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

**Abstract**

A total of 48 bronchoalveolar lavage fluid (BALF) metatranscriptomes from 18 COVID-19 positive patients in five studies were compared to 114 BALF metatranscriptomes from 114 individuals in four studies, as well as peripheral blood mononuclear cell (PBMC) metatranscriptomes from three healthy and three COVID-19 positive individuals (https://osf.io/7nrd3/wiki/home/). For the samples from each study, considerations were made for the primary research question, sample type and storage method, lysis protocol, type of sequencing library preparation, sequencing instrument, enrichment method (if any), and negative control data (if any). After removal of RNA reads from the human host and contaminating organisms, the remaining SARS-CoV-2 viral and microbiome RNA reads were analyzed according to their taxonomic classifications and functional characterizations. In addition to detecting taxonomic shifts that may be driven by COVID-19 microbiome dysbiosis or drugs being taken by COVID-19 patients, notable functional profiles were detected within microbial communities through the analysis of predicted proteins and gene ontology terms assigned to sequence reads. This study has demonstrated that imperfect datasets can contribute to knowledge about how the human microbiome influences susceptibility to and severity of COVID-19 and helped to generate important hypotheses that warrant further investigation.

**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,)
* Trim reads (Trimmomatic)
* Filter human reads (Kraken2)
* Filter low complexity (FastP)
* Taxonomic Analysis (Kraken2)
  + Decontam
  + Dmm clustering
* Gene Ontology (Seqscreen)
  + Parent propagation (bit)
* Analysis

Parent propagated Geno ontology count tables were loaded into R 4.03 and converted into a working phyloseq object alongside collected and curated metadata . 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[1]. Additionally, Geno ontology counts were square root transformed and subjected to community typing with Dirichlet Multinomial Mixtures [2] (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) [3].

* 1. 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.
* 2. Holmes I, Harris K, Quince C: **Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics**. *PLOS ONE* 2012, **7**(2):e30126.
* 3. Kolde R: **Pheatmap: pretty heatmaps**. *R package version* 2012, **1**(2).