**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**

* Raw sequencing reads were download from either SRA or CRA followed by
* Read preprocessing consisted of adapter trimming,QA/QC, and filtering of human and low complexity reads using Trimmomatic, fastQC, Kraken2, FastP, followed by
* Taxonomic classification via Kraken2 and functional annotation via Seqscreen
* Taxonomic classifications were decontaminated against negative controls when applicable using the library decontam in R, followed by statistical analysis and visualization using the bioinformatic software package phyloseq , vegan, and metacoder in R.
* Functional annotations gene ontologies counts derived from seqscreen outputs were parent propagated using covirt-micro conda package generated by Mike Lee at NASA.
* Parent propagated Gene ontology terms counts were then subjected to a Dirichlet mixture modeling (or Dmm clustering) and comparative analysis using multivariable associations with linear models Maaslin2 in R.
* All Scripts, intermediate files, and results can be found in the OSF, github and covert-micro githubs

Parent propagated Geno ontology abundances were imported into a working phyloseq [1]hylo object alongside collected and curated metadata using R 4.03. Samples case types “unknown”, “Sick”, and negative controls were pruned from subsequent analysis. Samples from Michalovich et. al and samples 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 (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[2]. Additionally, Geno ontology counts were square root transformed and subjected to community typing with Dirichlet Multinomial Mixtures [3] (Supplementary File 1b). Statistically significant GO terms were then ordered by parents, and visualized (I THINK MIKE DID SOME OTHER STUFF NORMALIZAITON STUFF HERE) alongside consensus DMM clusters and metadata columns publication, case, and outcome using the bioinformatic software package pheatmap (v1.0.12) [4].

30 out of a total of 13,534 Gene ontologies were associated with COVID19 when compared to community acquired pneumonia and uninfected patients using a max significance cutoff of q < 0.05 with Benjamini-Hochberg multiple test correction.

Significant Gene ontologies are comprise of 6 Depth 1 Parents involving catalytic activity, binding, metaboloic and cellular process,biological regulation, and interspecies interaction between organisms.

Terms of interest associated with COVID19 include hydrolase/transferase activity transferring phosphorus, and nucleotidyltransferase activity, and ion binding

comprise of 6 Depth 1 Parents involving catalytic activity, binding, metaboloic and cellular process,biological regulation, and interspecies interaction between organisms.

Significant Gene ontologies of interest associated with COVID19 **include hydrolase activity**

**transferase activity**

**transferase activity, transferring phosphorus**

**nucleotidyltransferase activity**

**ion binding**

**hydrolase, Metal ion binding (mg,zn,etc), Nucleotide and binding terms, and Lytic activity. [Next slide]**

Additionally, Geno ontology counts were square root transformed and subjected to community typing with Dirichlet Multinomial Mixtures which determined 3 Distinct clusters that were significantly associated with each case type p<0.0001\*\*\*

We then conducted a subset Maaslin2 analysis amongst COVID19 samples with known survival outcomes analysis \*\*\* revealing GO Terms associated **Phosphate / phosphorylation, Metal ion binding (mg,zn,etc) Nucleotide terms (DNA/RNA) Lytic activity (hydrolase, endopeptidase,etc)**

Maaslin2 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 controlling for random effects of publication and sample name, max significance cutoff of q < 0.05 with Benjamini-Hochberg multiple test correction[2].

Additionally, Geno ontology counts were square root transformed and subjected to community typing with Dirichlet Multinomial Mixtures which determined 3 Distinct clusters that were significantly associated with each case type p<0.0001\*\*\*

*Maaslin2 comparison by outcome.* An analysis of outcome amongst COVID19 positive patients via Maaslin2 revealed XX GO Terms associated with deceased outcome (Table X) (Figure X).

*DMM clustering.*

DMM modeling revealed statistically significant clustering of community types associated with case type and outcome (p<XXX).

*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*, which is commonly known as an opportunistic pathogen found in healthcare-associated pneumonia.

. Amongst COVID19 patientst, species XXX,XXX,XXX were associated with the deceased outcome cohort.

**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)**

**Discussion**

*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**

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

2. 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.

3. Holmes I, Harris K, Quince C: **Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics**. *PLOS ONE* 2012, **7**(2):e30126.

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