SAND2020-10296

Title: COVID-19 biomarkers based on respiratory microbiome content.

Team: Steve Branda (PI), Kunal Poorey

Introduction

COVID-19 patient care management would greatly benefit from new tools that enable accurate assessment of disease severity and stage, potentially enabling a personalized medicine approach. Detection of the SARS-CoV-2 virus itself, or even quantitation of viral loads, is not sufficient for accurate assessment of disease state beyond diagnosis of infection [eg, doi:10.1093/cid/ciaa344]. Levels of usual-suspect protein biomarkers associated with host response to infection [eg, C-reactive protein (CRP); cytokines like IL-6, TNF-alpha, and IL-10; complement proteins like C3a and C5a], and of individual blood cell types (eg, leukocytes, lymphocytes, and subsets thereof), show limited correlation with disease severity and stage, with high patient-to-patient and study-to-study variability [eg, doi:10.1093/cid/ciaa248]. High-dimensional panels of biomarkers should have greater predictive power and resilience to unavoidable sources of variability; however, their assembly from proteins and cell types is extremely difficult, due to technical limitations in analyte measurement, especially with regard to starting material requirements and detection sensitivity. Host response profiling through Next Generation Sequencing (NGS) of gene expression patterns (ie, RNA-Seq) is a promising approach, but at the time of this project there were only two publicly available datasets of relevance [doi:10.1093/cid/ciaa203, doi:10.1080/22221751.2020.1747363], and close inspection of them revealed that each had at least one major flaw that severely undermined its value in supporting robust analysis of host response to SARS-CoV-2 infection. However, the first of these studies [doi:10.1093/cid/ciaa203] fortuitously collected NGS data not only from host cells, but also from bacteria present in bronchoalveolar lavage fluid (BALF) recovered from COVID-19 patients; and because the respiratory microbiome (in terms of bacterial species content) is far less complex than the human transcriptome, the NGS data collected were sufficient to provide coverage depth supporting robust analysis. Surprisingly, the authors of the study did not carry out a detailed analysis of these data and their potential for revealing important new information about COVID-19. Therefore, we carried out a meta-analysis of the dataset as a first step in evaluating the potential for profiling of respiratory microbiome dynamics as a means of accurately assessing COVID-19 disease state.

Results

The NGS data described in the report [doi:10.1093/cid/ciaa203] were generated through RNA-Seq profiling of BALF specimens recovered from 8 COVID-19 (CoV) patients, as well as from 25 community-acquired pneumonia (CAP) patients and 20 healthy subjects for comparison. We retrieved the 734M NGS reads (a mixture of 76bp single-end, 101bp single-end, and 151bp paired-end reads generated using Illumina HiSeq) from the BioSample database of CNCB-NGDC (BioProject PRJCA002202). The reads were quality filtered and mapped to reference genomes (human, bacterial, viral, and fungal) using our in-house bioinformatics pipeline (RapTOR) [Biotechniques 53:373 ’12; RNA Biol 10:502 ‘13], revealing that 102M of the reads (~14%) mapped to bacterial 16S ribosomal RNA (rRNA) genes. The mapping analysis also revealed that 9 of the BALF specimens (2 CAP and 7 healthy) yielded too few 16S rRNA hits (each <5000 hits, corresponding to <0.01% of total hits to 16S rRNA in the dataset) to support further analysis, leading us to leave them out of subsequent analyses. Using the 16S rRNA data from the remaining 44 BALF specimens, 312 different bacterial families were detected (setting ≥10 mapped hits as a limit-of-detection threshold). Of these 312 bacterial families, 191 (~61%) were detected in at least 1 of the BALF specimens from healthy subjects, but only 45 (~14%) were detected in ≥50% of these specimens. This indicates that a large number of bacterial families are represented in the respiratory microbiome of healthy subjects, but a much smaller subset can be detected in any given human subject; we refer to bacterial families comprising this subset as the "core" constituents of the respiratory microbiome. Similar analyses revealed that 282 bacterial families were detected in the BALF specimens from CAP patients, with 98 of them belonging to the core respiratory microbiome of CAP patients; and 292 bacterial families were detected in COVID-19 specimens, with 183 of them belonging to the core respiratory microbiome of COVID-19 patients.

Remarkably, all 45 core families in healthy subjects also qualified as core families in CAP and COVID-19 patients (**Figure 1**). 16 of these "shared" core families showed significant differences in relative abundance in healthy *versus* CAP *versus* COVID-19 subjects (one-way ANOVA, p<0.05) (**Table 1**). Of these, 2 families (Actinomycetaceae, Lactobacillaceae) were >5-fold more abundant in COVID-19 patients than in healthy subjects or CAP patients; 2 families (Brucellaceae, Neisseriaceae) were >5-fold less abundant in COVID-19 patients than in healthy subjects or CAP patients; and 1 family (Methylobacteriaceae) was >5-fold more abundant in COVID-19 patients than in healthy subjects, but >5-fold less abundant in COVID-19 patients than in CAP patients.

Additionally, 85 of the core families in COVID-19 patients did not meet the criteria to be considered core families in healthy subjects or CAP patients. 7 of these core families "unique" to COVID-19 (Chitinophagaceae, Comamonadaceae, Enterobacteriaceae, Micrococcaceae, Moraxellaceae, Staphylococcaceae, Streptococcaceae) were detected in high abundance (>1% of total 16S rRNA hits), with 3 of these (Comamonadaceae, Enterobacteriaceae, Streptococcaceae) detected in extremely high abundance (>5% of total 16S rRNA hits) (**Table 2**).

Conclusions

The results from this meta-analysis clearly indicate that the respiratory microbiome of COVID-19 patients is radically different from those of healthy subjects and CAP patients. We identified the 12 bacterial families that best illustrate this difference. For 5 of these bacterial families, measurement of their relative abundance in BALF should enable recognition of COVID-19; and for the remaining 7 bacterial families, detection of them at all should be sufficient for recognition of COVID-19. It is important to keep in mind that while the BALF RNA-Seq dataset used as the basis for this study served as an excellent starting point, analysis of additional studies will be required to determine whether the observed differences in respiratory microbiome content are resilient to study-to-study variability. However, our results strongly suggest that classifiers based on the 12 bacterial families of greatest interest (or, ideally, small subsets of them) should enable diagnosis of COVID-19 through BALF analysis (eg, using qPCR or microarrays, which are faster than RNA-Seq). Moreover, these classifiers, or variations on them, may also enable accurate assessment of COVID-19 severity and stage, a hypothesis that should be tested in future work. Successful verification of the observed trends and proposed classifiers in future prospective studies should lead to commercial licensing for assay implementation and rapid translation to clinical use, potentially improving patient management in the ongoing and/or future COVID-19 outbreaks.

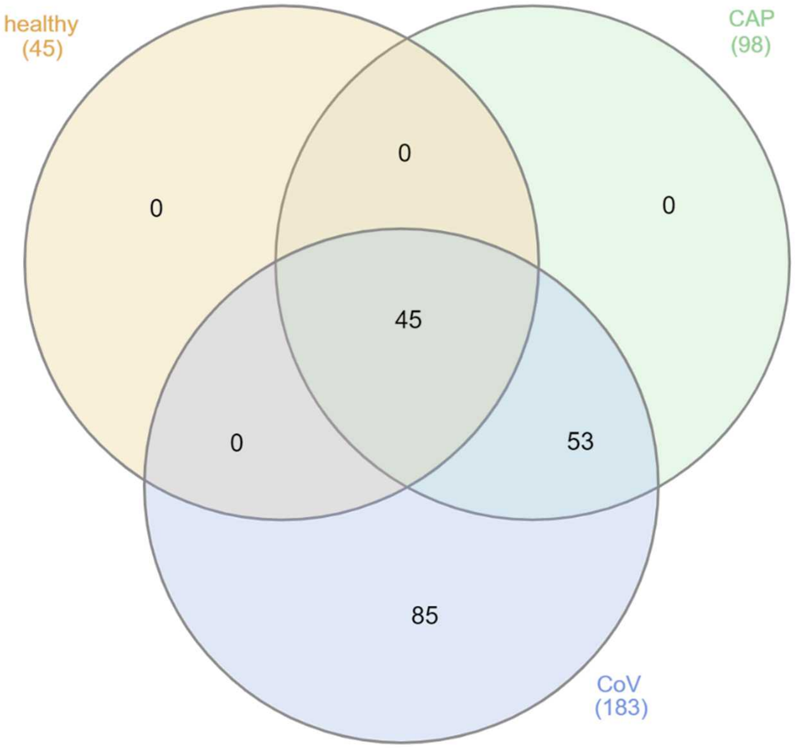
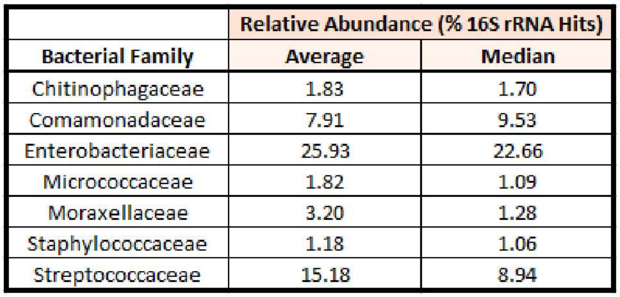


Figure 1: Numbers of bacterial families consistently detected in BALF specimens recovered from healthy *vs* CAP *vs* COVID-19 subjects.

	Relative Abundance (% 16S rRNA Hits)				Ratio of Relative Abundances	
Bacterial Family	**Healthy**	**CAP**	**COVID-19**	**ANOVA *p*-value**	**COVID-19 vs Healthy**	**COVID-19 vs CAP**
Actinomycetaceae	0.07	0.21	1.39	0.0307	**18.86**	**6.52**
Bradyrhizobiaceae	0.53	2.60	0.62	0.0005	1.17	0.24
Brucellaceae	24.56	10.82	0.09	0.0252	**0.00**	**0.01**
Comamonadaceae	0.16	8.85	7.91	0.0003	50.33	0.89
Enterobacteriaceae	2.85	8.37	25.93	0.0016	9.08	3.10
Enterococcaceae	0.08	0.08	0.39	0.0105	4.94	4.78
Lachnospiraceae	0.15	0.22	0.99	0.0292	6.71	4.50
Lactobacillaceae	0.02	0.14	1.28	0.0002	**55.79**	**8.87**
Methylobacteriaceae	0.09	33.80	0.54	0.0001	**6.09**	**0.02**
Microbacteriaceae	0.19	0.15	0.54	0.0249	2.83	3.49
Neisseriaceae	6.05	2.03	0.32	0.0488	**0.05**	**0.16**
Phyllobacteriaceae	0.08	0.05	0.20	0.0077	2.54	3.82
Porphyromonadaceae	5.59	1.08	0.24	0.0052	0.04	0.22
Rhodospirillaceae	0.12	1.50	0.04	0.0002	0.32	0.03
Sphingomonadaceae	6.99	3.66	1.00	0.0240	0.14	0.27
Streptococcaceae	5.04	2.63	15.18	0.0089	3.01	5.76

Table 1: Relative abundances of shared core bacterial families in BALF specimens recovered from healthy vs CAP vs COVID-19 subjects.



	Relative Abundance (% 16S rRNA Hits)	
Bacterial Family	Average	Median
Chitinophagaceae	1.83	1.70
Comamonadaceae	7.91	9.53
Enterobacteriaceae	25.93	22.66
Micrococcaceae	1.82	1.09
Moraxellaceae	3.20	1.28
Staphylococcaceae	1.18	1.06
Streptococcaceae	15.18	8.94

Table 2: Relative abundances of core bacterial families “unique” to COVID-19 patients.