**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**

* *Download Raw reads (SRA,CRA)*
  + *Who all did we download*
  + *Where did we get them*
  + *Medical clinical stats analysis.*
* *Trim reads (Trimmomatic)*
* *Filter human reads (Kraken2)*
* *Filter low complexity (FastP)*
* *Taxonomic Analysis (Kraken2)*
  + *Decontam*
  + *Dmm clustering*
* *Gene Ontology (Seqscreen) (Kristen)*
  + *Parent propagation (bit) (Mike)*
  + *Comparative Analyses*
* 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].

**Results**

*Maaslin2 Comparison by case.* Results from the Maaslin2 comparison across case types reveled XX GO Terms associated with COVID19 when compared to community acquired pneumonia and uninfected patients (Table X) (Figure X).

*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 the presence of XXX, XXX,XXX in COVID19 patients. Amongst COVID19 patientst, species XXX,XXX,XXX were associated with the deceased outcome cohort.

**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).