Final Team Project Writeup: Chiraag Gohel and Varun Subramaniam

Understanding gene heterogeneity across COVID-19 severity through scRNA-seq data

1. Research Strategy

Question

We seek to understand biological differences between different severities of COVID-19 infection. Specifically speaking, is similarity in gene expression driven by COVID-19 infection type or other confounding variables? Additionally, we are interested in exploring differences in gene expression conditioning upon COVID-19 infection type.

Technology

We believe that these questions can be best answered from data collected through single-cell sequencing technologies (scRNA-seq). We specifically use the Drop-Seq single cell protocol from Illumina.[1]

Data

scRNA-seq was performed on nasopharyngeal swabs from 58 healthy and COVID-19 participants. Metadata variables include COVID-19 status, COVID-19 intensity, Sex, Age, and Bloody Swab (a variable indicating if the swab used to collect nasopharyngeal cells also had blood residues), among several other indicators. Our data come from the paper: “Impaired local intrinsic immunity to SARS-CoV-2 infection in severe COVID-19” by Ziegler et al.[2].

Approach

We used `Seurat`[3] for the initial data preprocessing, cleaning, quality control, normalization, and dimensionality reduction of our transcript data. We used `omeClust`[4] for clustering of transcript data, association of clustering to metadata, and clustering visualization. We used `tweedieverse`[5] for gene differential expression testing across COVID-19 severity.

Interpretation

Figure C demonstrates differences in gene regulation across COVID-19 severity. We see that certain genes, such as HOPX, are upregulated in severe COVID, but not moderate COVID. This analysis can be used to identify biomarkers for severe COVID, which is important for developing tests and therapeutics. Our analysis also specifically corroborates the finding from Ziegler et al. that KRT13 is upregulated with more intense COVID severity[2].

2. Specific Aims

The overall goal of our investigation is to determine the interplay between COVID-19 and the expression of various genes in the nasopharynx. In pursuit of this overarching objective, we have identified four specific aims, outlined below.

Specific Aim 1: Clean scRNA-seq data and metadata for further analysis.
The scRNA-seq data is recorded in a large matrix file, accompanied by a barcode (cell label) and feature (gene label) file. These files were combined and loaded into a singular Seurat object. The original dataset contained 32,871 genes, and 32,588 cells. For quality control, we subsetted cells that contained at least 200 unique genes, and less than 7,500 unique genes (to avoid doublet contamination). We also subsetted genes with less than 5% mitochondrial gene expression, to prevent downstream analysis of dead cells. The cleaned dataset was normalized using `SCTransform`[6]. Normalized data was aggregated across patient samples, and used as input to create a distance matrix between samples (see `loading-data.R`).

Specific Aim 2: Create and visualize initial clusters from expression data.
The distance matrix generated from the cleaned and normalized data was used to calculate clusters in `omeClust`[3]. The first output from `omeClust` identified 5 major clusters, with two further clusters containing data from fewer than three participants. These clusters are shown in Figure A below. The following code was written and used in Terminal (OSX) to produce initial cluster output. The directory was set as a locally hosted folder linked to this shared GitHub repository `pubh6885-team-project`.

```
omeClust -i data/dist/dist_matrix_norm.tsv -o data/omeclust/norm --plot
```

Specific Aim 3: Incorporate disease ontology labels into cluster visualization.
Before producing further cluster visualizations, we cleaned the metadata by removing contextually irrelevant columns, isolating unique rows, and setting row names equal to donor IDs for matching with the distance matrix produced by `omeClust` (See `cleaning_metadata.R`). Given that the focus of this investigation was the interplay between COVID-19 status and gene expression, we ran `omeClustviz` to shape points on the cluster plot by the disease ontology label indicator in the metadata[3]. This variable holds the specific disease status–normal, COVID-19, long COVID-19, or respiratory failure–of each patient in the dataset. The updated cluster plot generated from `omeClustviz` is shown in Figure B below. The following code was once again written in Terminal to produce this visualization with the same directory as above.

```
omeClustviz data/omeclust/norm/adist.txt data/omeclust/norm/clusters.txt --metadata
    data/tweedieverse/cleaned_metadata.tsv --shapeby disease__ontology_label -o
                            output/omeclustviz/disease
```

Specific Aim 4: Compare differentially expressed genes across COVID-19 severity
Differential gene expression testing was performed using `Tweedieverse`, using raw count data from the cleaned Seurat object as input (See `tweedieverse.R`). Testing was performed using normal control cases as a reference group against COVID-19, long COVID-19, and respiratory failure COVID-19 patients. Figure C demonstrates differences in gene regulation across COVID-19 severity

and was produced using R (see `tweedie_viz.R`). Our analysis primarily highlights genes that are only upregulated or downregulated for a specific type of COVID-19 severity. For example, we found that HOPX is significantly upregulated in severe COVID-19, moderately upregulated in long COVID-19, and is actually downregulated in short COVID-19. Thus, HOPX can be understood as a biomarker distinguishing the severity of a potential infection.

3. Figures

(A) Principal Component Analysis Plot

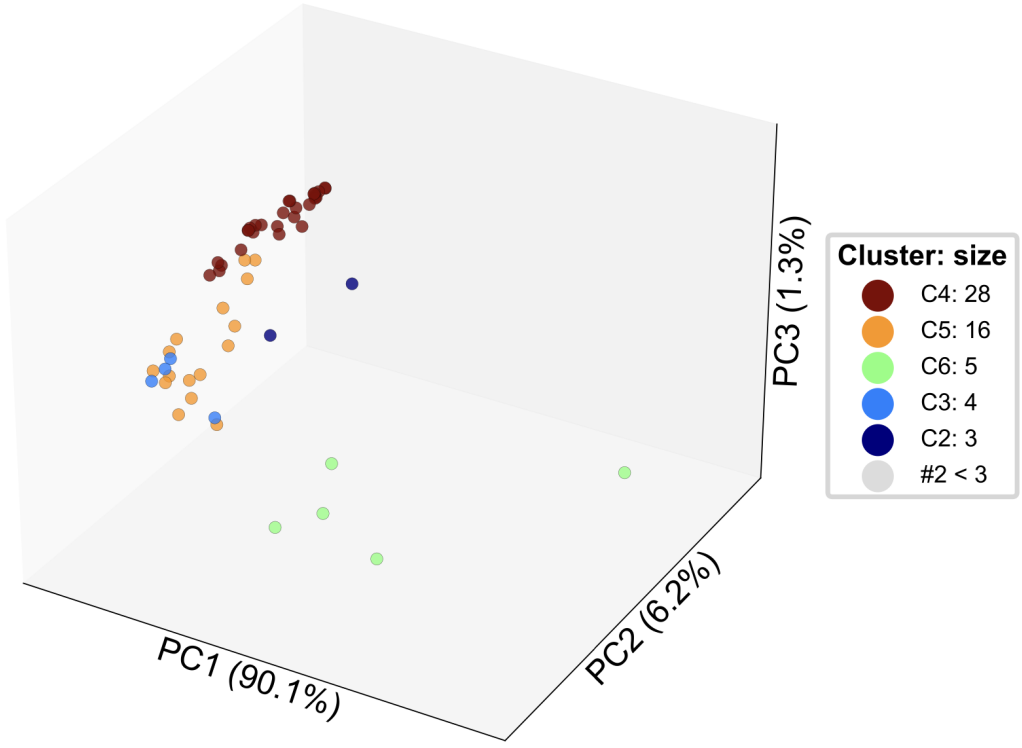

Fig A. Output from `omeClust` of normalized, aggregated distance matrix data for nasopharyngeal counts. There are five major clusters visible, denoted by color. The largest cluster C4 contains gene data from 28/58 participants. There are two minor clusters (not plotted), each containing data from fewer than three participants.

(B) Principal Component Analysis Plot Shaped By Disease Ontology Label

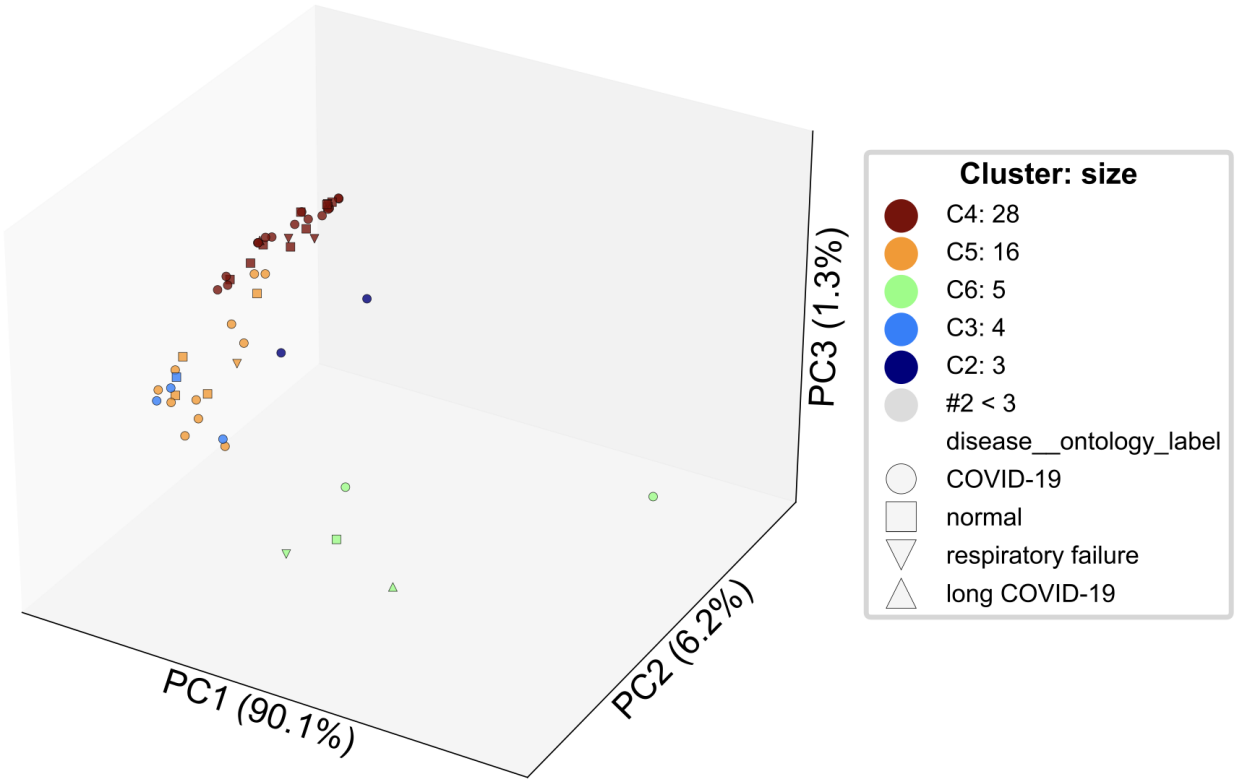

Fig B. Output from `omeClustviz` of total distance matrix data from initial omeClust output. Identical clusters to those in Figure A have been plotted, this time shaped by disease ontology labels. Clearly, these shapes are distributed across the plot, warranting further statistical (rather than visual) evaluation of relationships between COVID-19 severity and gene expression.

(C) Tweedieverse Differentially Expressed Genes

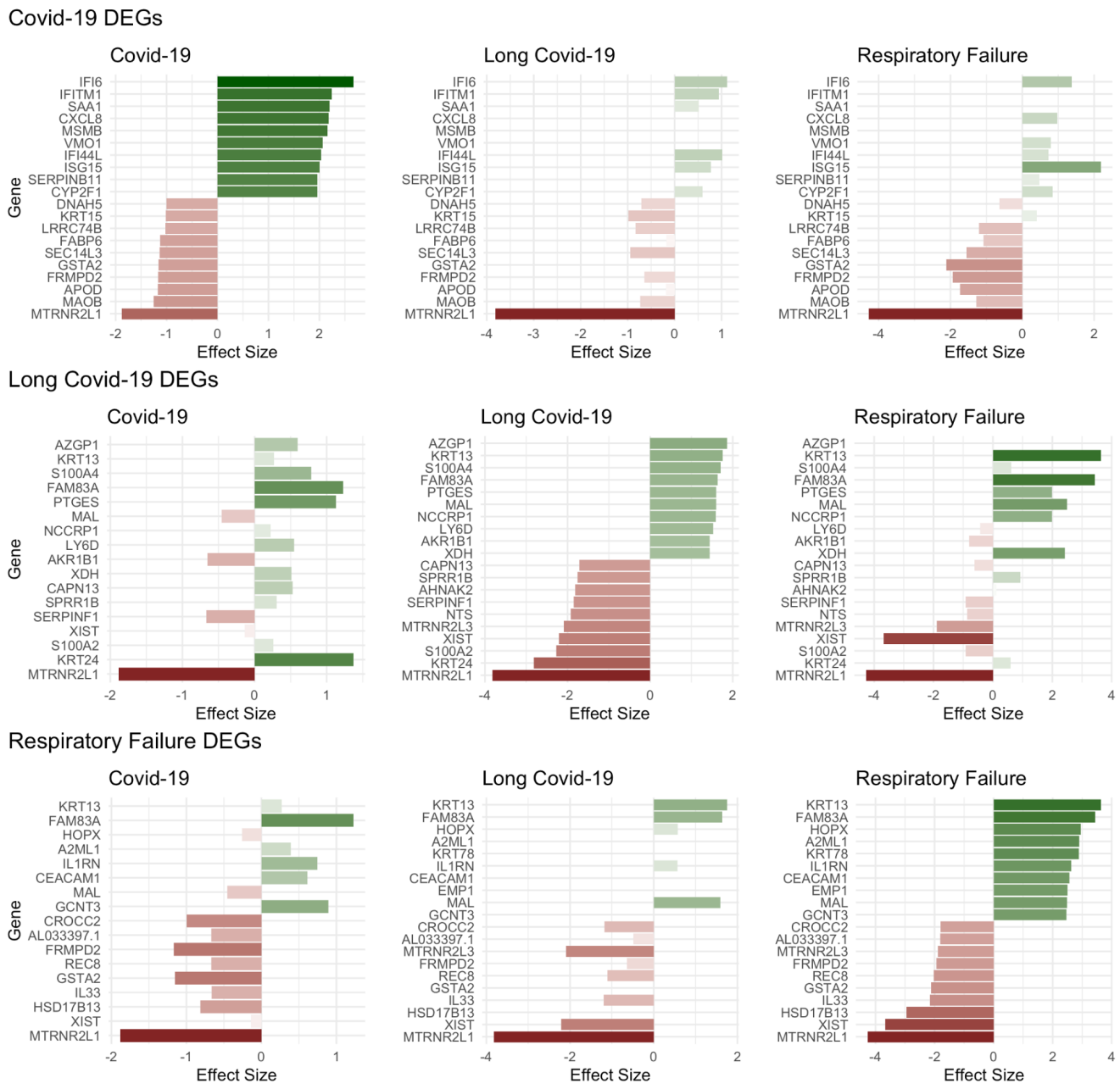

Fig C. `Tweedieverse` effect size coefficients for various genes conditioning on COVID-19 severity against the normal reference. The first row compares the top ten upregulated and top ten downregulated genes for moderate COVID. The second row compares the top ten upregulated and top ten downregulated genes for long COVID. The third row does the same for severe COVID. Several genes, such as HOPX, demonstrate upregulation in certain severities, yet not others.

4. References

1. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).
2. Ziegler, C. G. K. *et al.* Impaired local intrinsic immunity to SARS-CoV-2 infection in severe COVID-19. *Cell* **184**, 4713-4733.e22 (2021).
3. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
4. Rahnavard, A. *et al.* Omics community detection using multi-resolution clustering. *Bioinformatics* **37**, 3588–3594 (2021).
5. Mallick, H. *et al.* Differential expression of single-cell RNA-seq data using Tweedie models. *Stat. Med.* **41**, 3492–3510 (2022).
6. Choudhary, S. & Satija, R. Comparison and evaluation of statistical error models for scRNA-seq. *Genome Biol.* **23**, 27 (2022).