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

**Abstract**

**Introduction**

In order to better understand the potential relationship between COVID-19 moderate to severe disease severity and the microbial community dynamics / functional profiles from a hologenome standpoint, we conducted an analysis using human bronchoalveolar lavage fluid (BALF) metatranscriptomes sample sequences sourced from 8 different publications that were made available from public repositories.

These samples comprise 3 main cohorts case types consisting of uninfected control cohort, community acquired pneumonia or CAP patients, or COVID19 patients, with a secondary analysis of disease severity amongst a subset of the COVID19 cohort broken down by survival outcome. The objectives of the study are to compare the BALF metatranscriptomes in the COVID19 cohort amongst uninfected and CAP patient cohorts and identify taxonomic changes in microbial derived community dynamics / and functional changes derived from gene ontologies associated with COVID19 moderate to severe disease or its treatment and identify predictors of disease outcome amongst COVID19 based on metatranscriptomes profiling, with the overarching hypothesis that there is a potential informative relationship between the BALF microbiome and the severity of COVID-19 disease onset, progression, and outcome.

**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). 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 were pruned from subsequent analysis. Taxonomic classifications were decontaminated against negative controls when negative congrols were present using the library decontam to identify and remove potential contaminating organisms (14). 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). After read filtering and batch effect sample removal, sample cohorts consisted of 29 uninfected sampled, 25 CAP samples, and 32 COVID19 samples, bringing the total *n* to 86. Amongst the COVID19 cohort with known survival outcomes, 10 were deceased and 15 were survived. GO term abundances from the remaining 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 packages pheatmap (v1.0.12) (22). Heat tress taxonomic comparisons were visualized using the bioinformatic software packages and metacoder (<https://cran.r-project.org/web/packages/metacoder/citation.html#:~:text=metacoder%20citation%20info,doi%3A%2010.1371%2Fjournal>) .

**Results**

*Case type comparisons.*

After controlling for random effects of publication and patient, Results from the Maaslin2 comparison across case types reveled 35 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]. 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 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).

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].

*Stratification based on survival.*

A stratified analysis amongst the COVID19 samples with known survival outcomes via Maaslin2 which revealed 25 significant GO Term when comparing the deceased to the survived cohorts, with Depth 1 parents Cellular and Metabolic processes [XXXXXXXXXXXX], Catalytic Activity [XXXXXXXXXX], and Binding [XXXXXXXXXX].

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).

Features of particular interest associated with morbidity in the Depth 1 parent Metabolic and cellular processes include decreases in carbohydrate metabolic processes, Increases in RNA metabolic processes and RNA phosphodiester bond hydrolysis, decreases in phosphorylation, and increases in nucleobase containing compound biosynthetic processes.

Gene ontology comparisons amongst those who died to those who survived reported major decreases in oxoreductase activity, increases in catalytic activity acting on RNA, and endopeptidase activity. Lastly with respect to the Go Terms belonging to the Depth 1 parent of binding, we observed decreases in organic cyclic compound binding and increases in RNA binding transition metal ion binding, magnesium ion binding, and zinc ion binding amongst those who died compared to those who survived.

Results of an Analysis of Variance of a predictive model of outcome based and clusters derived from the unsupervised machine learning Dirichlet mixture modeling clustering analysis revealed a statistically significant associated between dmm\_cluster number and outcome with a p value of <0.001 and an Adjusted R-squared value of 0.3238. A posthoc Tukey- Kramer multiple comparison of means test with 95% confidence intervals showed a statistically significance amongst samples that deceased compared with samples that survived with an adjusted p value of 0.001.

Results from the Taxonomic comparison analysis revealed a statistically significant increase in log2 median ratio of the family *Comanomonadacea*, belonging to the genus *Variovorax* and decreases in the family *Bacteriodales* when comparing the deceased to the survive (p<0.0001, q <0.001) (table X).

**Discussion**

* *What are these go terms / what are they telling us*
* *Who else has found similar stuff*
* *What are these taxa telling us*

*Who else has found similar stuff*

*Functional*

We observed unique and taxonomic and functional discriminant features in the brochoalveolar lavage metatranscriptomes distinctive of COVID19 moderate to serve disease or its treatment, and predictors of COVID19 mortality. Functionally 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). 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.

*Taxonomic comparisons*

Distinct Taxonomic features of COVID19 disease and mortality include increases in log2 median ratios of genera *Sphingomonas* and *Variovorax* belonging to the *Sphingomonadacae* and *Comonomonadacea* families, and decreases in the class *Bacteroidia* belonging to the order *Bacteroidiales*. These finding support previous reports regarding an association with *Sphingomonas* **[CITE ME]**, which is commonly known as an opportunistic pathogen found in healthcare-associated pneumonia.

**[Define and tie the taxa to the GO terms throught the use of the Uniprot things here]**

Proteins derived from *Sphingomonas* contributed to the Sig. GO terms of interest 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] amongst the COVID-19 cohort. The catalase protein decomposes hydrogen peroxide into water and oxygen; serves to protect cells from the toxic effects of hydrogen peroxide, which may suggest that *Sphingomonas* spp. responds to COVID-19 conditions in the patient by expressing genes that help it to survive well in evironments undergoing great amounts of oxidative stress.

[PUT OTHER STUFF ABOUT VARIOVORAX and BACTEROIDIA HERE]

[DEFINED GO TERMS]

* Nucleobase containing compound biosynthetic process is defined as: “The chemical reactions and pathways resulting in the formation of nucleobases, nucleosides, nucleotides and nucleic acids.”
* RNA phosphodiester bond hydrolysis exonucleolytic is defined as “The chemical reactions and pathways involving the hydrolysis of terminal 3',5'-phosphodiester bonds in one or two strands of ribonucleotides.”
* Catalytic activity acting on RNA is defined as Catalytic activity that acts to modify RNA., Oxidoreductase activity is defined as “*Catalysis of an oxidation-reduction or redox reaction, a reversible chemical reaction in which the oxidation state of an atom or atoms within a molecule is altered via one substrate acting as an electron donor and becoming oxidized, while the other acts as an electron acceptor and becomes reduced.”*
* The Terminal GO Term endopeptidase activity is defined as “*Catalysis of the hydrolysis of internal, alpha-peptide bonds in a polypeptide chain.”*
* RNA binding is defined as *Interacting selectively and non-covalently with an RNA molecule or a portion thereof.* *Ion binding is defined as Interacting selectively and non-covalently with magnesium (Mg) and (Zn) ions.*

[MIKE LEE ALSO SUGGESSTED MAYBE REDO THE ANALAYSIS AND DON’T JUST LOOK AT SIG. TAXA TO COVID19 VS CAP & UNINFECTED]

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 thank 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).