Functional pathways in respiratory tract microbiome separate COVID-19 from community-acquired pneumonia patients

Niina Haiminen[1], Filippo Utro[1], Ed Seabolt[2], and Laxmi Parida[1,*]

[1]IBM T. J. Watson Research Center, Yorktown Heights, NY, USA
[2]IBM Almaden Research Center, San Jose, CA, USA
[*]parida@us.ibm.com

ABSTRACT

In response to the global pandemic of the last four months, some progress has been made in understanding the molecular-level host interactions of the new coronavirus SARS-CoV-2 responsible for COVID-19. However, when the virus enters the body it interacts not only with the host but also with the micro-organisms already inhabiting the host. Understanding the virus-host-microbiome interactions can yield additional insights into the biological processes perturbed by the viral invasion. We carry out a comparative functional analysis of bronchoalveolar lavage fluid of eight COVID-19, twenty-five community-acquired pneumonia (CAP) patients and twenty healthy controls. The resulting functional profiles clearly separate the cohorts, even more sharply than just their corresponding taxonomic profiles. We also detect distinct pathway signatures in the respiratory tract microbiome that consistently distinguish COVID-19 patients from both the CAP and healthy cohorts. These include increased vitamin, drug, nucleotide, and energy metabolism during SARS-CoV-2 infection, contrasted with decreased amino acid and carbohydrate metabolism. This comparative analysis indicates consistent differences in COVID-19 respiratory tract metatranscriptomes compared to CAP and healthy samples.

Introduction

An impressive number of scientific studies have rapidly been published on the genomics and molecular-level host interactions of the new coronavirus SARS-Cov-2[1] of reported bat origin,[2] responsible for the COVID-19 disease pandemic. Additionally, understanding changes in the microenvironment within the host can yield insights into related biological processes perturbed by the virus. The respiratory tract microbiome community composition during pathogen infections has been studied previously, along with its predictivity of clinical outcomes and associated potential probiotic interventions.[3–7] Studies profiling the respiratory microbiome community in COVID-19 patients are also beginning to appear.[1,8–10]

To better understand the role of the microbiome in COVID-19, we introduce a functional analysis of recently published metatranscriptomes[8] from bronchoalveolar lavage fluid (BALF) of COVID-19 patients, healthy subjects, and community-acquired pneumonia (CAP) cases. While Shen et al.[8] focused on the SARS-Cov-2 genomes and taxonomic profiling of the microbiomes, here we perform *functional profiling* to characterize active biological processes.

The sequenced reads were classified with PRROMenade[11] using a vast amino acid sequence collection of 12 million bacterial and viral protein domains from the IBM Functional Genomics Platform,[12] annotated with KEGG enzyme codes from a corresponding functional hierarchy.[13] Post-processing and robust rank-based RoDEO[14] projection of the annotations onto a unified scale made the resulting profiles comparable. The overall analysis workflow is depicted in Fig. 1A.

The functional profiling separated samples by cohort, and the more abundant enzymes codes in COVID-19 mapped to vitamin, drug, nucleotide, and energy metabolism pathways, while the less abundant ones mapped mostly to amino acid and carbohydrate pathways. In related literature, similarly altered pathways were indicated in the lower respiratory tract microbiomes of mice infected with the influenza virus.[15] The results from this robust comparative analysis highlight consistent differences in microbial pathways in COVID-19 compared to community-acquired pneumonia and healthy control samples.

Results

Microbiome functional profiles cluster by cohort

The functional profiles were first visualized with Krona[16] to examine their average distribution per cohort (Fig. 1B, see Supplemental File 1 for interactive version including all samples). While there are certain differences detectable in these coarsely averaged profiles, a subsequent robust comparative analysis reveals specific functions that are consistently altered between cohorts.

Multi-dimensional scaling of pairwise Spearman distances between samples clearly separates the cohorts, with just a few co-locating healthy control and CAP samples (Fig. 1C). A significant difference in functional profiles was observed between the COVID-19, CAP, and healthy control groups according to PERMANOVA test ($p \leq 0.0001$). Functional profiling separated the groups more sharply than taxonomic profiling ($R^2 = 0.12$ from functional profiling vs. $R^2 = 0.07$ from taxonomic profiling by Shen et al[8]). Nearly identical results were also obtained when utilizing only the bacterial subset of the database for read matching; the separation is not merely due to COVID-19 sample reads matching SARS-CoV-2 protein domains (Supplemental Fig. S1).

Differentially abundant functions distinguish COVID-19 samples

The sequencing data had varying total number of reads and human content per sample (Supplemental Fig. S2). Therefore we used RoDEO[14] to project the samples onto a robust comparable scale. To examine the features that differentiate COVID-19 from others, we extracted the top-ranked features from the COVID vs. CAP and COVID vs. healthy control comparisons, and considered the union of top 30 features from each comparison, resulting in 51 EC numbers. Their clustering shows three main feature clusters that correspond to the sample clusters of i) 20 healthy controls (with 8 CAP samples included), ii) 15 CAP, and iii) 8 COVID-19 (with 2 CAP samples included) (Fig. 2A, three outlined rectangles).

The CAP patient samples were collected from different hospital sources, yet these samples cluster together and separate from the COVID-19 patient samples. While the features were selected as those differentiating COVID-19 from CAP and healthy controls, they also tend to separate CAP from the controls. Although two CAP samples cluster here with COVID-19, note that the CAP samples were collected prior to the current pandemic and represent distinctly different pneumonia cases.

The bottom five COVID-19 samples in Fig. 2A (3,4,5,7,8) separate from other samples, with more abundant features including 4.4.1.3 "Sulfur metabolism" and 1.7.2.4 "Nitrogen metabolism" related to energy metabolism, and 2.7.4.25 "Prokaryotic cytidylate kinase" related to pyrimidine metabolism. The result is also an indication that the samples do not merely cluster by number of sequenced reads or fraction of microbial annotated reads, since according to those statistics the COVID-19 samples 6 and 7 appear quite similar (Supplemental Fig. S2) while here are in different COVID-19 subclusters.

Altered lung microbiome pathways indicated in COVID-19

We further checked the 51 top differential functions shown in Fig. 2A for extract those whose average abundance changed in a consistent direction in the COVID-19 cohort compared to both the CAP and the healthy cohort. The average difference between COVID-19 vs. CAP and COVID-19 vs. healthy controls was visualized against their corresponding pathways from the KEGG pathway mapping (Fig. 2B).

Functions with increased relative abundance are linked to energy metabolism (methane, nitrogen, sulfur), metabolism of cofactors and vitamins (B6, folate, lipoic acid) and purine – related to one-carbon metabolism, nucleotide metabolism (purine, pyrimidine), and drug metabolism. Several functions with decreased relative abundance are related to amino acid and carbohydrate metabolism. Decreased amino acid and carbohydrate metabolism with increased metabolism of cofactors and vitamins and nucleotide metabolism were also indicated among the changes in lower respiratory tract microbiomes at the acute phase (day 7) of mice infected with influenza virus, with increased energy metabolism in the early days of infection (day 3).[15]

Discussion

Analyzing COVID-19 respiratory tract metatranscriptomes from a functional perspective offers an additional view into their differences from healthy controls and subjects with community-acquired pneumonia (CAP). Querying translated reads against amino acid sequences allows for more flexibility in sequence matching, as synonymous mutations will not affect the matching. This also supports the detection of functions performed by unknown organisms yet to be characterized.[17]

The resulting functional profiles clearly distinguish COVID-19 from CAP and healthy controls, even more so than the original taxonomy-based analysis of the community members.[8] From this profiling, we detected consistent differences in COVID-19 samples compared to both the healthy control and CAP samples.

Features with increased relative abundance mapped to metabolism of vitamins, drugs, purine, pyrimidine, and energy metabolism. Several features with decreased relative abundance were related to amino acid and carbohydrate metabolism. Similar observations were recently made in a comprehensive study of microbiomes in mice infected with the influenza virus.[15] With limited clinical data available for these patients it is unclear if administered therapies had an effect on the detected microbiome differences and the detected subclusters of COVID-19 samples. The comparative approach taken here carefully avoids possible experimental variation and resulting biases within individual samples, and focuses on detecting robust and consistent differences between cohorts.

These preliminary observations call for further in-depth analysis of microbial functions and host-microbiome interactions from studies coupling sequencing and clinical data, as viewing them with the lens of functional annotation can yield additional insights into the altered biological processes.

Methods

Sequence data and functional database

The recently published bronchoalveolar lavage fluid (BALF) metatranscriptomic sequencing data of 8 COVID-19 patients, 20 healthy controls, and 25 samples of community-acquired pneumonia (CAP) were obtained from the National Genomics Data Center.[8, 18] Pre-processing was performed as described by Utro et al.[11] including Trim Galore[19] trimming and bowtie2[20] human read filtering, resulting in 40k to 32M input reads for microbial functional annotation per sample (Supplemental Fig. S2). Human reads were already filtered by Shen et al.[8], still we removed additional reads that mapped to human with the bowtie 2 local alignment mode.

The KEGG Enzyme Nomenclature code (EC) reference hierarchy[13] was used as the functional annotation tree. EC numbers define a four-level hierarchy. For example, 1.5.1.3. = "Dihydrofolate reductase" is a fourth (leaf) level code linked to top level code 1 = “Oxidoreductases”, via 1.5. = “Acting on the CH-NH group of donors" and 1.5.1 = “With NAD+ or NADP+ as acceptor”.

A PRROMenade[11] database was constructed from a total of 11.88M bacterial and 51k viral annotated protein domain sequences, obtained from the IBM Functional Genomics Platform[12] (formerly known as OMXWare). The virus domain sequences (including from SARS-Cov-2) were added to the previously described bacterial sequence collection,[11] totaling 3.72 billion amino acids.

Functional annotation and downstream analysis

Metatranscriptomic sequencing reads were annotated with PRROMenade by locating the maximal length exact match for each read and processed as described by Utro et al.[11] Minimum match length cutoff of 9 AA (27 nt) was employed. Classified read counts (13k to 5.0M per sample, see Supplemental Fig. S2–S3) were post-processed to summarized the counts at the leaf level of the functional hierarchy. Leaf nodes contributing $\geq 0.1\%$ of total annotated reads in at least one sample were retained, 595 nodes. Functional profiles were processed with RoDEO[14] for robust comparability (parameters P=10, I=100, R = 10^7). A two-sample Kolmogorov-Smirnov test was applied to identify differentially abundant features between COVID-19 samples and CAP, healthy control samples.

References

1. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).

2. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nat Medicine* **26**, 450–452 (2020).

3. Hanada, S., Pirzadeh, M., Carver, K. Y. & Deng, J. C. Respiratory viral infection-induced microbiome alterations and secondary bacterial pneumonia. *Front. Immunol.* **9**, 2640 (2018).

4. Mendez, R., Banerjee, S., Bhattacharya, S. K. & Banerjee, S. Lung inflammation and disease: A perspective on microbial homeostasis and metabolism. *IUBMB Life* **71**, 152–165 (2019).

5. Dickson, R. P. *et al.* Lung microbiota predict clinical outcomes in critically ill patients. *Am. J. Respir. Critical Care Medicine* **201**, 555–563 (2020).

6. Zolnikova, O., Komkova, I., Potskherashvili, N., Trukhmanov, A. & Ivashkin, V. Application of probiotics for acute respiratory tract infections. *Italian J. Medicine* **12**, 32–38 (2018).

7. Fanos, V., Pintus, M. C., Pintus, R. & Marcialis, M. A. Lung microbiota in the acute respiratory disease: from coronavirus to metabolomics. *J Pediatr Neonat Individ. Med* **9**, e090139 (2020).

8. Shen, Z. *et al.* Genomic Diversity of Severe Acute Respiratory Syndrome–Coronavirus 2 in Patients With Coronavirus Disease 2019. *Clin. Infect. Dis.* (2020).

9. Chen, L. L. *et al.* RNA based mNGS approach identifies a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak. *Emerg. Microbes & Infect.* **9**, 313–319 (2020).

10. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).

11. Utro, F. *et al.* Hierarchically labeled database indexing allows scalable characterization of microbiomes. *iScience* **23** (2020).

12. Seabolt, E. *et al.* IBM Functional Genomics Platform, A Cloud-Based Platform for Studying Microbial Life at Scale. *arXiv:1911.02095* (2019). https://arxiv.org/abs/1911.02095.

13. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353–D361 (2017).

14. Haiminen, N. *et al.* Comparative exomics of Phalaris cultivars under salt stress. *BMC Genomics* **15 Suppl 6**, S18 (2014).

15. Gu, L. *et al.* Dynamic changes in the microbiome and mucosal immune microenvironment of the lower respiratory tract by influenza virus infection. *Front. Microbiol.* **10**, 2491 (2019).

16. Ondov, B., Bergman, N. & Phillippy, A. Interactive metagenomic visualization in a web browser. *BMC Bioinformatics* **12** (2011).

17. Kaufman, J. H. *et al.* Insular microbiogeography: Three pathogens as exemplars. *Curr Issues Mol Biol* **36**, 89–108 (2020).

18. National Genomics Data Center Members and Partners. Database Resources of the National Genomics Data Center in 2020. *Nucleic Acids Res.* **48**, D24–D33 (2020).

19. Krueger, F. TrimGalore. https://github.com/FelixKrueger/TrimGalore.

20. Langdon, W. B. Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. *BioData Min* **8** (2015).

Author contributions statement

N.H. and F.U. conducted the experiments, analyzed the results and wrote the manuscript. E.S. provided reference data from the IBM Functional Genomics Platform. N. H. and L.P. conceived the study and analyzed the results. All authors reviewed the manuscript.

Additional information

Competing interests The PRROMenade methodology is associated with patent applications currently pending review at the USPTO.

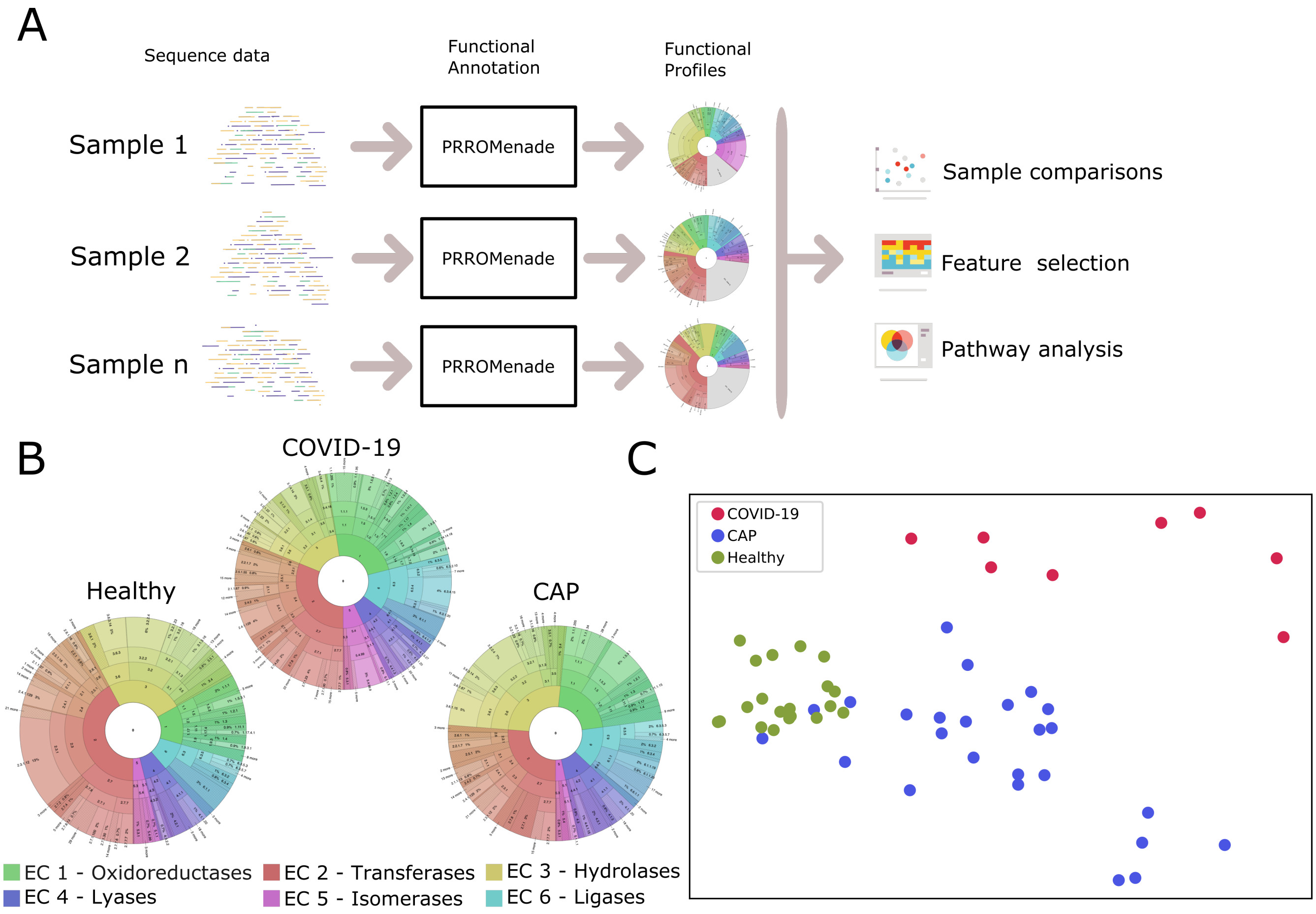

Figure 1. **A**. Overall analysis workflow. Each microbiome sequencing sample is annotated with PRROMenade, utilizing labeled reference data from the IBM Functional Genomic Platform. The resulting functional profiles are visualized and compared in downstream analyses. **B**. Visualization of averaged functional profiles per cohort with Krona. Interactive versions including all samples are included in Supplemental File 1. **C**. Multidimensional scaling of the functional profiles using the Spearman distance. Each sample is represented by a dot colored by cohort.

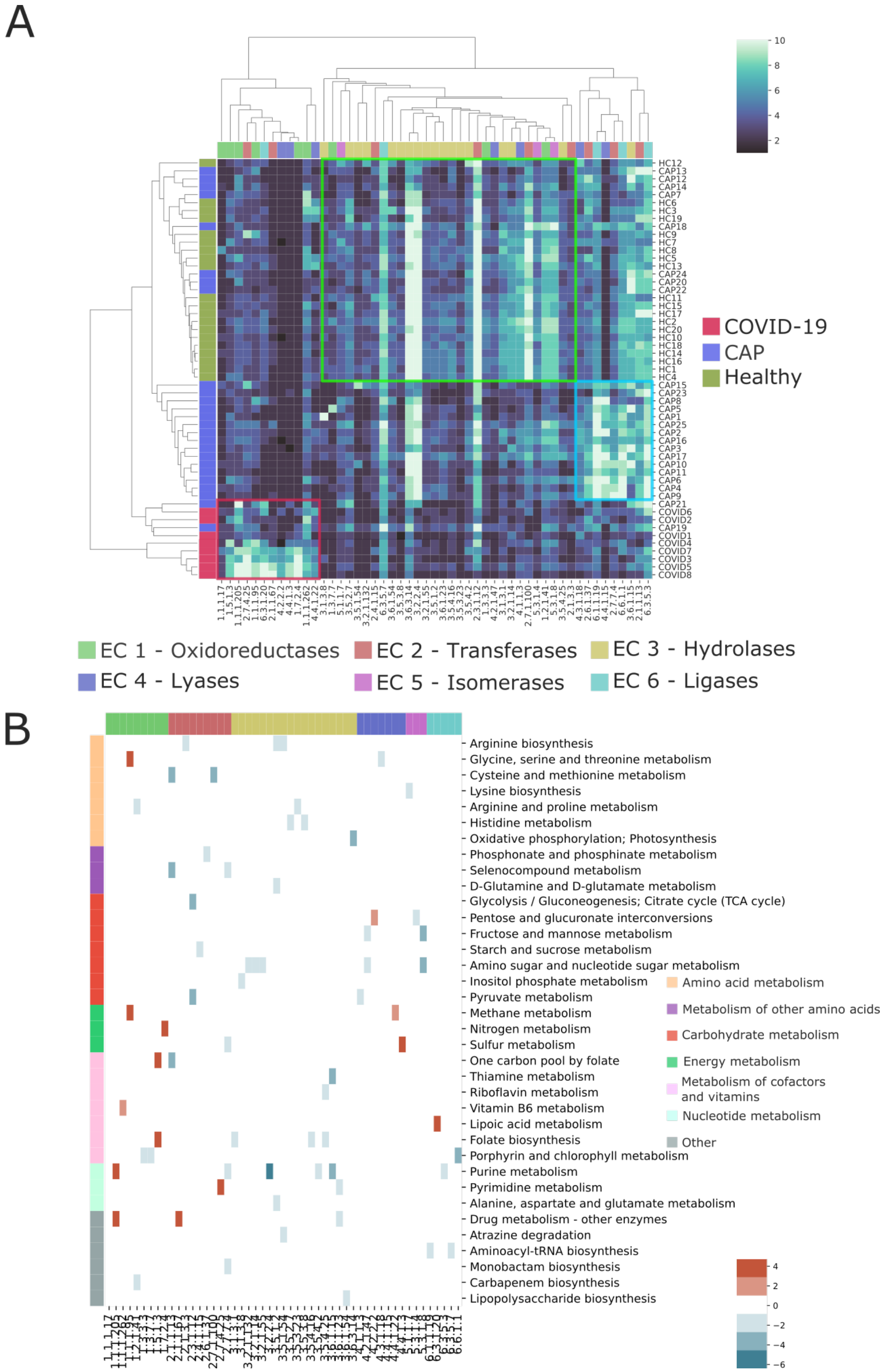

Figure 2. **A**. RoDEO processed EC abundance values (10 denotes highest possible value), for 51 features differentiating COVID-19 from CAP and healthy controls. Columns and rows are ordered independently by average linkage hierarchical clustering of features and samples. The colors attached to the dendrogram on the left reflect the cohort labels. The colors attached to the dendrogram on top reflect the top level codes in the functional hierarchy (e.g. Oxidoreductases). Three main blocks are shown within the heatmap with colored outlines for healthy, CAP, and COVID-19 clusters. **B**. The same 51 features as in A, with their pathway mappings and additional labels on the most frequent high level concepts, e.g. Amino acid metabolism. The heatmap colors reflect the average RoDEO-processed EC abundance change from CAP and healthy control samples to COVID-19 (red indicates more abundant in COVID-19). White indicates the EC number is not mapped to the pathway, or does not change from CAP and healthy to COVID-19 in the same direction. The columns are ordered by EC number and their corresponding top level codes indicated by the colors at the top of the figure.