**Analysis of Bronchoalveolar Lavage Fluid Metatranscriptomes Among Patients with COVID-19 Disease**

**Abstract**

**Introduction**

Metatranscriptomes from diseased host tissues represent a rich source of information to evaluate the role of the microbiome in disease onset and progression. Early in the SARS-CoV-2 outbreak, scientists openly published metatranscriptome sequences from Bronchoalveolar Lavage Fluid (BALF) of patients with COVID-19 disease; however, limitations in the sample numbers and lack of uniformity in study designs across different laboratories prevented a robust statistical analysis from taking place. In this paper, we evaluate what insights can be drawn from these valuable samples early in an outbreak scenario, as well as what questions are not able to be answered.

**Methods**

Supplemental Tables 1 and 2 describe the publicly available Illumina reads that were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) or the China National Center for Bioinformation (CNCB) National Genomics Data Center (NGDC), as well as the original publications where the clinical information was obtained for downstream analysis (1-8). Supplemental Table 3 lists the specific commands and additional details used for downstream analysis of the reads. After the raw reads were downloaded, the quality of the reads was assessed before and after trimming with FastQC (9) and quality control was performed on the downloaded sequence reads with Trimmomatic (10). To control different datasets being paired or single-end, all paired-end reads were converted to single-end by merging reads with flash (11), and then all merged and unmerged forward reads were combined into one file after being processed with Trimmomatic. Human and PhiX reads were filtered out with a custom Kraken2 database (12) and low complexity sequences were removed with fastp (13). Taxonomic analysis was subsequently performed with Kraken2 (12) and the decontam package was employed to identify potential contaminating organisms (14). The processed fastq datasets with human and PhiX reads removed were converted to fasta files and analyzed with SeqScreen (15) to obtain a list of leaf node molecular function and biological process Gene Ontology (GO) terms and proteins present within each of the samples. The CoV-IRT-Micro conda package (16) was used to propagate parent GO terms, parse GO terms by kingdom-level domains, and summarize Kraken2 taxonomic results and SeqScreen-reported protein identifiers.

Parent-propagated GO term counts for all domains other than eukaroytes were imported into a working phyloseq (17) object, alongside collected and curated clinical metadata using R 4.03 (18). Samples types of “unknown”, “sick”, and “negative controls” were pruned from subsequent analysis. Samples from Michalovich *et. al* (6) and samples from Shen et al. (5) that were viral enriched (PRJNA605907) were also pruned from subsequent analysis due to observed batch effects (Supplementary File 1a). GO term abundances from the remaining 86 samples were then compositionally transformed and compared by case type (min abundance=0.01, min prevalence=0.1 normalization=CLR, and outcome (COVID19 only) via Maaslin2 (19) (Supplementary File 1b), controlling for random effects of publication and sample name, max significance cutoff of q < 0.05 with Benjamini-Hochberg multiple test correction (20). Additionally, GO term counts were square root transformed and subjected to community typing with Dirichlet Multinomial Mixtures (21) (Supplementary File 1b). Statistically significant GO terms were then ordered by parental lineage and visualized alongside consensus DMM clusters and metadata columns publication, case, and outcome using the bioinformatic software package pheatmap (v1.0.12) (22).

**Results**

*Maaslin2 Comparison by case.* After controlling for random effects of publication and patient, Results from the Maaslin2 comparison across case types reveled 30 out of 13534 GO Terms associated with COVID19 when compared to community acquired pneumonia and uninfected patients (Table X) (Figure X). Significant GO terms were comprised of 6 Depth 1 Parents involving catalytic activity [GO:0003824], binding, metabolic [GO:0008152] and cellular processes [GO:0009987], biological regulation [GO:0065007], and interspecies interaction between organisms [GO:0044419]. Significant Terms of interest associated with COVID19 include hydrolase [GO:0016787] / transferase [GO:0016740] activity transferring phosphorus [GO:0016772], nucleotidyltransferase activity [GO:0016779], and ion binding [GO:0043167].

*Maaslin2 comparison by outcome.* An analysis of disease outcome amongst COVID-19 positive patients via Maaslin2 revealed XX GO Terms associated with deceased outcome (Table X), with notable functional profiles associated phosphate / phosphorylation [GO:0016310], metal ion binding (mg,zn,etc) [GO:0046914;GO:0000287;GO:0008270], RNA binding [GO:0003723], and lytic activity (hydrolase, endopeptidase, oxidoreductase, etc) [GO:0016491;GO:0016817; GO:0140098]. (Figure X).

*DMM clustering.* Results from the Dirichlet Multinomial Mixtures clustering analysis using all 13,534 Gene ontologies counts resulted in a best model fit using 3 distinct clusters that were significantly associated with each case type p<0.0001 (Figure X, Supplementary Table X).

*Taxonomic Analysis.* Taxonomic analysis revealed a statistically significant decrease in log2 median ration of several species belonging to the genus *Sphingomonas* when compared to both the uninfected (p<0.0001, q <0.001) and CAP cohorts (p<0.005,q <0.05) cohorts (Table X). This finding supports previous reports regarding an association with *Sphingomonas* **[CITE ME]**, which is commonly known as an opportunistic pathogen found in healthcare-associated pneumonia.

Analysis of the GO Terms derived from *Sphingomonas* proteins in the COVID19 samples were

hydrogen peroxide catabolic process [GO:0042744]; response to oxidative stress [GO:0006979]

catalase activity [GO:0004096]; heme binding [GO:0020037]; and metal ion binding [GO:0046872].

**Discussion**

* The findings from the disease outcome analysis were similar in nature to the GO Terms associated with COVID19 versus the uninfected and community acquired pneumonia patient cohorts.
* Sphingomonas
  + The catalase protein decomposes hydrogen peroxide into water and oxygen; serves to protect cells from the toxic effects of hydrogen peroxide.
  + So as a discussion point perhaps Sphingomonas is responding to COVID-19 conditions in the patient by expressing genes that help it to survive well under the COVID-19 disease conditions.
* *What are these go terms telling us*
* *Who else has found similar stuff*
* *What are these taxa telling us*
* *Who else has found similar stuff*
* *Whats next*

**Conclusion**

In conclusion, we observed unique and taxonomic and functional discriminant features in the brochoalveolar lavage metatranscriptomes associated with COVID19 disease and death.  Taxa of interested included genera from the Sphingomonadacae Class, and function annotated Gene ontologies of interest included associated with:Phosphate / phosphorylation, metal ion binding (mg,zn,etc), nucleotide terms (DNA/RNA), Lytic activity (hydrolase, endopeptidase,etc). Collectively, while this data does cannot speak to causality or directionality of the association, it does demonstrate a significant relationship between the human microbiome and severity of COVID-19, rendering further testable hypotheses that warrant further investigation.

**Acknowledgments**

We would like to acknowledge the COVIRT microbial subgroup team members and give special acknowledgment to John Fonner and the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported.

**References**

1. Chen L, Liu W, Zhang Q, Xu K, Ye G, Wu W, et al. RNA based mNGS approach identifies a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak. Emerg Microbes Infect. 2020;9: 313–319.

2. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579: 265–269.

3. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579: 270–273.

4. Xiong Y, Liu Y, Cao L, Wang D, Guo M, Jiang A, et al. Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in COVID-19 patients. Emerg Microbes Infect. 2020;9: 761–770.

5. Shen Z, Xiao Y, Kang L, Ma W, Shi L, Zhang L, et al. Genomic Diversity of Severe Acute Respiratory Syndrome–Coronavirus 2 in Patients with Coronavirus Disease 2019. Clinical Infectious Diseases. 2020. doi:10.1093/cid/ciaa203

6. Michalovich D, Rodriguez-Perez N, Smolinska S, Pirozynski M, Mayhew D, Uddin S, et al. Obesity and disease severity magnify disturbed microbiome-immune interactions in asthma patients. Nat Commun. 2019;10: 5711.

7. Huang W, Yin C, Wang G, Rosenblum J, Krishnan S, Dimitrova N, et al. Optimizing a Metatranscriptomic Next-Generation Sequencing Protocol for Bronchoalveolar Lavage Diagnostics. J Mol Diagn. 2019;21: 251–261.

8. Ren L, Zhang R, Rao J, Xiao Y, Zhang Z, Yang B, et al. Transcriptionally Active Lung Microbiome and Its Association with Bacterial Biomass and Host Inflammatory Status. mSystems. 2018;3. doi:10.1128/mSystems.00199-18

9. FastQC

10. Trimmomatic

11. flash

12. Kraken2

13. fastp

14. decontam

15. SeqScreen

16. CoV-IRT-Micro conda package

17. McMurdie PJ, Holmes S: **phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data**. *PLOS ONE* 2013, **8**(4):e61217.

18. R 4.03

19. Maaslin2

20. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful**

**approach to multiple testing**. *Journal of the Royal statistical society: series B (Methodological)*

1995, **57**(1):289-300.

21. Holmes I, Harris K, Quince C: **Dirichlet Multinomial Mixtures: Generative Models for**

**Microbial Metagenomics**. *PLOS ONE* 2012, **7**(2):e30126.

22. Kolde R: **Pheatmap: pretty heatmaps**. *R package version* 2012, **1**(2).