Assessing the impact of cell type deconvolution on differential gene expression analysis of COVID-19

Li Qing Wang, Jana Osea, Erick Navarro, Daniela Yanez
STAT540 Group 2

Apr 6, 2022

Outline

1. **Introduction & Aims**
2. **Aim 1:** Cell type deconvolution
3. **Aim 2:** Differential gene expression analysis and gene set enrichment analysis
4. **Aim 3:** Quantitative assessments of cell type model
5. **Conclusion & Limitations**

INTRODUCTION

- Mechanism of COVID yet to be fully elucidated
- Cell type information is hidden in bulk tissue sequencing

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

In vivo antiviral host transcriptional response to SARS-CoV-2 by viral load, sex, and age

Nicole A. P. Lieberman, Vikas Peddu, Hong Xie, Lasata Shrestha, Meei-Li Huang, Megan C. Mears, Maria N. Cajimat, Dennis A. Bente, Pei-Yong Shi, Francesca Bovier, Pavitra Roychoudhury, Keith R. Jerome, Anne Moscona, Matteo Porotto, Alexander L. Greninger

Published: September 8, 2020 • https://doi.org/10.1371/journal.pbio.3000849

- Sample size (n = 409):
  - 356 positive, 53 negative

We hypothesize that including cell type as a covariate will influence differential gene expression analysis results.

AIMS

AIM1

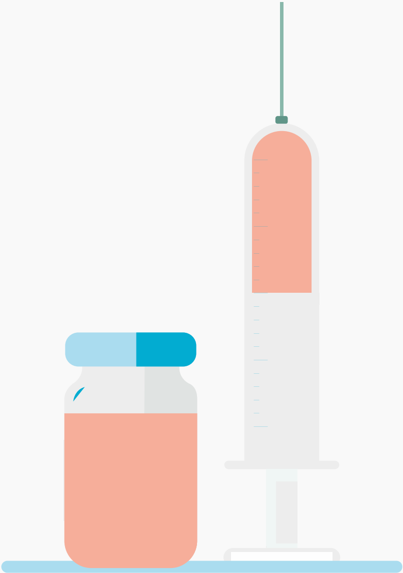
Perform cell type deconvolution to the dataset in Lieberman *et al.* paper.

AIM2

Perform a DE analysis based on infection status and using sex, age and cell type as covariates.

AIM3

Evaluate the statistical importance of including cell type. Compare our GSEA with the Lieberman *et al.* paper.



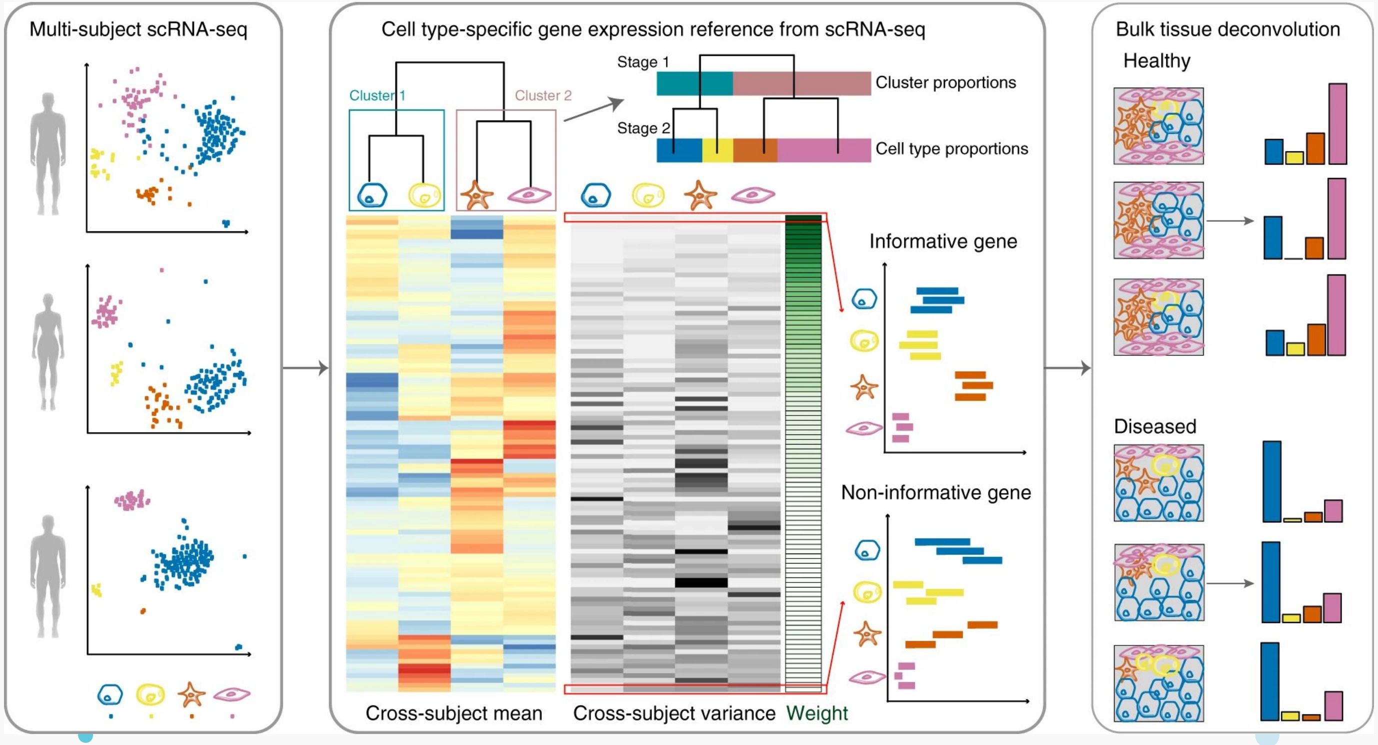
Aim 01

Cell type deconvolution

Multi-subject Single-cell Deconvolution (MuSiC)

scRNAseq reference

Bulk tissue data

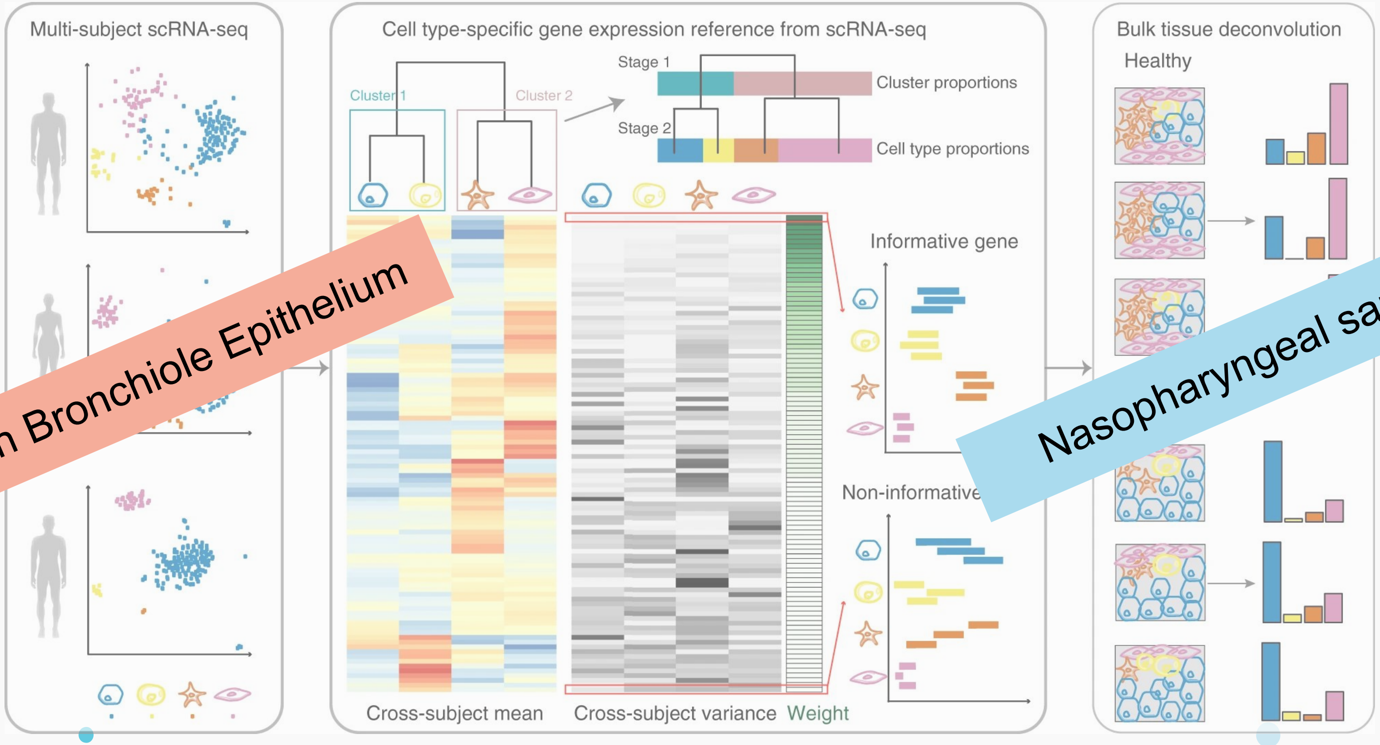


Nature Communications (Nat Commun) ISSN 2041-1723 (online)

Multi-subject Single-cell Deconvolution (MuSiC)

scRNAseq reference

Bulk tissue data



Human Bronchiole Epithelium

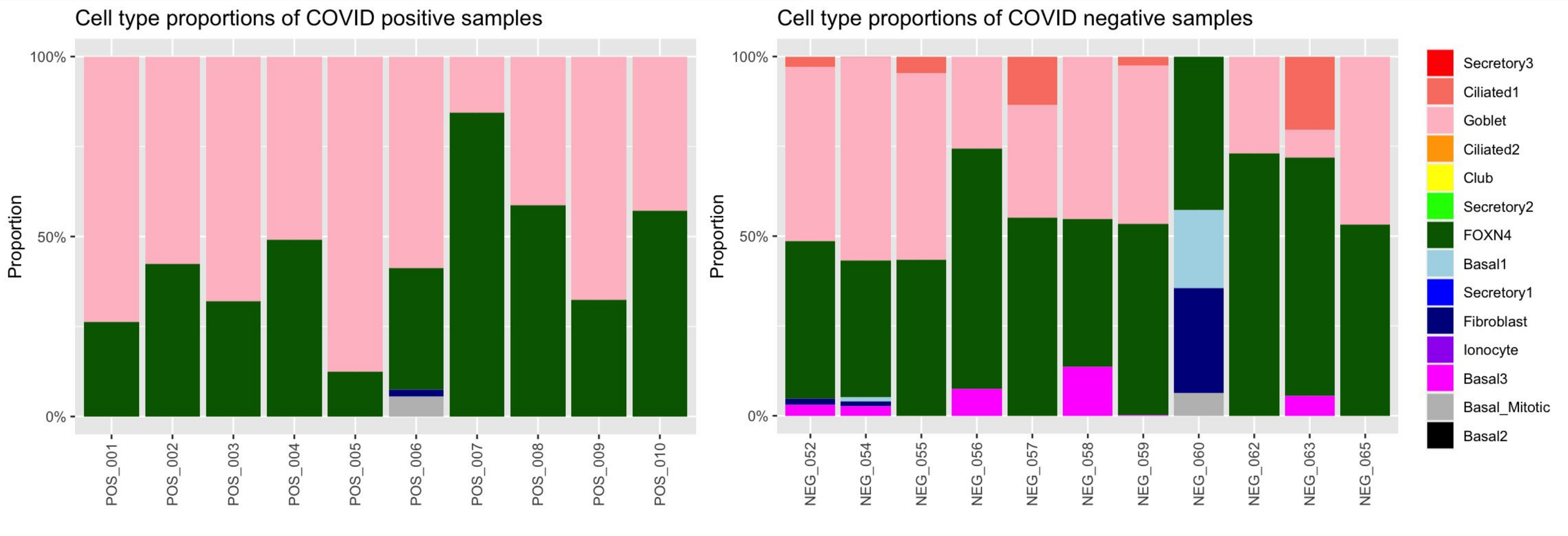
Nasopharyngeal samples

Nature Communications (Nat Commun) ISSN 2041-1723 (online)

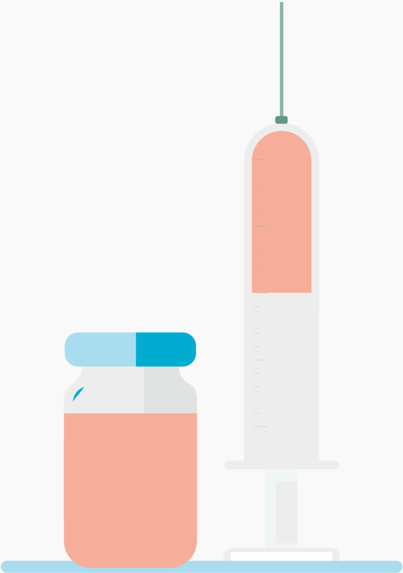
Overview of variables

Variables	Median or Count (range or %)
Age (years)	53 (2 - 89)
Gender (N females)	217 (53.06%)
Covid status (N positive)	356 (87.04%)
Cell type proportions (%)	Ciliated1: 0% (0-0.41%) Goblet: 55.98% (0-100%) FOXN4: 41.58% (0-100%) Basal1: 0% (0-25.72%) Fibroblast: 0% (0-29.38%) Basal3: 0% (0-29.16%)

COVID status is associated with cell type



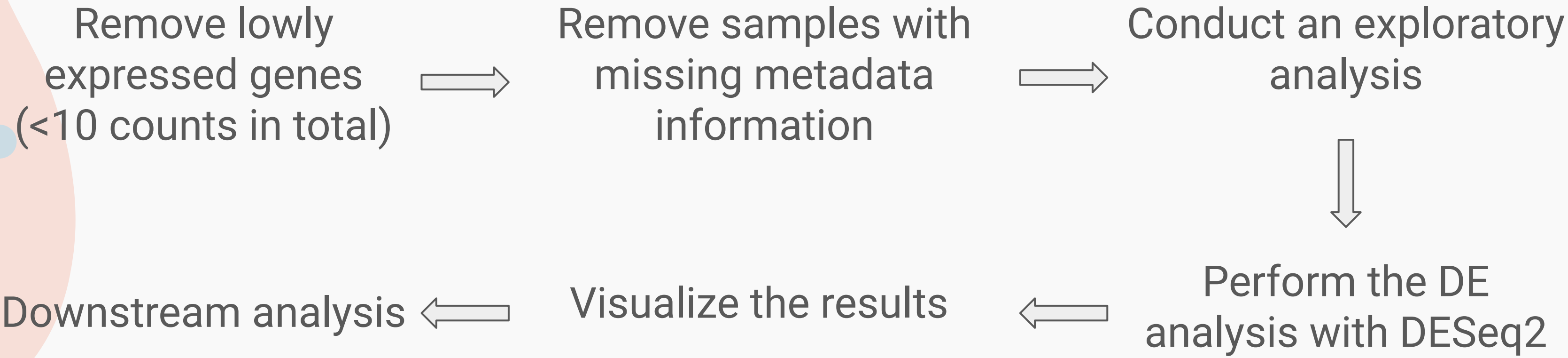
- Cell types with low proportions in the overall dataset were dropped



Aim 02

Differential gene expression

Workflow



Remove lowly expressed genes (<10 counts in total)

Remove samples with missing metadata information

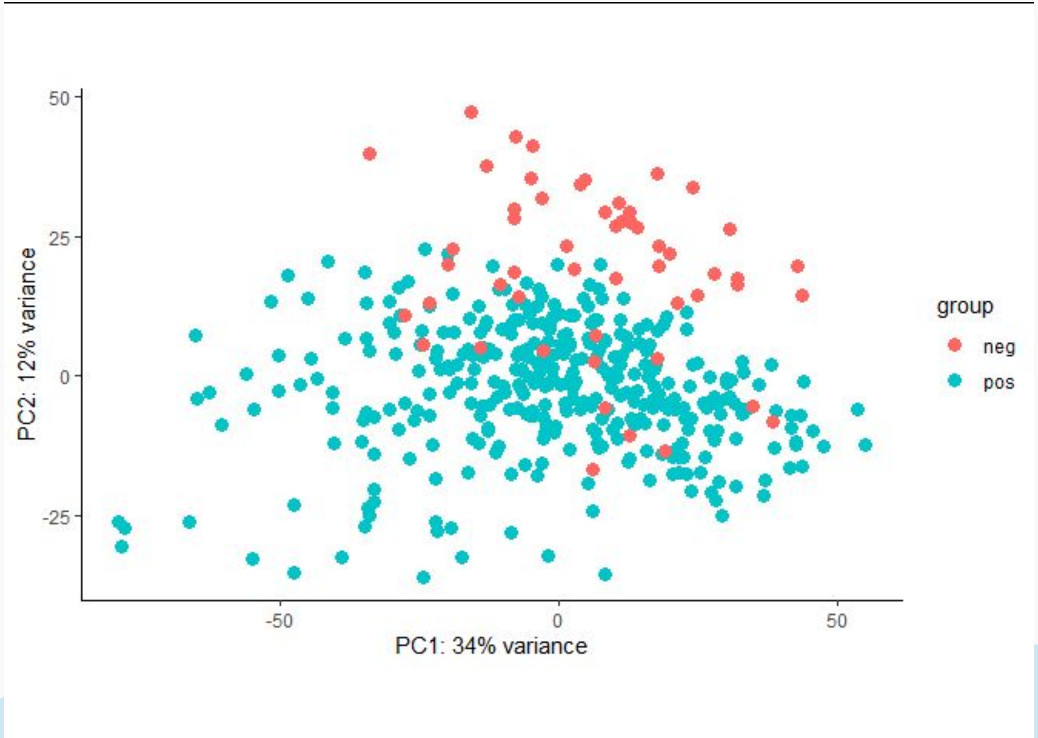
Conduct an exploratory analysis

Perform the DE analysis with DESeq2

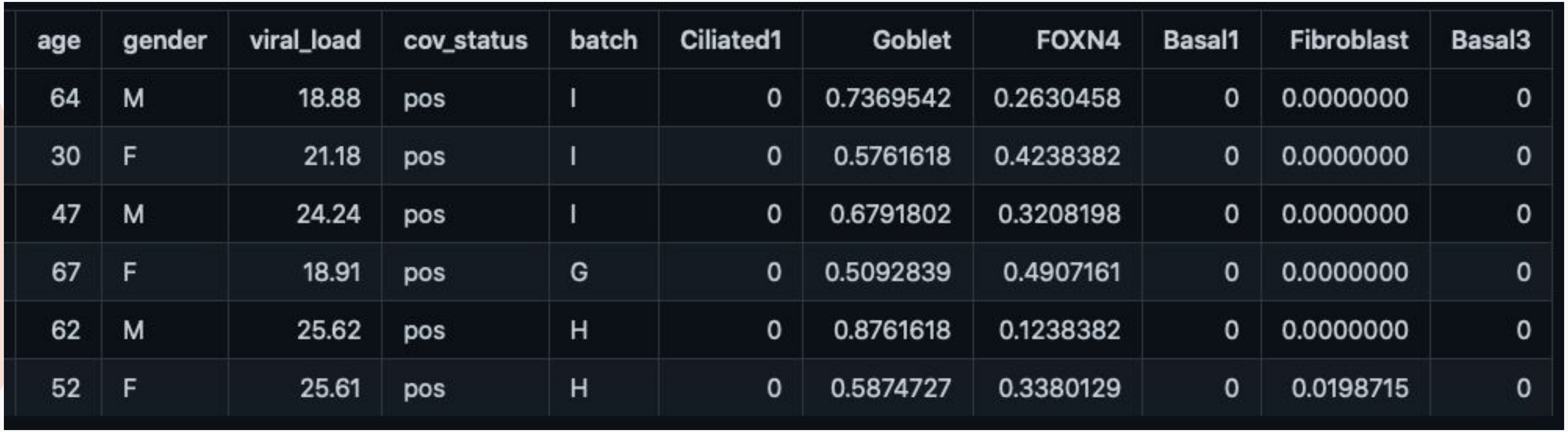
Visualize the results

Downstream analysis

Samples clustered based on their covid status in the PCA

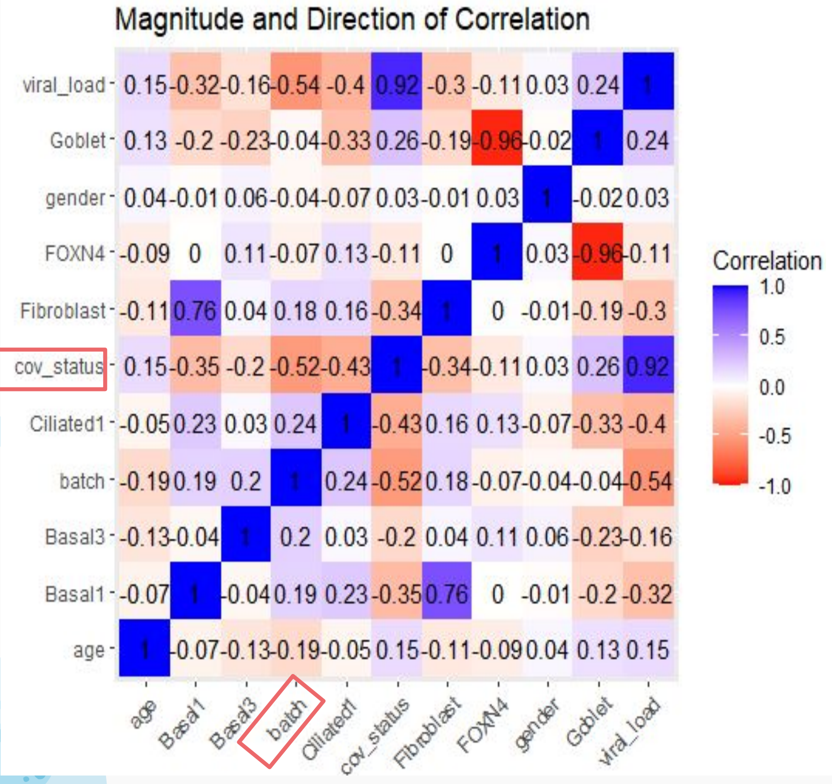


Model design

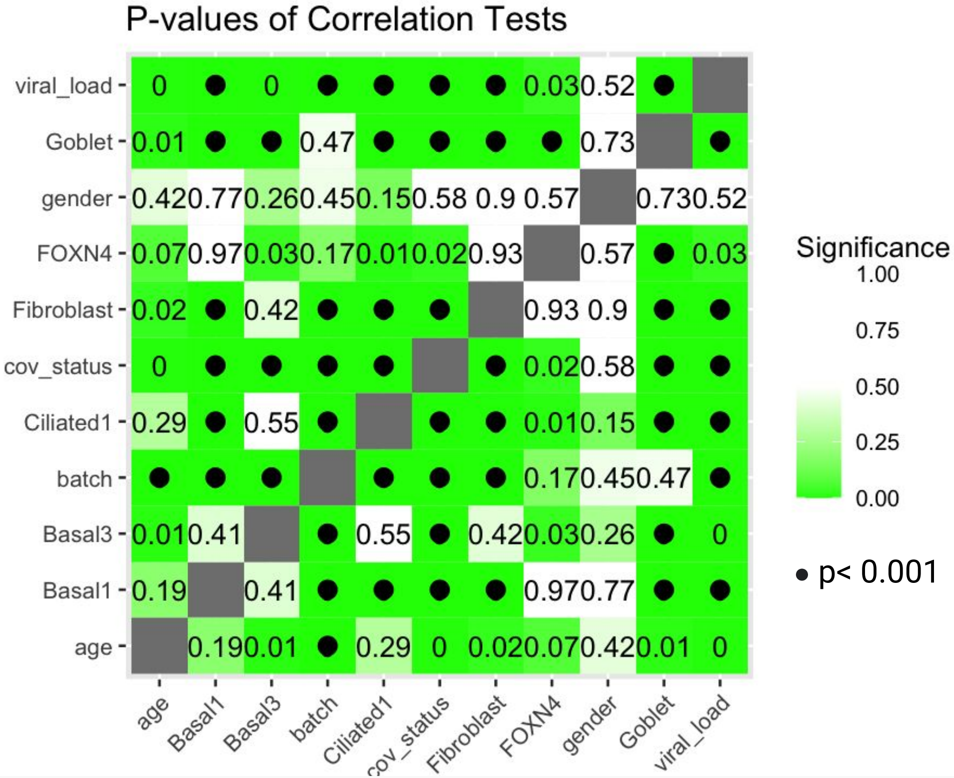


age	gender	viral_load	cov_status	batch	Ciliated1	Goblet	FOXN4	Basal1	Fibroblast	Basal3
64	M	18.88	pos	I	0	0.7369542	0.2630458	0	0.0000000	0
30	F	21.18	pos	I	0	0.5761618	0.4238382	0	0.0000000	0
47	M	24.24	pos	I	0	0.6791802	0.3208198	0	0.0000000	0
67	F	18.91	pos	G	0	0.5092839	0.4907161	0	0.0000000	0
62	M	25.62	pos	H	0	0.8761618	0.1238382	0	0.0000000	0
52	F	25.61	pos	H	0	0.5874727	0.3380129	0	0.0198715	0

High collinearity between covid status and batch



	age	Basal1	Basal3	batch	Ciliated1	cov_status	Fibroblast	FOXN4	gender	Goblet	viral_load
viral_load	0.15	-0.32	-0.16	-0.54	-0.4	0.92	-0.3	-0.11	0.03	0.24	1
Goblet	0.13	-0.2	-0.23	-0.04	-0.33	0.26	-0.19	-0.96	-0.02	1	0.24
gender	0.04	-0.01	0.06	-0.04	-0.07	0.03	-0.01	0.03	1	-0.02	0.03
FOXN4	-0.09	0	0.11	-0.07	0.13	-0.11	0	1	0.03	-0.96	-0.11
Fibroblast	-0.11	0.76	0.04	0.18	0.16	-0.34	1	0	-0.01	-0.19	-0.3
cov_status	0.15	-0.35	-0.2	-0.52	-0.43	1	-0.34	-0.11	0.03	0.26	0.92
Ciliated1	-0.05	0.23	0.03	0.24	1	-0.43	0.16	0.13	-0.07	-0.33	-0.4
batch	-0.19	0.19	0.2	1	0.24	-0.52	0.18	-0.07	-0.04	-0.04	-0.54
Basal3	-0.13	-0.04	1	0.2	0.03	-0.2	0.04	0.11	0.06	-0.23	-0.16
Basal1	-0.07	1	-0.04	0.19	0.23	-0.35	0.76	0	-0.01	-0.2	-0.32
age	1	-0.07	-0.13	-0.19	-0.05	0.15	-0.11	-0.09	0.04	0.13	0.15

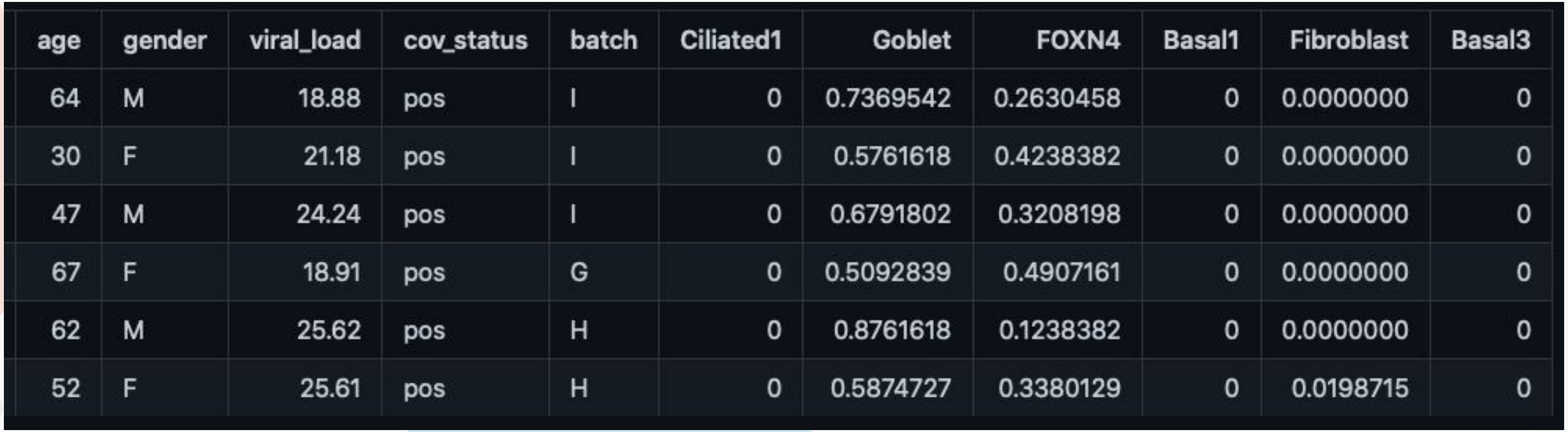


	age	Basal1	Basal3	batch	Ciliated1	cov_status	Fibroblast	FOXN4	gender	Goblet	viral_load
viral_load	0	●	0	●	●	●	●	0.03	0.52	●	
Goblet	0.01	●	●	0.47	●	●	●	●	0.73		●
gender	0.42	0.77	0.26	0.45	0.15	0.58	0.9	0.57		0.73	0.52
FOXN4	0.07	0.97	0.03	0.17	0.01	0.02	0.93		0.57	●	0.03
Fibroblast	0.02	●	0.42	●	●	●		0.93	0.9	●	●
cov_status	0	●	●	●	●		●	0.02	0.58	●	●
Ciliated1	0.29	●	0.55	●		●	●	0.01	0.15	●	●
batch	●	●	●		●	●	●	0.17	0.45	0.47	●
Basal3	0.01	0.41		●	0.55	●	0.42	0.03	0.26	●	0
Basal1	0.19		0.41	●	●	●	●	0.97	0.77	●	●
age		0.19	0.01	●	0.29	0	0.02	0.07	0.42	0.01	0

- **All COVID+ samples were run on the same batches, all COVID- on other batches**

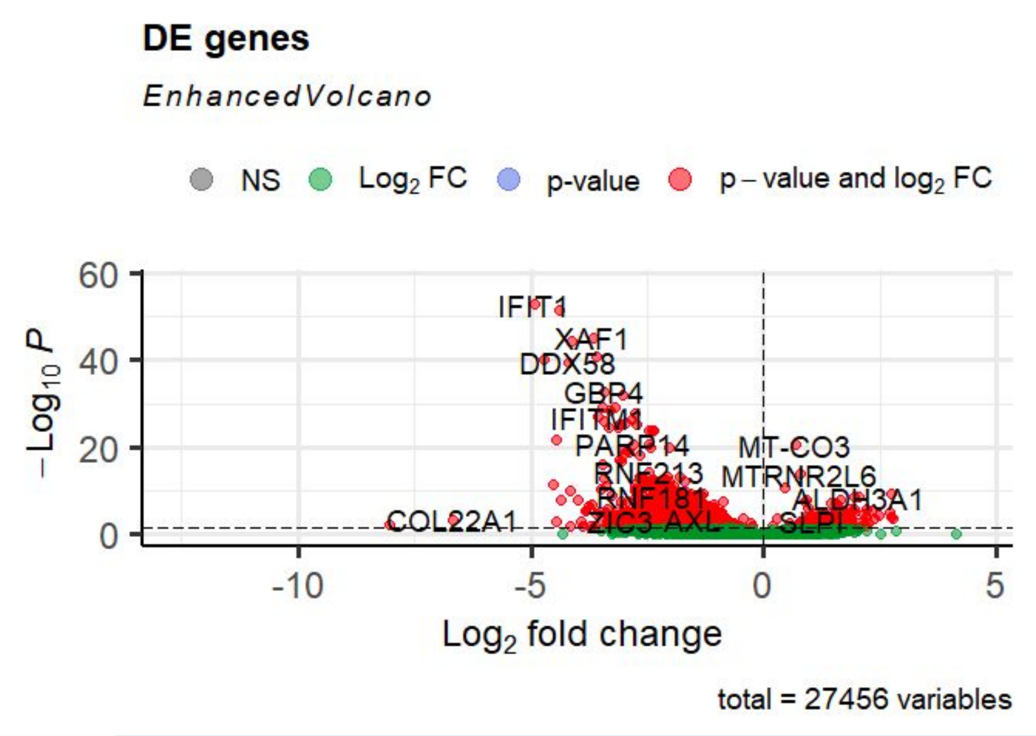
Model design

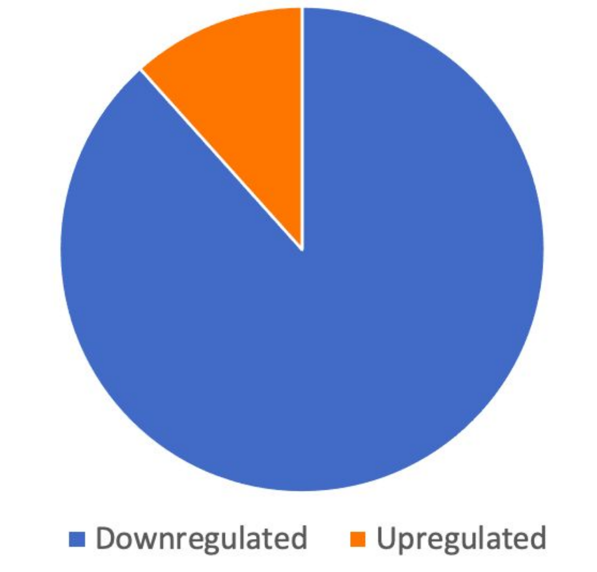
design = ~cov_status + age + gender + Ciliated1 + Goblet + FOXN4 + Basal1 + Fibroblast + Basal3



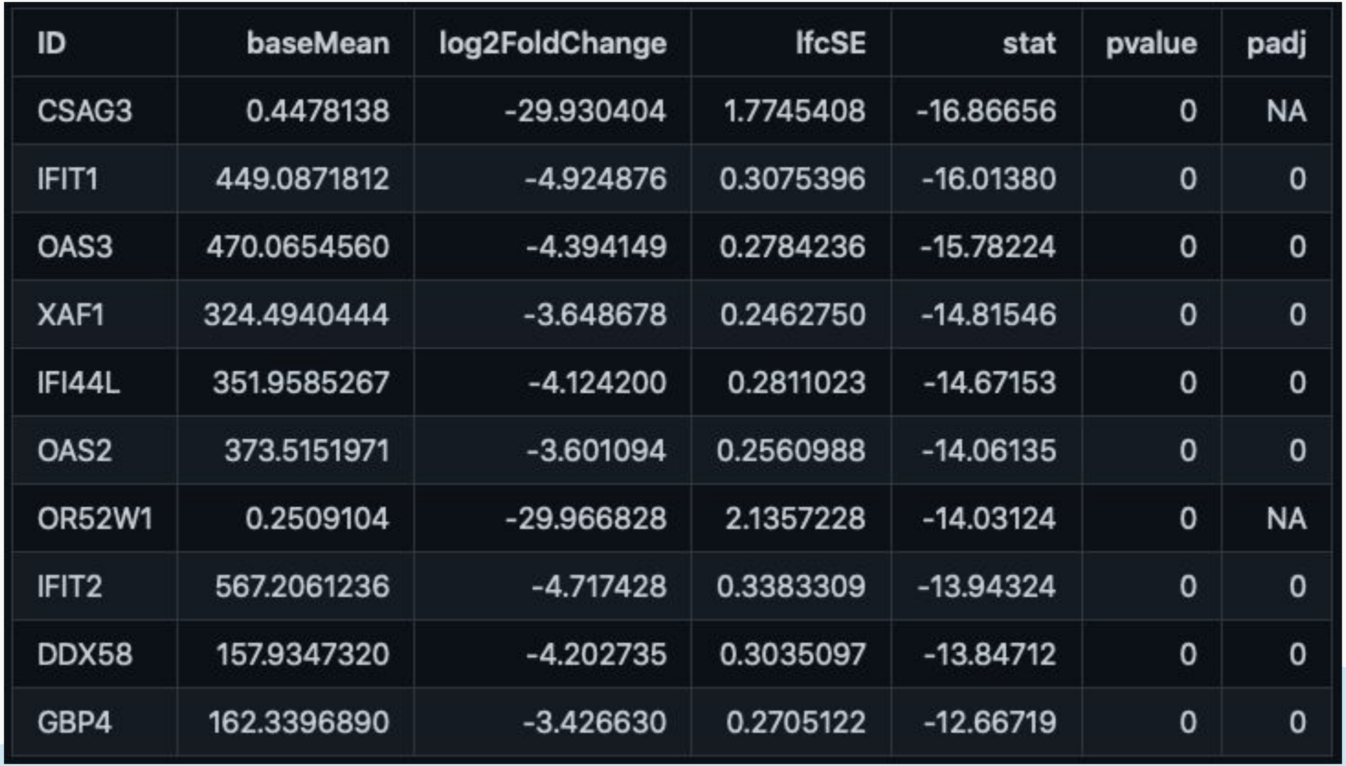
age	gender	viral_load	cov_status	batch	Ciliated1	Goblet	FOXN4	Basal1	Fibroblast	Basal3
64	M	18.88	pos	I	0	0.7369542	0.2630458	0	0.0000000	0
30	F	21.18	pos	I	0	0.5761618	0.4238382	0	0.0000000	0
47	M	24.24	pos	I	0	0.6791802	0.3208198	0	0.0000000	0
67	F	18.91	pos	G	0	0.5092839	0.4907161	0	0.0000000	0
62	M	25.62	pos	H	0	0.8761618	0.1238382	0	0.0000000	0
52	F	25.61	pos	H	0	0.5874727	0.3380129	0	0.0198715	0

Most genes were downregulated





Results



ID	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
CSAG3	0.4478138	-29.930404	1.7745408	-16.86656	0	NA
IFIT1	449.0871812	-4.924876	0.3075396	-16.01380	0	0
OAS3	470.0654560	-4.394149	0.2784236	-15.78224	0	0
XAF1	324.4940444	-3.648678	0.2462750	-14.81546	0	0
IFI44L	351.9585267	-4.124200	0.2811023	-14.67153	0	0
OAS2	373.5151971	-3.601094	0.2560988	-14.06135	0	0
OR52W1	0.2509104	-29.966828	2.1357228	-14.03124	0	NA
IFIT2	567.2061236	-4.717428	0.3383309	-13.94324	0	0
DDX58	157.9347320	-4.202735	0.3035097	-13.84712	0	0
GBP4	162.3396890	-3.426630	0.2705122	-12.66719	0	0

Gene Set Enrichment Analysis

1. **Hypergeometric enrichment**
   - MsigDB
2. **Gene set enrichment analysis (GSEA)**
   - MsigDB

Broad Molecular Signatures Database (MSigDB) gene sets

a. **Hallmark:** summarize and represent specific well-defined biological processes.

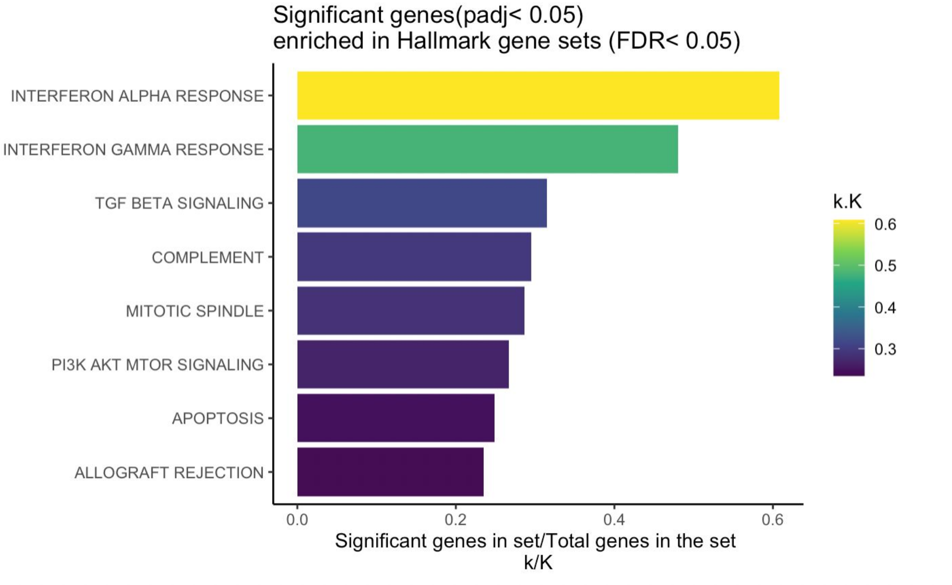
Total: 50

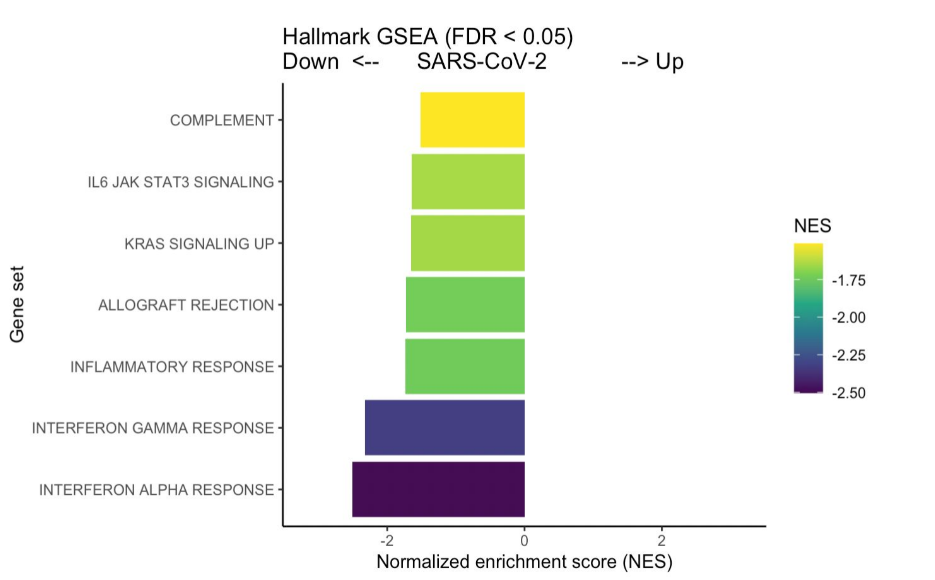
b. **Gene ontology [GO]**

- Biological process: molecular-level activities performed by gene products. **Total: 7481**

3. **Disease enrichment analysis (GSEA)**
   - Cluster profiler
   - DOSE: used for DO(Disease Ontology)

Hallmark Gene Set Results

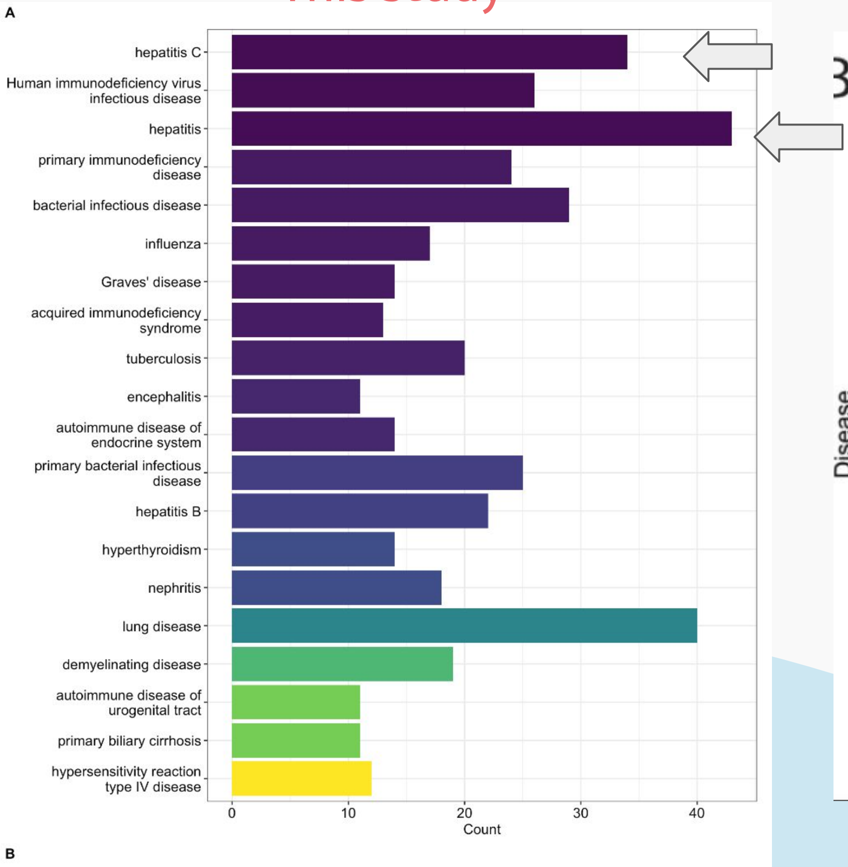




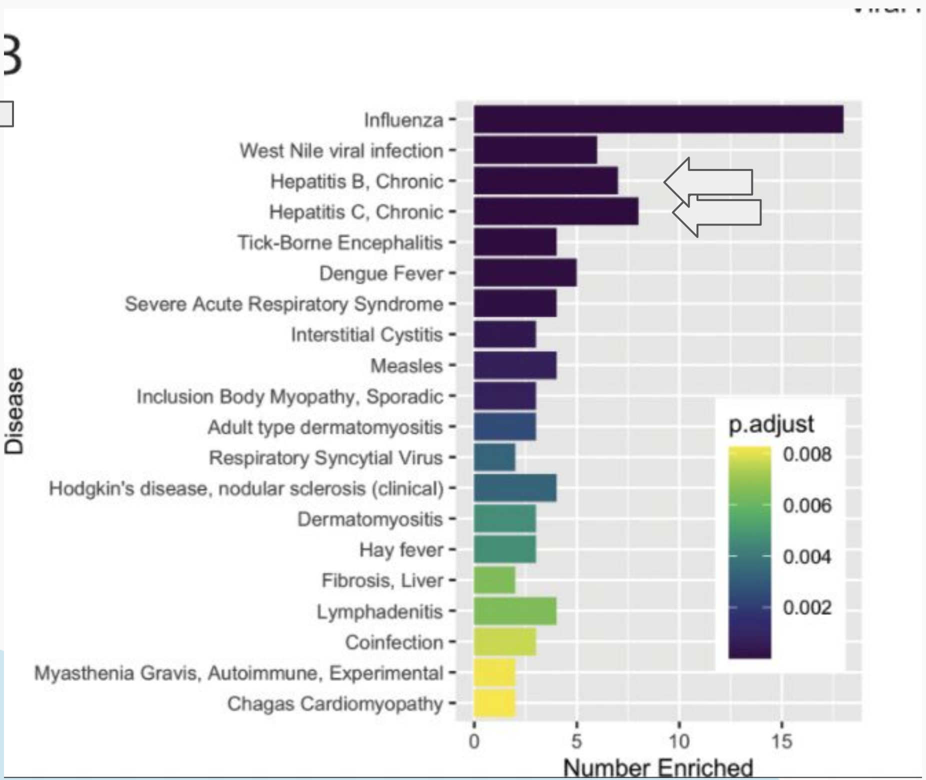
Enrichment of signature genes of Interferon Alpha, Interferon Gamma, and Inflammatory Responses

Disease Enrichment Analysis

This study

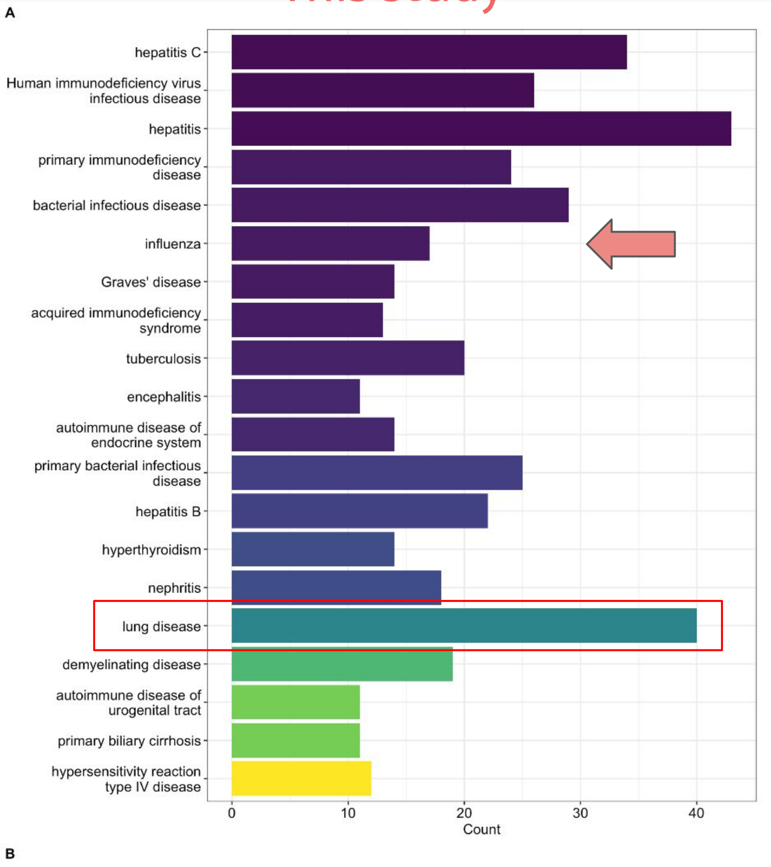


Liberman's findings

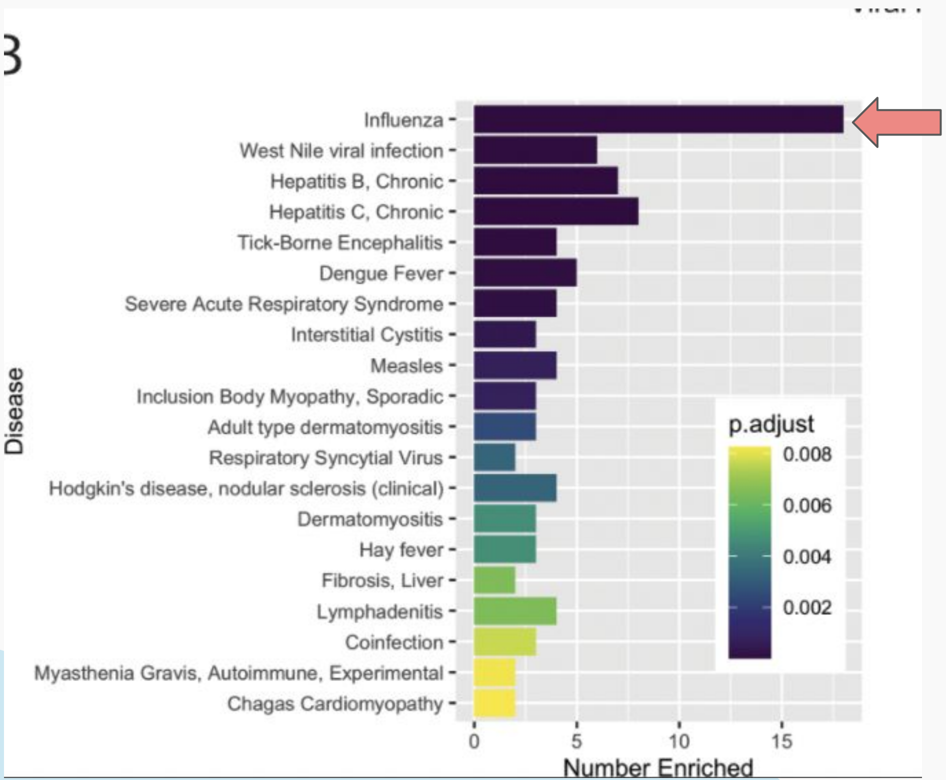


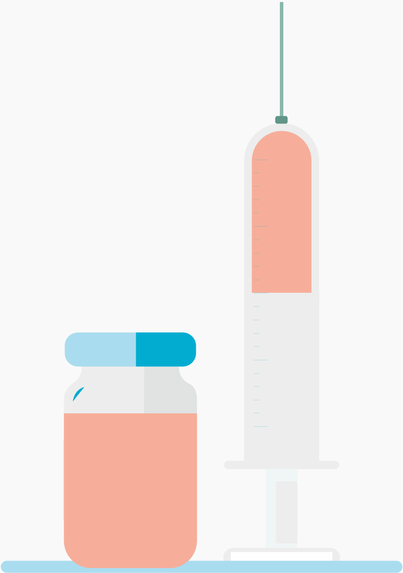
A high degree of overlap with lung disease signature

This study



Liberman’s findings



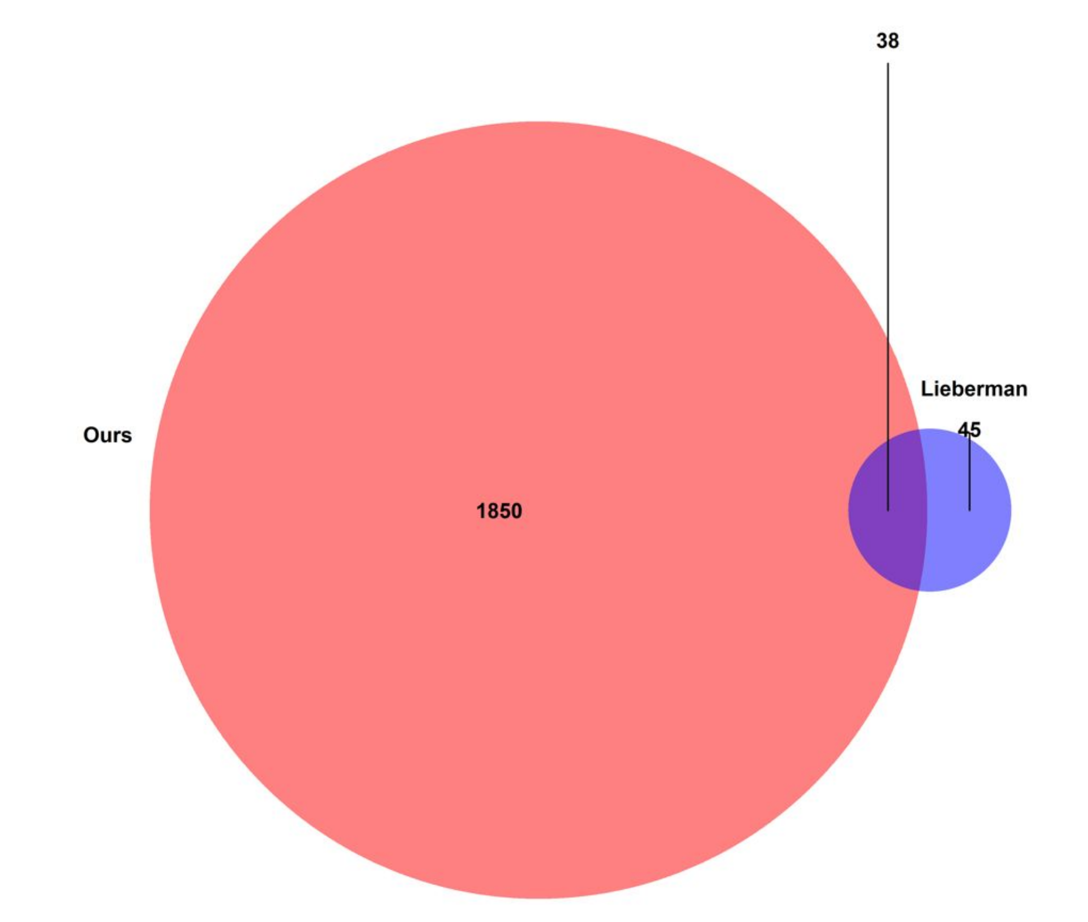


Aim 03

Quantitative comparison

1. **How many DE genes overlap between ours and theirs?**
2. **How many of our DE genes were better explained by the cell type model?**

How many overlapping DE genes?



Likelihood Ratio Test (LRT)

- Tests the following hypothesis

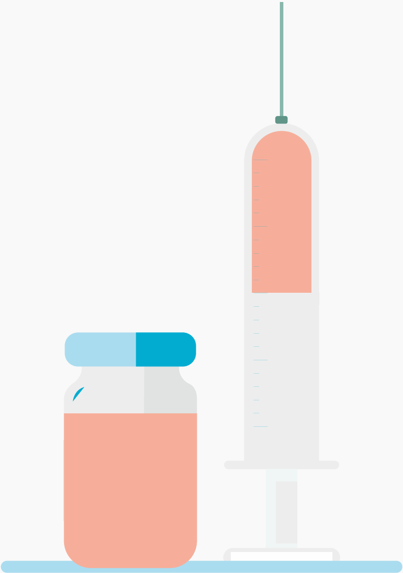
  H0: covid status + age + gender (reduced model)

  H1: covid status + age + gender + cell types (full model)

- Results
  - 7630 genes with significantly increased likelihood with full model
  - This is 27.7899% of all gene

How many of our DE genes were better explained by the cell type model?

- Results
  - Out of the 2960 our DE genes, we find 1888 with significantly increased likelihood with full model
  - This is 85.0636% of our findings
- Observation
  - Full model better characterizes significant DE genes



Conclusion

Limitations and Final remarks

Conclusion

- Limitations
  - Large correlation between batch and Covid status, could not reproduce author’s exact model
  - Reference data set is lower airways versus batch data set is upper airways
  - The experiment was not a balanced design (not possible to circumvent statistically)

- Final Remarks
  - Overall, cell type is a possible confounder associated with covid status and gene expression
  - Adding cell type corrects for possible confounding

References

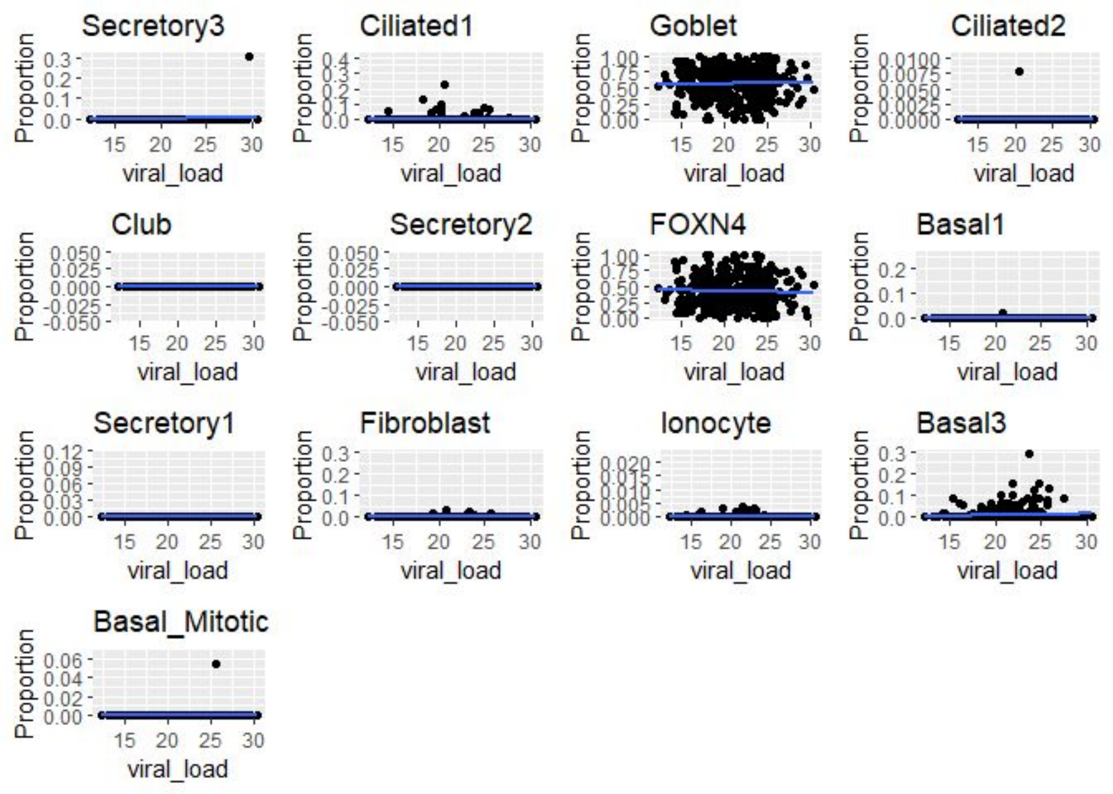
- Deprez, Marie, Laure-Emmanuelle Zaragosi, Marin Truchi, Christophe Becavin, Sandra Ruiz García, Marie-Jeanne Arguel, Magali Plaisant, et al. 2020. “A Single-Cell Atlas of the Human Healthy Airways.” *American Journal of Respiratory and Critical Care Medicine* 202 (12): 1636–45.
- Lieberman, Nicole AP, Vikas Peddu, Hong Xie, Lasata Shrestha, Meei-Li Huang, Megan C Mears, Maria N Cajimat, et al. 2020. “In Vivo Antiviral Host Transcriptional Response to SARS-CoV-2 by Viral Load, Sex, and Age.” *PLoS Biology* 18 (9): e3000849.
- Lukassen, Soeren, Robert Lorenz Chua, Timo Trefzer, Nicolas C Kahn, Marc A Schneider, Thomas Muley, Hauke Winter, et al. 2020. “SARS-CoV-2 Receptor ACE 2 and TMPRSS 2 Are Primarily Expressed in Bronchial Transient Secretory Cells.” *The EMBO Journal* 39 (10): e105114.
- Wang, Park, X. 2019. “Bulk Tissue Cell Type Deconvolution with Multi-Subject Single-Cell Expression Reference.” *Nat Commun* 10 (380).

Thank you!
Questions?

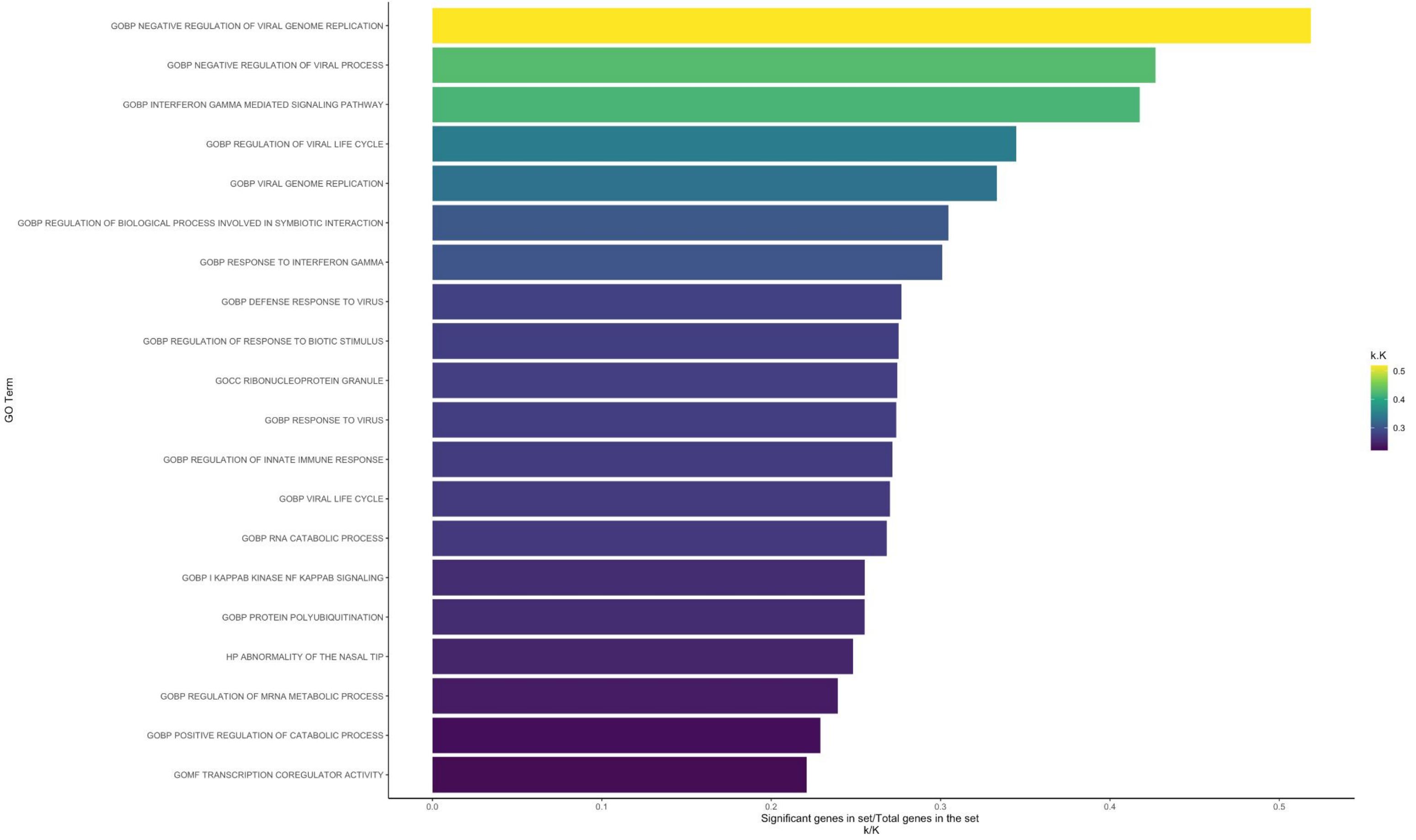
Extra slides

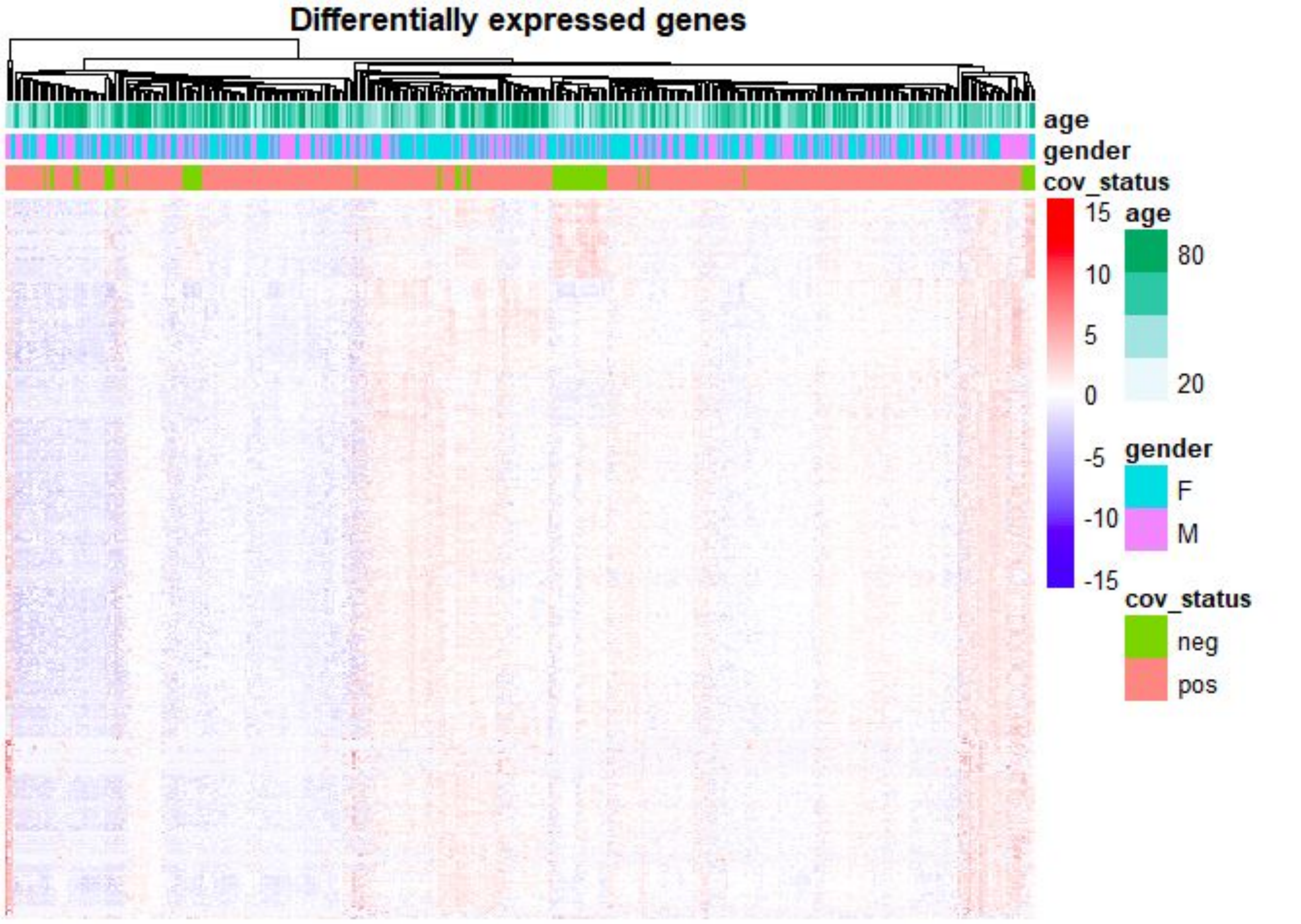
Figures that we made but are not part of the main presentation

Cell type by viral load



GO Terms


GOBP NEGATIVE REGULATION OF VIRAL GENOME REPLICATION
GOBP NEGATIVE REGULATION OF VIRAL PROCESS
GOBP INTERFERON GAMMA MEDIATED SIGNALING PATHWAY
GOBP REGULATION OF VIRAL LIFE CYCLE
GOBP VIRAL GENOME REPLICATION
GOBP REGULATION OF BIOLOGICAL PROCESS INVOLVED IN SYMBIOTIC INTERACTION
GOBP RESPONSE TO INTERFERON GAMMA
GOBP DEFENSE RESPONSE TO VIRUS
GOBP REGULATION OF RESPONSE TO BIOTIC STIMULUS
GOCC RIBONUCLEOPROTEIN GRANULE
GOBP RESPONSE TO VIRUS
GOBP REGULATION OF INNATE IMMUNE RESPONSE
GOBP VIRAL LIFE CYCLE
GOBP RNA CATABOLIC PROCESS
GOBP I KAPPAB KINASE NF KAPPAB SIGNALING
GOBP PROTEIN POLYUBIQUITINATION
HP ABNORMALITY OF THE NASAL TIP
GOBP REGULATION OF MRNA METABOLIC PROCESS
GOBP POSITIVE REGULATION OF CATABOLIC PROCESS
GOMF TRANSCRIPTION COREGULATOR ACTIVITY
GO Term
0.0
0.1
0.2
0.3
0.4
0.5
Significant genes in set/Total genes in the set
k/K
k.K
0.5
0.4
0.3


Differentially expressed genes
age
gender
cov_status
15
10
5
0
-5
-10
-15
age
80
20
gender
F
M
cov_status
neg
pos