ECBM E4060 - Intro to Genomic Information for Science and Technology

Project - Common human genetic variants of APOE impact murine Covid-19 mortality

Gemy Sethaputra[1] (gs3221), Eleftherios Koutsilianos[1] (ek3299), Eileen Choi[1] (sc4792), Onni Rauhala[2] (ojr2107), Alexander Ranschaert[2] (anr2157)

[1]Department of Biomedical Engineering [2]Department of Electrical Engineering

Abstract

Motivation: It is understood that genetic variants play a crucial role in the heterogeneity of Covid-19 outcomes. Quantifying the unexplored underlying mechanisms of genetic mutations, presents a great opportunity to further elucidate the causality between the genetic variants and mortality rates from Covid-19.

Results: We illustrate that mutations of the APOE gene in mice (APOE2,APOE4), exhibit an increased Covid-19 progression and viral loads relative to mice having the APOE3 gene. Additionally our findings suggest that APOE2,APOE4 negatively affect antiviral immunity.

Availability: https://github.com/aranscha/APOE_impact

Contact: ojr2107@columbia.edu, gs3221@columbia.edu, ek3299@columbia.edu, sc4792@columbia.edu, anr2157@columia.edu

1 Introduction

The fight to contain and cure COVID-19 has highlighted the relevance of bioinformatics and the ability to effectively analyze large volumes of genomics data. This rapid push to utilize computational tools in understanding biological systems and how they are affected by disease is showcased by the voluminous new literature on identifying genetic correlates with the varying symptom profiles of SARS-CoV-2 infection (van der Made *et al.,* 2022; Franke *et al,* 2020). In the effort to discover genomic markers for expected severity of the COVID-19 disease, some groups have also identified correlations between genetic traits that are traditionally linked with very different types of diseases, such as Alzheimer's or heart disease.

One such group recently described a strong association with the *APOE2/APOE4* variants of the *APOE* gene with worse COVID-19 disease outcomes using a genetic knock-in mouse model to evaluate the effects of the human genetic variants of *APOE* (Ostendorf *et al.,* 2022). By infecting mice with a knock-in human variant of *APOE* (*APOE2,APOE3 or APOE4*) with a subtype of SARS-CoV-2, Ostendorf *et al.* found that *APOE2/APOE4* variants confer significantly worse clinical outcomes than the *APOE3* variant in terms of survivability when controlling for age and sex, which are known to affect the outcome of COVID-19 (Ostendorf *et al.,* 2022, Fig 1). These results were further supported by evaluating disease progression as measured by fluorescent biomarkers for infection and necrosis in lung tissue from mice with each different genotype. Again, the analysis showed that possession of the *APOE2/APOE4* variants correlated with significantly faster and more severe disease progression (Ostendorf *et al.,* 2022, Fig 2). Secondly, the group identified modules of co-expressed genes from RNA sequenced lung tissue whose expression was selectively up- or downregulated depending on the *APOE*- genotype with *APOE2/APOE4* and *APOE3* conditions showing opposite modulations of these gene modules (Ostendorf *et al.,* 2022, Fig 3). Finally, single-cell RNA analysis revealed that *APOE2* and *APOE4* genotypes were associated with enrichment of different signaling pathways relative to *APOE3* suggesting that *APOE2* and *APOE4* confer their deleterious effects through different genetic and molecular mechanisms (Ostendorf *et al.,* 2022, Fig 3).

In this work, our objective is to reproduce the findings from Ostendorf *et al.* using the same murine data. The article also included human data from the UK Biobank, which they analyzed to find similar correlations between *APOE* variant and survival probability for COVID-19. However, due to access issues, we could not obtain the human data and hence focus only on the mouse data provided by the authors.

2 Results

2.1 *APOE* variants modulate outcome of mouse SARS-CoV-2 MA10 infection

To assess the impact of *APOE* variants in SARS-CoV-2 infection on mice, we used pre-processed data provided by the authors. (Ostendorf *et al.,* 2022). The Kaplan-Meier analysis approach was used to analyze the data, which allows the estimation of survival over time, even when data points were omitted (e.g. animal death in the study) or are studied for different lengths of time.

The statistical significance *P*-values of Figures 1a and 1b were computed by log-rank tests, similarly to the original paper's method. The log rank test allows us to test the null hypothesis of no difference in survival between two or more independent groups. The test compared the entire survival experience between the groups and could

be thought of as a test of whether the survival curves are identical (overlapping) or not.

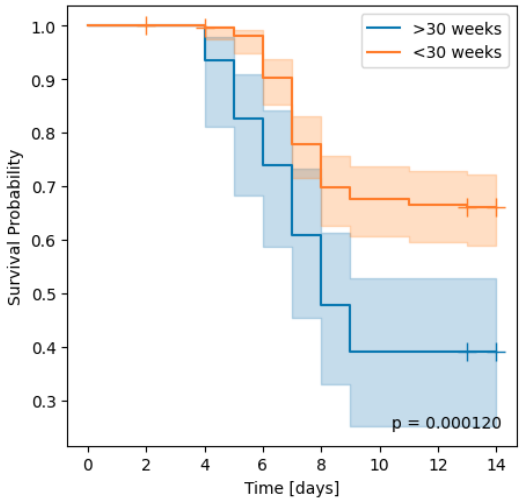

Fig. 1a. Survival rate of COVID-19 infected mice by age

Figure 1a demonstrated the survival of combined male and female SARS-CoV-2 MA10-infected *APOE*-knock-in mice stratified by age with the cutoff at 30 weeks, with obtained *P*-value of 1.2e-4.

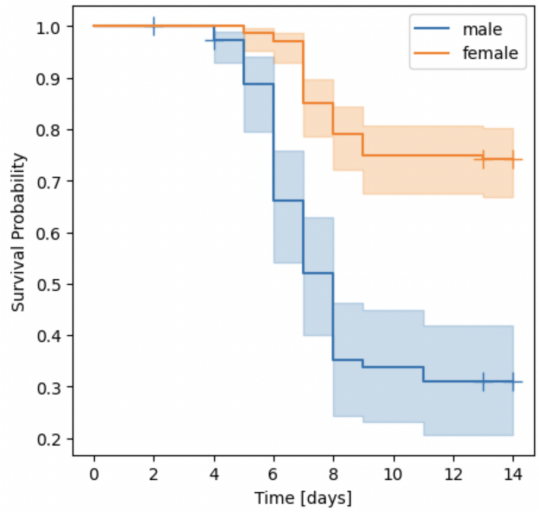

Fig. 1b. Survival rate of COVID-19 infected mice stratified by gender

Figure 1b demonstrated the survival SARS-CoV-2 MA10-infected *APOE*-knock-in mice stratified by gender, male and female with *P*-value of 2.6e-13.

The statistical significance *P* values of Figures 1c, 1d, and 1e were calculated by Cox proportional hazards (CPH) model, to compare the significance between multiple groups of variables, the three genotypes including *APOE2*, *APOE3*, and *APOE4*. The CPH model approach let us evaluate simultaneously the effect of several factors on survival rate. Simply put, it examined how specific factors influence the rate of a particular event happening (e.g., genotype) at a particular point in time. This rate is commonly referred to as the hazard rate. Predictor variables or factors are usually termed covariates in the survival-analysis literature.

While the CPH model and log-rank test methods are usually considered very similar, our rationale for using the CPH model for the following three comparisons was a drawback of the log-rank test: it cannot analyze other independent variables affecting the survival time. However, the CPH model is a semiparametric model that could analyze multiple independent variables for estimating differences between the survival curves. Independent variables can include the variable of interest (e.g. treatments) and other potential confounders (e.g. age of the patients). For this reason we decided to use the CPH method for the following 3 figures.

Additionally, the CPH test simultaneously evaluates the effect of several factors on survival within the model, whereas the log-rank test assumes the factors to be the same throughout. Since animal deaths were included in this study, we agreed that the CPH method would be a more appropriate approach.

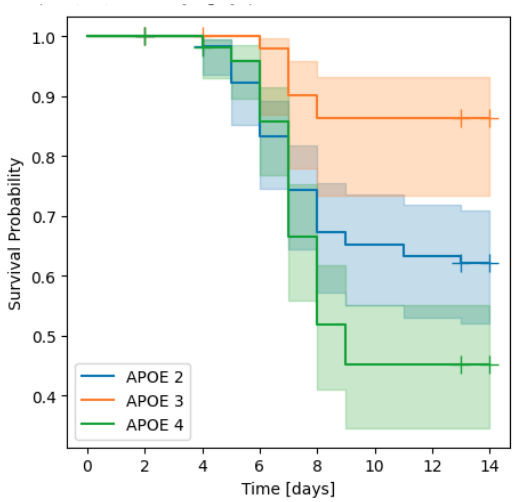

Fig. 1c. Survival rate of COVID-19 infected mice stratified by genotype

Figure 1c demonstrated the survival of combined male and female SARS-CoV-2 MA10-infected *APOE*-knock-in mice stratified by genotype including *APOE2*, *APOE3*, and *APOE4* with *P*-value of 3.1e-5.

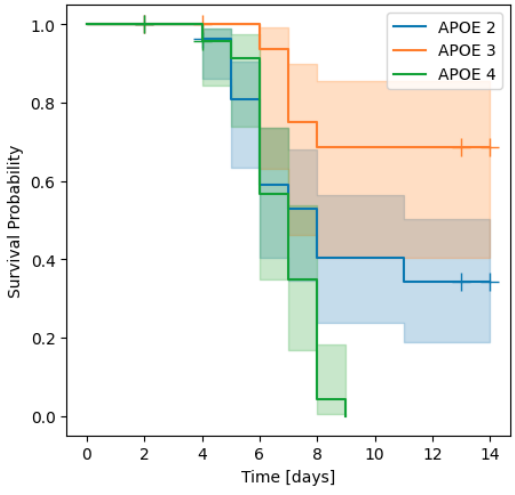

Fig. 1d. Survival rate of COVID-19 infected male mice stratified by genotype

Figure 1d demonstrated the survival of male SARS-CoV-2 MA10-infected *APOE*-knock-in mice stratified by genotype including *APOE2*, *APOE3*, and *APOE4*, with *P*-value of 3.6e-4

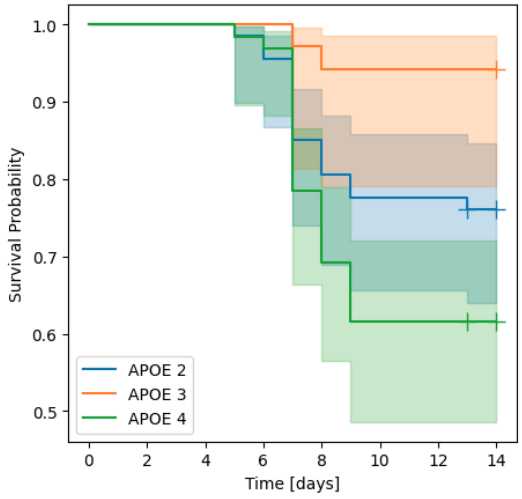

Fig. 1e. Survival rate of female subjects stratified by genotype

Figure 1e demonstrated the survival of female SARS-CoV-2 MA10-infected *APOE*-knock-in mice stratified by genotype including *APOE2*, *APOE3*, and *APOE4*, with *P*-value of 2.1e-3

2.2 Viral load analysis of *APOE3* against *APOE2/APOE4* mice for COVID-19

To perform viral load analysis, TaqMan quantitative real-time PCR was initially conducted 4 days post infection. The approach of the performed Taqman assays is beyond the scope of this study, and explicitly outlined in (Nagy *et al.*).

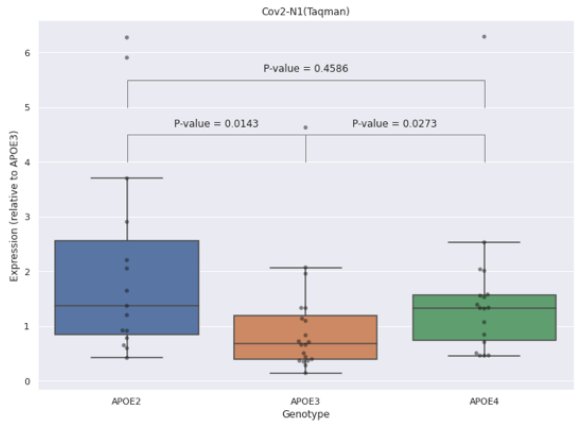

Fig. 2a. Statistical Analysis of Viral loads from Taqman PCR

Figure 2a illustrates the viral load for mice of each *APOE*-type gene, with the *P*-values calculated utilizing the Mann-Whitney U Test (H. & D., 1947), between *APOE2/-3,APOE4/-3* and *APOE2/-4*.

In accordance with the previously observed faster disease progression, substantial and consistent increase in viral loads in *APOE2*, *APOE4* mice can be deduced when compared to the viral load of *APOE3* mice. Their statistical significance is substantiated by *P*-values of <0.05 for both *APOE2/-3* and *APOE4/-3*.

To further validate our findings, we statistically assessed the data obtained from the nucleocapsid immunofluorescence staining ("An Introduction to Performing Immunofluorescence Staining," 2018). Similarly to Figure 2a, we observe a distinct increase in viral loads for *APOE2* and *APOE4*, with statistical significance guaranteed from the computed *P*-values.

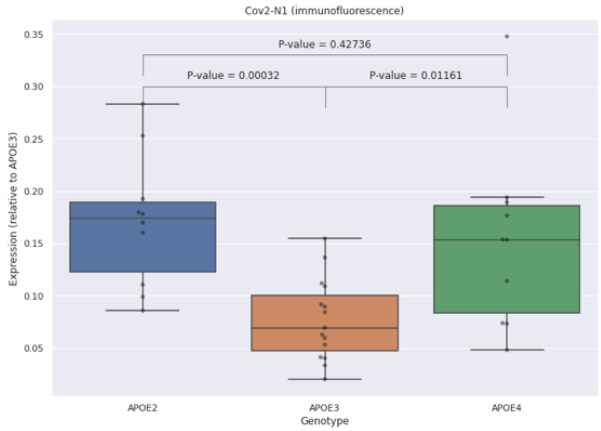

Fig. 2b. Statistical analysis of viral loads from immunofluorescence staining

Furthermore, to more specifically examine the effect of each *APOE* gene on the mice, several conditions were examined (Alveolar Damage, Fibrin Deposition, Bronchiolar Necrosis). Their level of expression was assessed, when compared to the presence of a specific *APOE* gene.

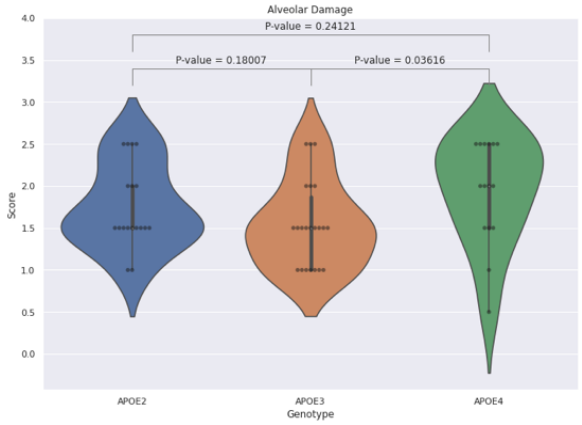

Fig. 2c. Scoring comparison of Alveolar Damage for *APOE*-type gene

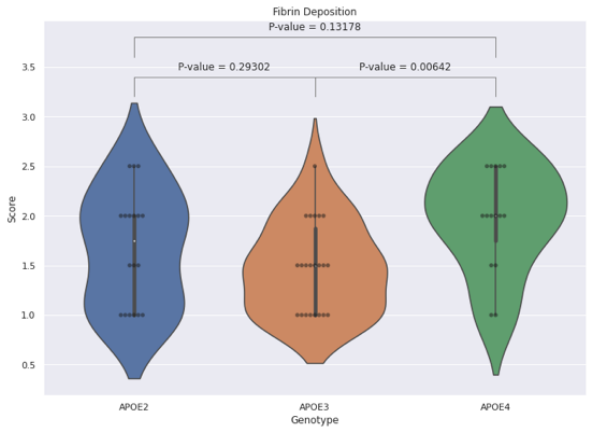

Fig. 2d. Scoring comparison of Fibrin Deposition for *APOE*-type gene

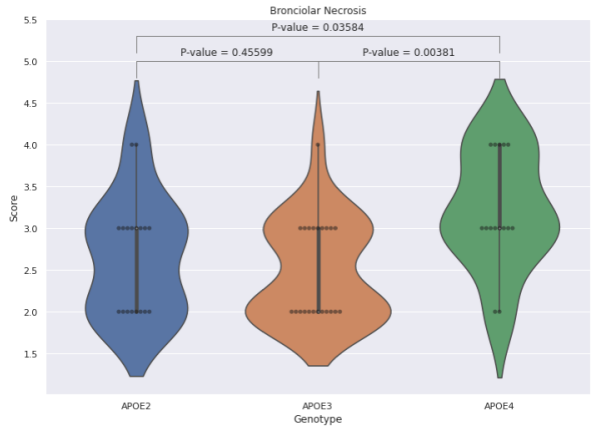

Fig. 2e. Scoring comparison of Bronchiolar Depo for *APOE*-type gene

The comparison analysis illustrates an increased presence of all three types of conditions in *APOE2* and *APOE4*, when compared against the *APOE3* mice. In particular, this increase is more pronounced in the case of *APOE4* where we observe a larger number of mice manifesting the injury conditions. In the case of *APOE2* mice, this increase is less substantial and could be deemed as not statistically significant, as documented by their respective *P*-values. As the authors do not further focus on this topic, it could potentially be an area of future research, in order to more specifically assess the relation between *APOE*-types of gene and lung conditions. Nevertheless, sufficient results have been provided to confidently suggest that there is a direct correlation between the *APOE2* and *APOE4* mice and the accelerated progression of COVID-19 when compared to the *APOE3* mice.

2.3 Impact of *APOE* on antiviral immunity

In order to investigate how *APOE* impacts antiviral immunity and viral infection, Ostendorf *et al.* sequenced both bulk RNA datasets and single cell datasets that were made publicly available. These datasets are analyzed in this section.

2.3.1 Bulk RNA Data

First, clusters of highly correlated genes are identified using weighted gene co-expression network analysis (WGCNA), first proposed as an R package in (Langfelder & Horvath, 2008). Because of this, the analysis here is also performed using R. These clusters can then be related to external characteristics to identify potential correlations between the clusters and these external factors. The analysis here was largely based on the tutorials provided by the authors of this R package (Langfelder & Horvath, 2016). The analyzed data is in the file *GSE184287_dds.rds*.

In a first step, the genes with a low expression were filtered from the expression matrix. It was decided here to keep genes with more than 50 total reads. After this first pre-processing step, 19,435/27,328 (71.12%) samples are kept. Along with normalization, a variance stabilizing transformation was performed, which could potentially reduce the skewness, and unmask relationships that were previously hidden. This transformation was also proposed in the original work.

Now, the power parameter for WGCNA can be determined. This power parameter is used to determine the adjacency between genes, based on the correlation matrix in the first phase of the algorithm, as the correlations are raised to this power to reduce the noise in the adjacency matrix. The package provides a function *pickSoftThreshold()* for this purpose. The output of this function is shown in Figure 3a.

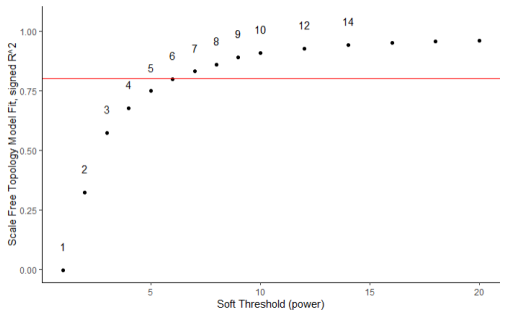

Fig. 3a. Selecting the GWCNA power parameter

The authors of the original package advise using a soft threshold power with an R^2 above 0.8 to avoid noisy results. The power should not be excessively large either, as this could reduce the number of identified clusters too much. The curve shown here has an inflection power around a power of 10 that has an R^2 above 0.8, so this is a suitable value for the power. Then, the WGCNA algorithm is executed.

The identified clusters are plotted in the dendrogram in Figure 3b. The dendrogram is plotted using the correlations between the eigenmodules, which are the first principal components of the expression matrix of one module, and thus function as a summary for this module. This is similar to Principal Component Analysis (PCA). 40 clusters or modules are identified, but from the dendrogram, it is apparent that many of these are similar. Therefore, modules are merged based on dendrogram correlations. A relatively low threshold of 0.15 reduced the number of modules from 40 to 10, which indeed confirms the similarity of some modules.

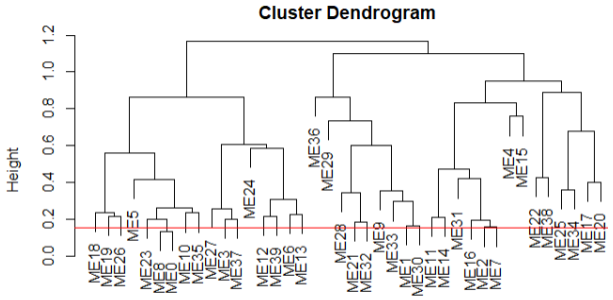

Fig. 3b. Cluster dendrogram to merge similar modules, based on their eigengene

Ostendorf identified 18 modules, so he was probably a bit less harsh on merging similar clusters. Moreover, he only used the top 30% most variably expressed genes as input to GWCNA. We now investigate the correlation between the module eigengenes and the *APOE* genotypes in Figure 3c. To compute the correlation with the genotype, this categorical variable was encoded by using the hazard rate of the genotype.

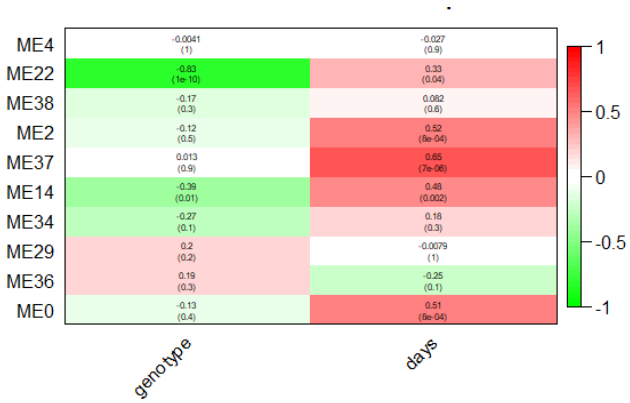

Fig. 3c. Correlations between the module's eigengene and the *APOE* genotype and days after infection.

Eigenmodule 22, which contains 84 genes, is distinctively negatively correlated with the *APOE* genotype, similar to the 'pink' cluster identified by Ostendorf. The top 10 hub genes of this pink cluster were all found to be in ME22. Similarly, ME2, slightly negatively correlated with the genotype and positively with the days after infection, contains all top 10 hub genes of the yellow module, and has a similar correlation pattern. On the other hand, ME14 contains all top 10 hub genes of midnightblue, but also the ones of the greenyellow cluster, so it is a combination of both. This makes sense, as the number of modules we have identified is smaller.

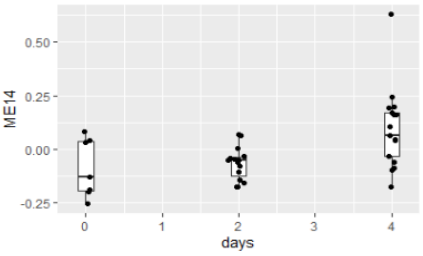

Fig. 3d. Upregulation of ME14 eigengenes during disease progression.

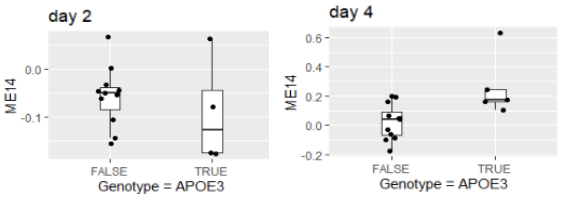

Fig. 3e. Upregulation of ME14 eigengenes during disease progression for the *APOE3* mice.

As in the homeworks, Reactome Pathway analysis was then performed on ME14, as there is an upregulation of gene expression in the *APOE3* mice during disease progression, as observed in Figures 3d and 3e. Table 1 shows that the expression of this module is related to the immune response, which is in line with the lower early immune response in *APOE2* and *APOE4* compared to *APOE3* (Fig. 3e). This was also found in the paper by Ostendorf.

Table 1. Reactome Pathway Analysis of ME14

Pathway	Gene ratio	*P*-value
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	14/116	1.29e-10
Signaling by the B Cell Receptor (BCR)	11/116	2.64e-08
Translocation of ZAP-70 to Immunological synapse	6/116	3.10e-08
Generation of second messenger molecules	7/116	5.50e-08
Antigen activates B Cell Receptor (BCR) leading to generation of second messengers	6/116	5.31e-07
Signaling by Interleukins	19/116	9.27e-07

2.3.2 Single cell Data

Single-cell RNA sequencing data was analyzed with the R code used by Ostendorf *et al.*, only modified to a single pipeline to reproduce the results from Figure 3e-g in the original paper (Ostendorf *et al.,* 2022). The raw data used in the R-analysis had been processed by the ParseBioscience pipeline (v0.9.6p) prior to deposition to the Gene Expression Omnibus for public access and analysis. The inputs for the R pipeline were a matrix containing the cell-wise sequenced expression counts per gene, a metadata file containing information on the cells and samples, and an info file containing the gene labels. Processing of the cell-gene matrix began by filtering out cells with fewer than 150 or more than 7,500 detected genes, cells with more than 40,000 unique molecular markers, and cells with more than 15% of mitochondrial data.

The pipeline, which used the Seurat R-package (Satija *et al.* 2015; Hao *et al.* 2021), then proceeded by performing log normalization on the matrix followed by detecting outliers on the mean variability plot. The data was then scaled before running PCA with the number of principal components set to 50. After PCA, Uniform Manifold Approximation & Projection (UMAP) was performed on the data with dimensions 1-30. Finally, clustering was performed by computing the nearest neighbors with 30 dimensions of reduction followed by computing the clusters with the original Louvain algorithm using a resolution value of 1.4.Three of the T-cell lineages (A, B & C T cells) identified in this first round of clustering were processed a second time with 20 dimensions for neighbors and a resolution of 0.5 for Louvain.

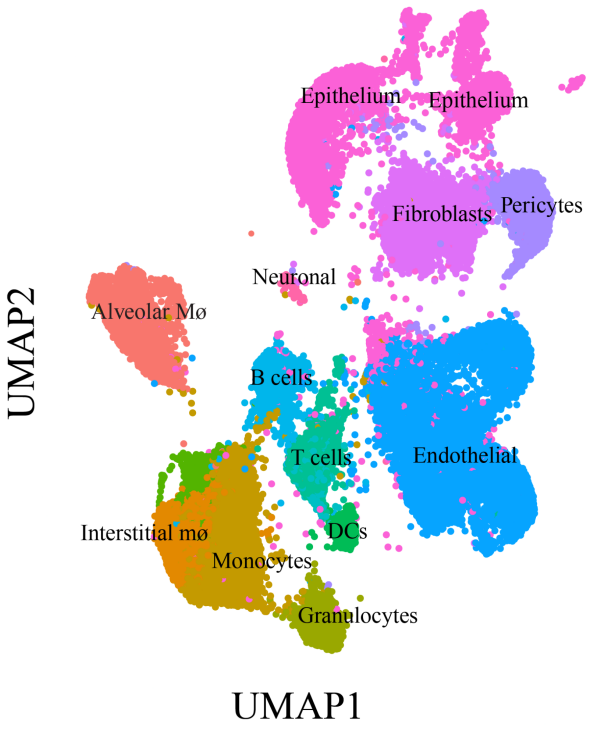

Fig. 3f. Cluster positions by cell type after UMAP

Cell types in each cluster were given by the authors based on previous literature evaluating the top expressed genes in single-cell data in specific lung cells. For group analysis, Ostendorf *et al.* further combined the cell types to the categories shown in Figure 3f, with each cell category labeled and color coded. By comparing the density of cells in each of these clusters across the experimental conditions, we find that mice infected with SARS-CoV-2 demonstrate an expansion in monocytes and endothelial cells and a reduction in epithelial cells as compared to the healthy controls (Figure 3g). Comparing the densities across *APOE* subtypes, there is also a notable expansion in monocytes in *APOE2* and *APOE4* as compared to *APOE3* but a reduction in epithelial cells (Figure 3h). These results are accurate replicas of the original paper and suggest differences in expression of different cells with the *APOE*-variants of interest and infected versus non-infected mice.

Next, gene set enrichment analysis (GSEA) to determine differentially expressed genes in *APOE2* vs *APOE3* and *APOE4* vs *APOE3* was performed. GSEA was also completed using the Seurat R-package and specifically the FindMarkers function. The ranking of differentially expressed genes was then calculated with (-log10(*P*-value)) / sign(log2(fold change)) and used with the clusterProfiler R-package to find relative enrichment of immune-related pathways in the cell groups. The pathways were selected from the Hallway gene set (www.gsea-msigdb.org).

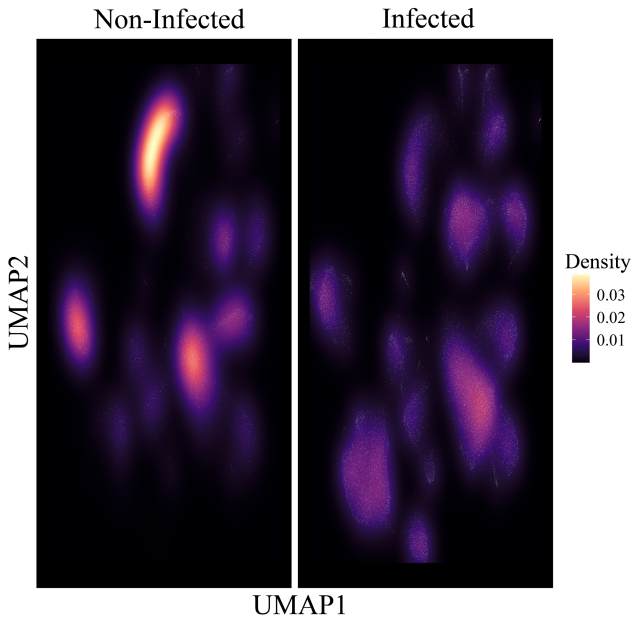

Fig. 3g. UMAP density for cells separated by SARS-CoV-2 infection status.

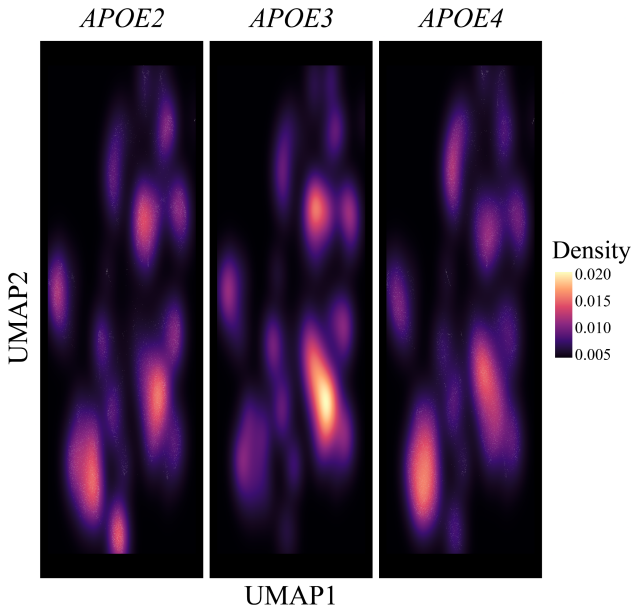

Fig. 3h. UMAP density for cells separated by *APOE* variant

With a few individual differences, the GSEA we ran yielded similar results to what was presented in Ostendorf *et al.* with *APOE4* showing lower enrichment in immune cells relative to *APOE3* (Figure 3i) while *APOE2* in fact had enhanced enrichment relative to *APOE3* (Figure 3j). As noted by the authors, this was an interesting finding as it highlights a potential difference in the ways *APOE2* and *APOE4* may confer their adverse outcomes with COVID-19. Taken together, these analyses successfully replicated the most important single-cell RNA sequencing results presented by Ostendorf *et al.* and

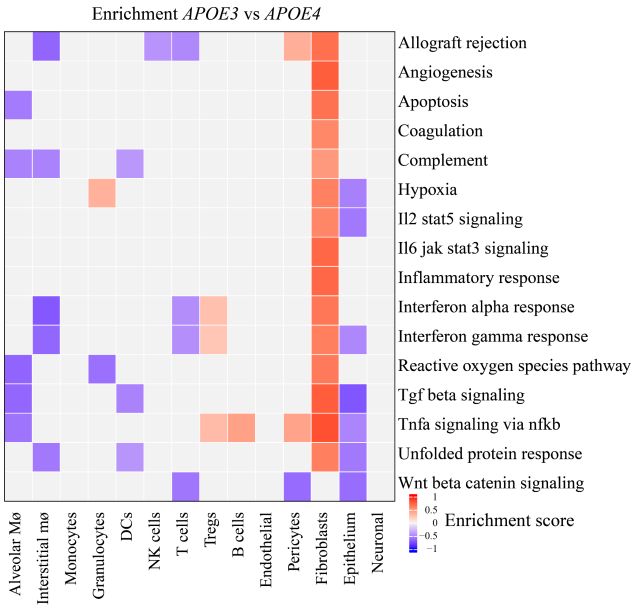

Fig. 3i. Enrichment scores for *APOE4* vs *APOE3* for various cell types

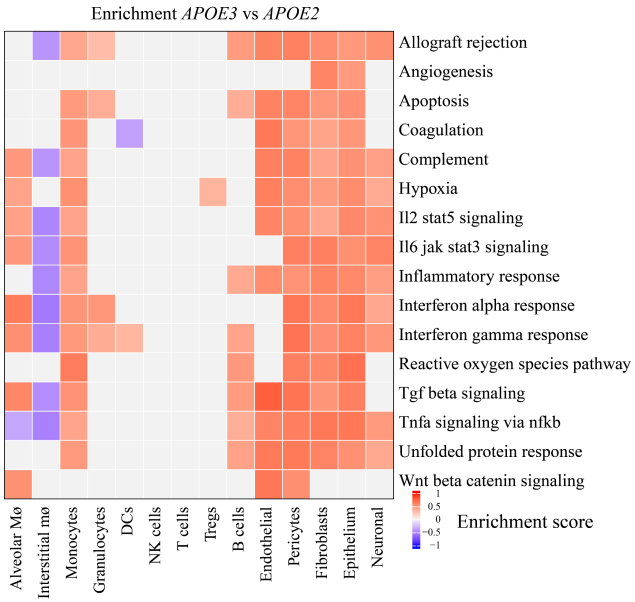

Fig. 3j. Enrichment scores for *APOE2* vs *APOE3* for various cell types

Conclusion

The objective of this work was to replicate and, where appropriate or possible, refine the results from Ostendorf *et al.* on the effect of *APOE* genotype on the clinical presentation of SARS-CoV-2 infection in mice. Overall, the survival, tissue marker and RNA sequencing analyses we ran, all paralleled and reproduced the findings presented by the authors of the original article. We found that *APOE2* and *APOE4* genotypes were significantly associated with greater mortality in mice with COVID-19 and that these genotypes also conferred a greater degree of lung damage in infection as compared to the *APOE3* genotype. Analysis of the bulk RNA sequencing data revealed eigengene modules correlated with the days after the infection and the genotype. One of these was found to be related to immune system activation, which showed a lower immune reaction onset in *APOE2* and *APOE4* mice. Finally, although we utilized the same code used by the authors with slight modifications, analysis for single cell RNA data was successfully replicated and showed the same expression patterns in different cell types in mice with the included *APOE*-variants and infection status as reported previously. Enrichment analysis also replicated successfully with the result suggesting a differential immune pathway recruitment between *APOE2* and *APOE4* when comparing the two to *APOE3* and thus a possible difference in how the two deleterious variants produce their effects. Taken together, our work supports the methodology and verifies the results given by Ostendorf *et al.*

References

Franke,L. *et al.* (2020) Genomewide association study of severe COVID-19 with respiratory failure. *NEJM*, **383**, 1522–1534.

Hao,Y. *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573-3587.

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 9(1). https://doi.org/10.1186/1471-2105-9-559

Langfelder, P., & Horvath, S. (2014). Tutorial for the WGCNA package for R: I. Network analysis of liver expression data in female mice 2.b Step-by-step network construction and module detection. https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/FemaleLiver-02-networkConstr-man.pdf

Ostendorf,B.N. *et al.* (2022) Common human genetic variants of *APOE* impact murine COVID-19 mortality. *Nature*, **611**, 346–351.

Satija,R. *et al.* (2015) Spatial reconstruction of single-cell gene expression data. *Nature Biotech.*, **33**, 495-502.

van der Made,C.I. *et al.* (2022) Clinical implications of host genetic variation and susceptibility to severe or critical COVID-10. *Genome Medicine*, **14**, 96.