upon infection of SARS-CoV-2 and compare the damage.

The data I chose to run came from a 2024 study conducted by Tsumita et al, who investigated the pathophysiology of lung endothelial cells in SARS-CoV-2 mouse models. The data set encompasses RNA-seq data using four types of isolated lung endothelial cells; uninfected and infected cells from young and middle aged mice. One challenge is that lung endothelial cells are largely heterogeneous, and so it can be difficult to use for studies taking into account the contamination of other cell types (Garlanda). This can cause issues with the purity of RNA-seq data and further analysis. A possible outcome would be negative implications of age on immune response of these lung endothelial cells. Future directions would include a statistical comparison of these cells with SARS-CoV-2 infection to note changes in gene expression and mark any deficiencies.

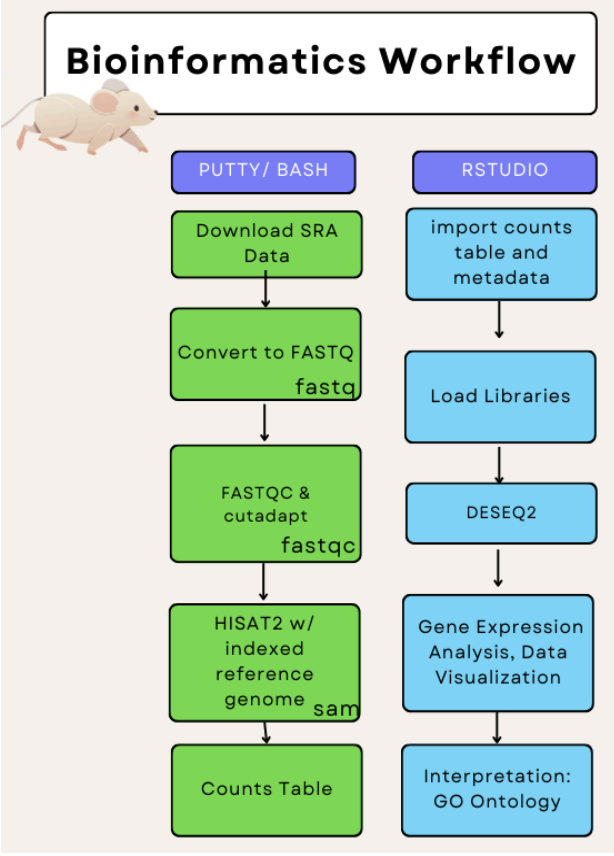


Figure 1. Bioinformatics Workflow

In order to understand the data, I decided to create a bash file to automate a script of commands and streamline this workflow to benefit from the accelerated speed of the KGI Server. I downloaded the NCBI SRA toolkit in order to extract 5 million reads from the four different SRA accession IDs provided and converted them to FastQ format. After renaming the files, I ran FastQ in order to

quality control the data and received the following FastQ results in Figure 2. From this, I used cutadapt to trim 200 bases from the beginning of each read and applied a quality control cutoff of 28, to provide some flexibility from the ideal 30 control cut off.

We then downloaded a mouse indexed reference genome, mm10, and used hisat2 to align the reference with the trimmed sequence reads, which gives a .sam file output. Lastly, I used this data against genome annotation data of Mus musculus genome from Ensembl to create a counts table of the expression profiles, seen in Figure 4. Moving away from PuTTY, the next steps of analysis were completed in R Studio.

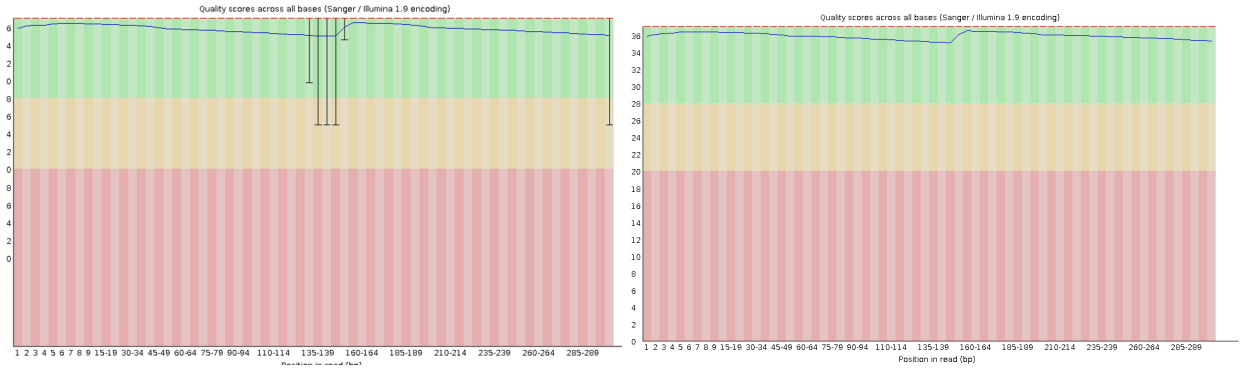


Figure 2. FastQC visual showing initial analysis of PBS Young and PBS Mid-Age mice

From the R Studio workspace, I visualized my counts table and created metadata to assign to DESeq2. DESeq2 is used for differential gene expression analysis in RNA seq data, which looks into the variability and provides statistical inference on the differential expressions. Upon analyzing the data, I found that there were no significant differences between the cell expressions of young PBS mice to mid age PBA mice (Figure 3). From here, I decided to move beyond my initial question, as GO enrichment would also not function on this data, to investigate the impact that SARS-CoV-2 has on differential gene expression of the sample cells.These results were then represented as a heatmap, volcano plot, GSEA graph, and in a gene enrichment chart.

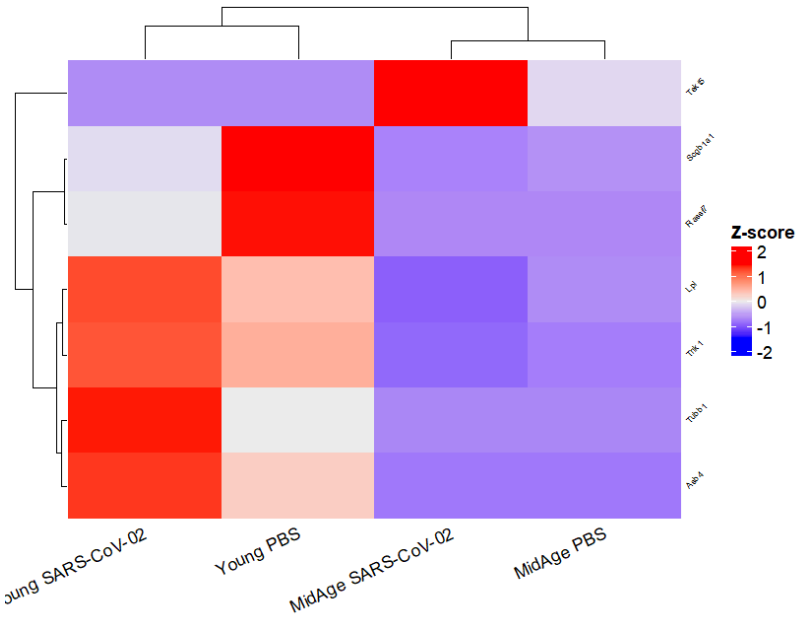


Figure 3. Heatmap with adjusted P valued at 0.5, gene expression of these few genes not statistically significant.

Interpretation & Conclusion

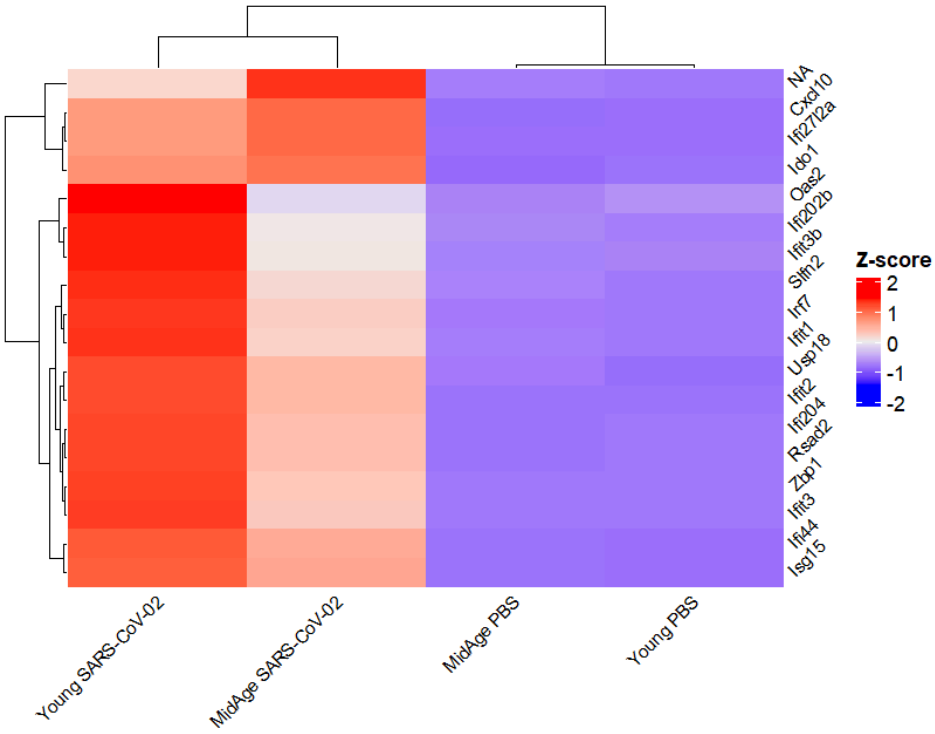


Figure 4 illustrates the top fifteen genes from the RNA-seq data and subsequent differences in gene expression profiles between the two groups. While it has been established that there is no statistical difference between age for the control PBS mice, it is clear that there are indications of gene expression variation between mice that have SARS-CoV-02 infection when compared to the PBS control group.

Figure 4. Heatmap illustrating top 15 fifteen genes and gene expression profiles from RNA seq.

The interpretation of this graph is higher expression in red color and lower expression in purple, indicating that the genes in the infected cells of young and mid age mice have an affected higher expression. Furthermore, figure 5 illustrates that these top 15 genes are involved in biological processes related to immune response, regulation, and defense.

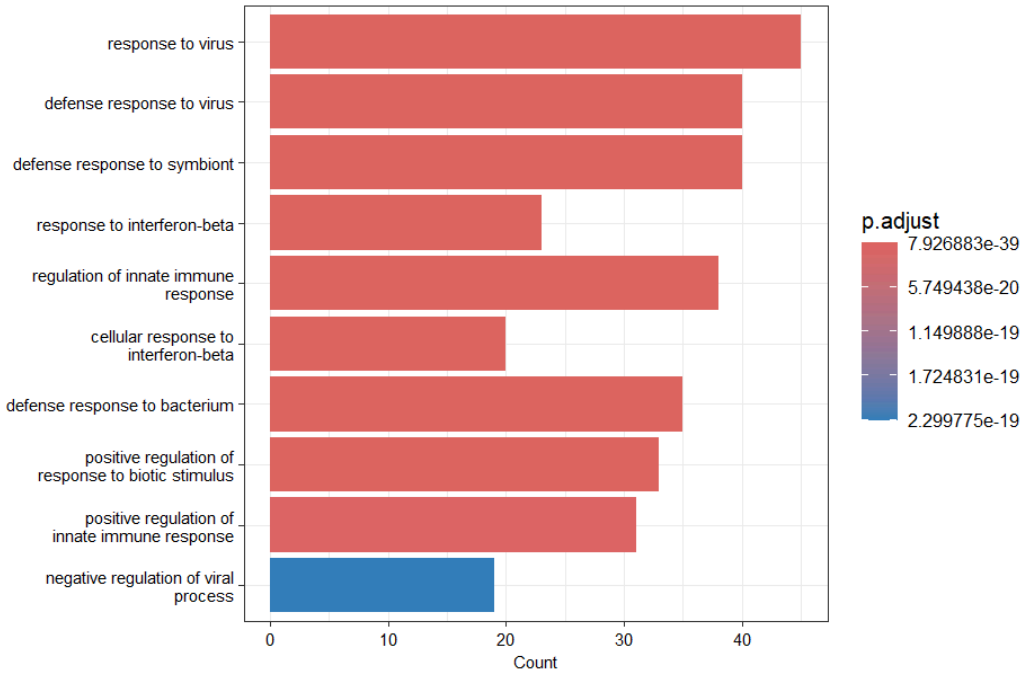


Figure 5. GO Ontology of SARS-CoV-02 RNA Seq Data; Biological Processes

This is really interesting to contextualize, as it indicates a direct involvement of these highly expressed genes in the immune response likely triggered by the infection of the SARS-CoV-02 virus. In addition, a GSEA plot then indicated which genes would contribute most to the enrichment score, those enriched by ontology, indicated by the peaks with positive and negative associations. For Figure 6, we can see that there is a varying association, with an ultimately positive association indicated by the position of the ranked list of genes.

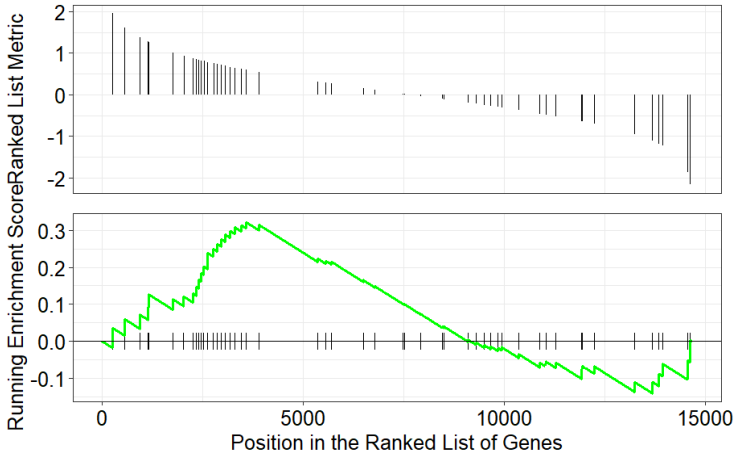


Figure 6. GSEA Plot for running enrichment score for ranked list of genes.

Lastly, we look at an enhanced volcano plot to further analyze this enrichment score ranking in correlation to differential expression analysis. The volcano plot of Figure 7 shows significance and magnitude change of gene expressions, with an interesting indication of genes IRF7, IFI44, and ZBP1 showcasing the highest magnitude change and significance. These genes are known for regulating myeloid derived suppressor cells, immune evasion biomarker, and positive regulation of defense response and programmed cell death respectively. These are key indicators of heightened gene expression resulting from SARS-CoV-02 infection in mice.

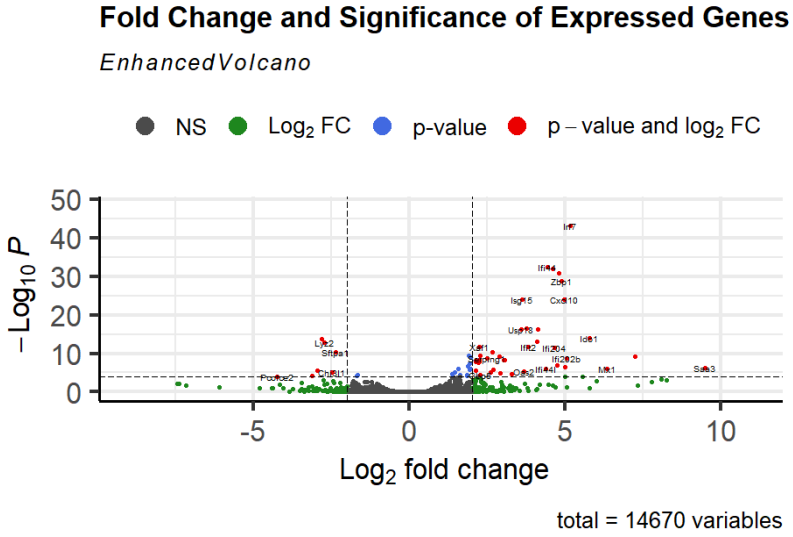
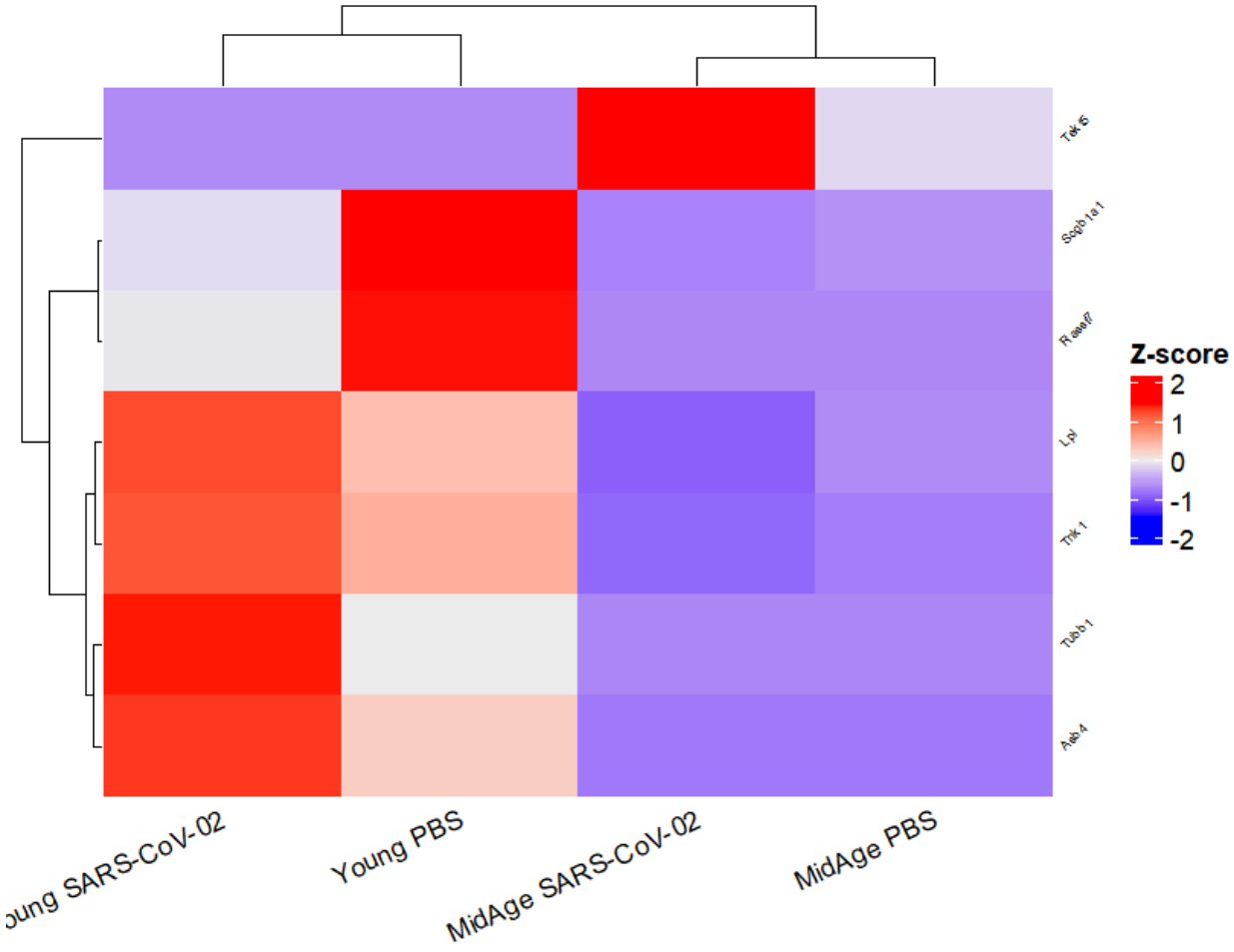


Figure 7. Enhanced volcano plot indicating magnitude and significance of gene expression.

Bonus:
I had a novel question of assessing the variations in age, as I was trying to determine if the concept of gene expression through development may play a role in the way that the body can be impacted by an infection of SARS-CoV-2.



This figure illustrates the results from that proposed question, which shows that there is no statistically significant difference between the sample cells we collected. This is insightful as we consider our perspective of mice models and analysis of disease, as the growth and development milestones have different implications across species. It can be determined that the development between a young and mid age mouse have no implications in regards to susceptibility or deterred immune response activity when faced with infections, a concept that was highly considered in regards to the infection affecting our most vulnerable populations for our human population and communities.

References:

Castanza, Anthony S., et al. “Extending Support for Mouse Data in the Molecular Signatures Database (MSigDB).” *Nature Methods*, vol. 20, no. 11, Nov. 2023, pp. 1619–20. *DOI.org (Crossref)*, https://doi.org/10.1038/s41592-023-02014-7.

EnhancedVolcano Function - RDocumentation. https://www.rdocumentation.org/packages/EnhancedVolcano/versions/1.11.3/topics/EnhancedVolcano. Accessed 5 May 2024.

EnhancedVolcano: Publication-Ready Volcano Plots with Enhanced Colouring and Labeling. https://bioconductor.org/packages/devel/bioc/vignettes/EnhancedVolcano/inst/doc/EnhancedVolcano.html . Accessed 5 May 2024.

Gao, Siqi, et al. “Endothelial SARS-CoV-2 Infection Is Not the Underlying Cause of COVID-19-Associated Vascular Pathology in Mice.” *Frontiers in Cardiovascular Medicine*, vol. 10, Sept. 2023, p. 1266276. *PubMed Central*, https://doi.org/10.3389/fcvm.2023.1266276.

Garlanda, Cecilia, and Elisabetta Dejana. “Heterogeneity of Endothelial Cells.” *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 17, no. 7, July 1997, pp. 1193–202. *ahajournals.org (Atypon)*, https://doi.org/10.1161/01.ATV.17.7.1193.

Ifi44 Interferon-Induced Protein 44 [Mus Musculus (House Mouse)] - Gene - NCBI. https://www.ncbi.nlm.nih.gov/gene/99899. Accessed 5 May 2024.

Irf7 Interferon Regulatory Factor 7 [Mus Musculus (House Mouse)] - Gene - NCBI. https://www.ncbi.nlm.nih.gov/gene/54123. Accessed 5 May 2024.

Perico, Luca, et al. “SARS-CoV-2 Spike Protein Induces Lung Endothelial Cell Dysfunction and Thrombo-Inflammation Depending on the C3a/C3a Receptor Signalling.” *Scientific Reports*, vol. 13, no. 1, July 2023, p. 11392. *www-nature-com.ccl.idm.oclc.org*, https://doi.org/10.1038/s41598-023-38382-5.

Zbp1 Z-DNA Binding Protein 1 [Mus Musculus (House Mouse)] - Gene - NCBI. https://www.ncbi.nlm.nih.gov/gene/58203. Accessed 5 May 2024.

Zheng, Qingcong, et al. “IFI44 Is an Immune Evasion Biomarker for SARS-CoV-2 and Staphylococcus Aureus Infection in Patients with RA.” *Frontiers in Immunology*, vol. 13, Sept. 2022. *Frontiers*, https://doi.org/10.3389/fimmu.2022.1013322.