Supplementary Materials:

TITLE:
Functional lower airways genomic profiling of the microbiome to capture active microbial metabolism

Imran Sulaiman[1], Benjamin G. Wu[1], Yonghua Li[1], Jun-Chieh Tsay[1], Maya Sauthoff[1], Adrienne S. Scott[1], Sergei B. Koralov[2], Michael Weiden[1], Jose Clemente[3,4], Drew Jones[5], Yvonne J. Huang[6], Kathleen A. Stringer[7], Lingdi Zhang[8], Adam Geber[8], Stephanie Banakis[8], Laura Tipton[8], Elodie Ghedin[8], Leopoldo N. Segal[,]*

[1] Division of Pulmonary, Critical Care, & Sleep Medicine, Department of Medicine, New York University School of Medicine, NY
[2] Department of Pathology, New York University School of Medicine, NY
[3] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, USA.
[4] Immunology Institute, Icahn School of Medicine at Mount Sinai, New York, USA
[5] Department of Biochemistry and Molecular Pharmacology and Department of Radiation Oncology, New York University School of Medicine, NY
[6] Division of Pulmonary and Critical Care Medicine, Department of Medicine, University of Michigan Medical School, Ann Arbor, MI, USA
[7] Department of Clinical Pharmacy, College of Pharmacy, and Division of Pulmonary and Critical Care Medicine, Department of Medicine, School of Medicine, University of Michigan, Ann Arbor, MI, USA
[8] Center for Genomics & Systems Biology, Department of Biology, New York University, New York, New York, USA
[9] Department of Epidemiology, School of Global Public Health, New York University, New York, New York, USA

Imran Sulaiman, MD	SheikMohammadImran.Sulaiman@nyumc.org
Benjamin G. Wu, MD	Benjamin.Wu@nyumc.org
Yonghua Li, MD, PhD	Yonghua.Li@nyumc.org
Jun-Chieh Tsay, MD	Jun-Chieh.Tsay@nyulangone.org
Maya Sauthoff	mayasauthoff@gmail.com
Adrienne Scott MS,	Adrienne.Scott@nyumc.org
Sergei B. Koralov, PhD	Sergei.Koralov@nyulangone.org
Michael Weiden, MD	Michael.Weiden@nyulangone.org
Jose Clemente	jose.clemente@mssm.edu
Drew Jones, PhD	Drew.Jones@nyulangone.org
Yvonne Huang, MD	yvjhuang@med.umich.edu
Kathleen Stringer, PharmD	stringek@umich.edu
Lingdi Zhang	lz967@nyu.edu
Adam Geber	adam.geber@gmail.com
Stephanie Banakis	sb6179@nyu.edu
Laura Tipton	Tipton.Laura@gmail.com
Elodie Ghedin, PhD	eg121@nyu.edu

Leopoldo N. Segal, MD[1] Leopoldo.Segal@nyumc.org

Corresponding Author/ Address for Reprints:
Leopoldo N. Segal, MD[1] Leopoldo.Segal@nyumc.org

NYU School of Medicine
462 First Ave 7W54
New York, NY 10016
Tel: (212) 562-3752
Fax: (212) 263-7445

Participants and samples

Inclusion criteria included >50 years old with significant smoking history. Exclusion criteria included recent infection or antibiotic use in the prior three months, no inhaler therapy, no diabetes, cardiovascular or renal disease. On all participants we performed research bronchoscopy to collect bronchoalveolar lavage (BAL) samples, upper airway (UA, which were either supraglottic or oral rinse) and bronchoscope control (BKG) as previously described [1, 2]. Technical background controls (DNA free water passed through DNA/RNA isolation kit) were included.

DNA/RNA isolation:

DNA was extracted from all samples using Qiagen DNA Mini Kit spin column protocol (Qiagen). RNA extraction was carried out with the miRNeasy Micro Kit (Qiagen). RNA was eluted in either 15 or 30µl. RNA quantity and integrity were tested with TapeStation 4200 (Agilent) for all samples. RNA quality control was established using an RNA Integrity Number (RIN) cut-off >5. The BAL of two subjects did not yield quality RNA (RIN < 5). Among BKG samples, only one yielded RNA with sufficient quality and quantity to undergo library and sequencing for the metatranscriptome. The Illumina Ribo-Zero Gold rRNA Removal Kit (Epidemiology) was used to deplete the rRNA from the samples with a modified low input version.

Bacterial Burden assessment

We measured bacterial burden in all samples using a QX200 Droplet Digital PCR System (BioRad, Hercules, CA). For this, primers were 5'- GCAGGCCTAACACATGCAAGTC-3' (63F) and 5'- CTGCTGCCTCCCGTAGGAGT-3' (355R). Cycling conditions included 1 cycle at 95°C for 5 minutes, 40 cycles at 95°C for 15 seconds and 60°C for 1 minute, 1 cycle at 4°C for 5 minutes, and 1 cycle at 90°C for 5 minutes all at a ramp rate of 2°C/second. The BioRad C1000 Touch Thermal Cycler was used for PCR cycling. Droplets were quantified using the Bio-Rad Quantisoft software. Two replicates were used per sample. Negative control specimens (detailed in the text) were used and were run alongside samples.

Microbial community characterization using 16S rRNA gene sequencing (16S rRNA):

High-throughput sequencing of bacterial 16S rRNA gene amplicons encoding the V4 region was performed as previously described.[2] Each unique barcoded amplicon was generated in pairs of 25µl reactions with the following reaction conditions: 11µl PCR-grade H2O, 10µl Hot MasterMix (5 Prime Cat# 2200410), 2µl of forward and reversed barcoded primer (5µM, 515F= 5'-GTGCCAGCMGCCGCGGTAA-3', and 806R= 5'-GGACTACHVGGGTWTCTAAT-3'), and 2µl template DNA. Amplification and detection of the 16S rRNA gene by qPCR was performed with the StepOne™ Real-time PCR System (Applied Biosystems, Foster City, CA, USA). The PCR reaction condition for amplification of DNA were: initial denaturing at 94°C for 3 min followed by 35 cycles of denaturation at 94°C for 45 seconds, annealing at 58°C for 1 minute, and extension at 72°C for 90 seconds, with a final extension of 10 min at 72°C. Amplicons were quantified using Agilent 2200 TapeStation system and pooled. Purification was then performed

using Ampure XT (Beckman Coulter Cat# A63882) as per the manufacturer instructions. Reagent controlled samples and mock mixed microbial DNA were sequenced and analyzed in parallel with the samples. Sequence was performed on the Illumina MiSeq (150bp read length, paired-end protocol) in one single run.

The 16S rRNA gene sequences were analyzed using the Quantitative Insights into Microbial Ecology (QIIME 1.9) package [3]. Reads were demultiplexed and quality filtered with default parameters. We required > 1,000 reads in any sample, a threshold that was achieved with all samples (31,524 [23,128 – 77,405] reads). Reads were demultiplexed and quality filtered with default parameters. Sequences were then clustered into operational taxonomic units (OTUs) using a 97% similarity threshold with UCLUST [4] and annotated using the Greengenes 16S rRNA gene reference dataset [5]. For each sample, the proportion of reads at the OTU or genus levels was used as a measure of the relative abundance of each type of bacteria in a specimen. No OTU was removed from the analysis. To infer the genomic potential of the microbial communities identified by 16S rRNA gene sequencing we computationally predicted the metagenome using Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt).[6] This software tool uses the obtained 16S rRNA sequence data to predict the functional profile of a bacterial community based on a database of existing reference genomes.

Microbial community characterization using Whole genome shotgun sequencing (WGS):

Extracted DNA was used as input into the NexteraXT library preparation kit following the manufacturer's protocol. Libraries were purified using the Agencourt AMPure XP beads (Beckman Coulter, Inc.) to remove fragments below 200 bp. The purified libraries were quantified using the Qubit dsDNA High Sensitivity Assay kit (Invitrogen) and the average fragment length for each library was determined using a High Sensitivity D1000 ScreenTape Assay (Agilent). Samples were added in an equimolar manner to form two sequencing pools. The sequencing pools were quantified using the KAPA Library Quantification Kit for Illumina platforms. The pools were then sequenced on the Illumina NextSeq500 (2x150 bp) in one single run.

The metagenomic sequences were filtered to remove the adaptor sequences, and human sequences by using Trimmomatic [7], SortMeRNA [8] and DeconSeq [9]. In addition, fungal and viral sequences were removed. The median sequence read count for raw reads was 32.5M (IQR=18.11M) per sample. After rRNA removal, the median read count was 30.2M (IQR=16.90M), and after human read removal it was 886368 [IQR: 1217360 – 3264527]. The remaining metagenomic sequences were processed for functional and taxonomic annotation using the HUMAnN2 pipeline [10] with the UniRef90 database [11]. The gene assignments were regrouped by KEGG terms and joined into a gene table.

Microbial community characterization using RNA metatranscriptomic sequencing (RNA)

NEB Ultra II RNA Library Prep kit was used for library preparation and each library was subjected to 14-17 cycles of PCR in a manner proportional to RNA input mass. Adaptor concentration was diluted 1:100 or 1:200 to maintain appropriate sample and adaptor

ratio. Library molarity was quantified on the LightCycler 480 (Roche) using the KAPA Library Quantification Kit. Each library was diluted to 20µM before pooling (5 libraries/pool, each pool intended for one lane of the flowcell) and the molarity of each pool was remeasured using the same assays before further dilution to 2µM in preparation for clustering. Two BAL samples failed to amplify and did not generate a typical fragment size distribution; they were excluded from sequencing. Sequence was performed in one single run using HiSeq 2500 in High Output mode with a TruSeq SBS v3-HS kit (2x100 paired-end approach).

Metatranscriptome sequences were filtered to remove the adaptor sequences, ribosomal RNA and human sequences by using Trimmomatic, SortMeRNA, and DeconSeq [7-9]. In addition, fungal and viral sequences were removed. The median read count for the raw reads was 84.8M (IQR=26.86M) per sample. After rRNA removal, the median read counts were 60M (IQR=14.30M); after removing the human reads, it was 1794919 [IQR: 766116 – 58743048]. The first three bases on the reverse sequences were removed using Seqtk due to low quality. The remaining metatranscriptomic sequences were processed for functional and taxonomic annotation using the HUMAnN2 pipeline [10] with the UniRef90 database [11]. The gene assignments were regrouped by KEGG terms and joined into one gene table.

Measurement of Short Chain Fatty Acids

Each sample (50µl) was diluted to 100µl with HPLC-grade water. A solution of hydrochloric acid (30mM) plus isotopically-labeled acetate (150µM), butyrate (10µM),

and hexanoate (2µM) in water was added to each sample and vortexed. Methyl tert-butyl ether (MTBE; 300µl) was added to each sample, and the mixture was vortexed (10 s) to emulsify, then held at 4°C for 5 min, and vortexed again (10 s). Samples were centrifuged (1 min) to separate the solvent layers. The MTBE layer was then removed and transferred to an autosampler vial for GC-MS analysis. A small volume (10µl) of the MTBE layer was removed from each sample and pooled in a separate autosampler vial for quality control purposes. A series of calibration standards were prepared along with samples to quantify metabolites.

GC-MS analysis was performed on an Agilent 69890N GC- 5973 MS detector (Agilent, Santa Clara, CA USA) using the following parameters: sample (1µl) was injected with a 1:10 split ratio on a ZB-WAXplus, 30m_ 0.25mm x 0.25 mm (Phenomenex Cat# 7HG-G013-11) GC column, with helium as the carrier gas at a flow rate: 1.1 ml/min. The injector temperature was 240°C, and the column temperature was isocratic at 310°C. Data were processed using MassHunter Quantitative analysis version B.07.00 (Agilent). Two replicated injections of five standards were used to create a linear calibration curve with accuracy better than 80% for each standard. Each detected SCFA was normalized to the nearest isotope labeled internal standard and quantitated from the linear phase of the standard curve. Normalization between sample type, by measuring urea, was not performed due previously published limitations with this method [12-14].

Mouse Experiment

Twenty female C57BL/6J mice (8-14 weeks of age, 18-22 grams/mouse) were purchased from a vendor (Jackson Research Laboratories, Bar Harbor, ME Cat#000664). All mice were allowed 2 weeks of acclimation to their facilities prior to the start of experiments and mice with different experimental condition arm were cohoused to limit potential cage effect on the microbiome and host immune tone.

Three mice were inoculated with PBS while the remaining 17 mice were inoculated with a mixture of human oral commensals (MOC) consisting of *Prevotella, Streptococcus and Veillonella.* The mice were anesthetized using isoflurane via VetFlo Anesthesia machine (Kent Scientific, Torrington, CT) sedated to 10-15 breaths per minute and monitored for any distress. The mice were then placed on an intubation platform and with blunt forceps, their tongue was gently pulled ventrally until the pharynx was exposed [15]. A human otoscope with a 2 mm ear cone (Welch Allyn 3.5V Hill-Rom, Inc., Skaneateles Falls, NY Model #20200) was introduced into the oral airway to expose and visualize the murine vocal cords and a gel-loading pipette tip loaded with a 50 ml aliquot was introduced through the vocal cords of the mouse and deployed into the lower airway. Then, the mouse was removed from the platform to recover from anesthesia on a heat pad. Mice were monitored every 2-4 hours following intra-tracheal challenge. The MOC exposed mice were sacrificed at each time point: 1Hr (n=4), 4Hr (n=4), 1Day (n=3), 3Day (n=3) and 7Day (n=3) (**Figure 6A**). The PBS exposed mice were all sacrificed at 1Hr. BAL samples were sent for 16S rRNA gene sequencing and RNA metatranscriptome sequencing. The animal studies described in these experiments were approved by the Institutional Animal Care and Use Committee at the

respective institutions (New York University School of Medicine IACUC# s16-00032). Laboratory animal care policies follow the Public Health Service Policy on Humane Care and Use of Laboratory Animals.

Statistical Analysis

For association with discrete factors, we used non-parametric tests (Mann-Whitney or Kruskal-Wallis ANOVA). We used the *vegan* package in R to construct Principal Coordinate Analysis (PCoA) based on Bray-Curtis distances [16, 17] and PERMANOVA to test statistical differences based on β-diversity. To cluster microbiome communities into exclusive 'metacommunities' we used a Dirichlet Multinomial Mixture Model using the R package *DirichletMultinomial* [18, 19]. For tests of association with continuous variables, we used non-parametric Spearman correlation tests. To evaluate differences between groups within each sequence data type, we evaluated differential expression with DESeq2 [20] with a false discover rate (FDR) <0.05 [21]. To compare the sequencing methods, we used Gene Set Enrichment Analysis (GSEA, [22]) and PROCRUSTES (an analytical pipeline that can overlay paired sample β diversity distribution obtained by two different sequence data types). Summary of KOs for pathway analysis was performed by aligning KOs with associated pathways from the KO Database (https://www.genome.jp/kegg/ko.html). With KOs with multiple pathways, the read count for that KO was added to each pathway. Data was then summarized for each pathway. Further analysis was also done to identify contaminants from the three different sequencing datasets. To perform this, a prevalence-based method using the R package *decontam[23],* with a threshold of 0.7, was used. In this process, all reads from BKG

(bronchoscope control) samples were identified as negative controls and prevalence of each sequence feature in all other sample types was identified as contaminants. All data is publicly available in Sequence Read Archive (SRA) under accession numbers PRJNA603592, PRJNA573853 and PRJNA603675. All codes utilized for the analysis included in this manuscript are available at:

https://github.com/segalmicrobiomelab/functional_microbiomics

Supplementary Results:

Taxonomic signature differences between sequencing data types

To evaluate similarities of taxonomic annotation at a global level we used PROCRUSTES with Monte-Carlo simulation. While there was high correlation in β-diversity between WGS and RNA taxonomic assignment (p=0.001), there was no significant correlation when 16S rRNA gene sequencing data was compared with WGS or RNA data (**Supplementary Figure 3**).

Based on 16S rRNA gene sequencing, several taxa were significantly enriched (FDR <0.05) in samples identified as BAL.SPT as compared with BAL.BPT (**Figure 2A, Supplementary Data 2**) with a number of known oral commensals, such as *Prevotella*, *Veillonella* and *Streptococcus*, enriched in BAL.SPT. Using 16S rRNA gene sequencing SPT/BPT assignment, we evaluated the taxonomic differences identified with the other sequencing methods. Comparing BAL.16S.SPT to BAL.16S.BPT, there were fewer significantly enriched taxa within the WGS (**Figure 2C, Supplementary Data 2**), most of

which were annotated to the genus *Streptococcus*. However, in the RNA sequencing data, there were a number of taxa significantly enriched in BAL.16S.SPT samples as compared with BAL.16S.BPT samples, including *Veillonella*, *Rothia*, and *Streptococcus* **(Figure 2E, Supplementary Data 2).** Using Gene-Set-Enrichment Analysis (GSEA), we compared the different sequencing methods for 16S.SPT/16S.BPT signatures in BAL samples (at the genus level) and identified very little overlap (**Figures 2B**, **2D** and **2F**).

Functional overlap and differences across sequencing data types

We used PROCRUSTES to evaluate similarities in global functional annotation between sequence datasets. Monte-Carlo simulation showed that there was statistically significant overlap between the three datasets ($p<0.01$ for all comparisons, **Supplementary Figure 4**). The 16S rRNA gene sequence data was used to infer the predictive metagenome allowing us to identify several KEGG Orthology (KO) pathways that should be enriched in BAL.16S.SPT samples (**Figure 1F**). We then used WGS and RNA sequencing KEGG annotated data to explore functional capacity on measured genes between BAL.16S.SPT and BAL.16S.BPT. WGS data revealed few differentially enriched pathways between BAL.16S.SPT and BAL.16S.BPT, most of which were associated with Ribosome function (**Supplementary Figure 5A, Supplementary Data 3**). A larger number of differentially enriched KEGGs were identified in the RNA metatranscriptome data, all enriched in BAL.16S.SPT samples. Among these pathways we identified lipid metabolism and fatty acid metabolism as enriched (**Supplementary Figure 5B, Supplementary Data 3**).

Ex Vivo analysis of Short Chain Fatty Acids

We first evaluated our ability to detect SCFA production by oral anaerobes frequently found in the lower airways. To this end, we measured SCFA levels in an *ex vivo* culture of *Veillonella parvula* grown under anaerobic conditions. Compared to media alone, the supernatant of *Veillonella parvula* had higher levels of 3/7 SCFAs: acetate, propionate and isovalerate (**Supplementary Figure 6A**). Given the concern of volatility of SCFAs and use of these metabolites by mammalian cells in human biological samples we then measured SCFA levels after the addition of butyrate and propionate to BAL cells *ex vivo*. Levels of butyrate and propionate remained measurable for at least 1 and 3 hours, respectively (**Supplementary Figure 6B-C**) which exceeds the time normally taken to process samples from our bronchoscopies (usually < 1hr).

Further Clustering of Samples by Sequencing Method

Clustering for each of these data types show 2 clusters as the best fit (**Supplementary Figure 8A & C**). However, within WGS only 3 BAL samples clustered with UA samples (BAL.WGS.SPT, **Supplementary Figure 8B**), while with the RNA metatranscriptome only 2 BAL samples clustered with UA samples (BAL.RNA.SPT, **Supplementary Figure 8D**), all of which were the samples with the highest levels of SCFAs, as shown in **Figure 4**.

Decontamination Analysis

In both the 16S rRNA and RNA Metatranscriptome, contaminants were identified as BPT taxa such as *Flavobacterium* and *Propionibacterium (***Supplementary Figure 9A, B, &**

Supplementary Table 4*).* SPT taxa such as *Streptococcus* and *Veillonella* were not identified as contaminants. In the WGS dataset, contaminants were genes that were unmapped or ungrouped and thus did not have a taxonomic annotation (**Supplementary Figure 9C & Supplementary Table 4**).

1. Segal LN, Alekseyenko AV, Clemente JC, Kulkarni R, Wu B, Chen H, et al. Enrichment of lung microbiome with supraglottic taxa is associated with increased pulmonary inflammation. Microbiome. 2013;1(1):19.
2. Segal LN, Clemente JC, Tsay JC, Koralov SB, Keller BC, Wu BG, et al. Enrichment of the lung microbiome with oral taxa is associated with lung inflammation of a Th17 phenotype. Nature microbiology. 2016;1:16031.
3. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7(5):335-6.
4. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460-1.
5. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. Isme J. 2012;6(3):610-8.
6. Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol. 2013;31(9):814-21.
7. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.
8. Kopylova E, Noe L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 2012;28(24):3211-7.
9. Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PloS one. 2011;6(3):e17288.
10. Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol. 2012;8(6):e1002358.
11. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, UniProt C. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015;31(6):926-32.
12. Rennard SI, Basset G, Lecossier D, O'Donnell KM, Pinkston P, Martin PG, et al. Estimation of volume of epithelial lining fluid recovered by lavage using urea as marker of dilution. J Appl Physiol (1985). 1986;60(2):532-8.
13. Marcy TW, Merrill WW, Rankin JA, Reynolds HY. Limitations of using urea to quantify epithelial lining fluid recovered by bronchoalveolar lavage. Am Rev Respir Dis. 1987;135(6):1276-80.

14. Dargaville PA, South M, Vervaart P, McDougall PN. Validity of markers of dilution in small volume lung lavage. Am J Respir Crit Care Med. 1999;160(3):778-84.
15. Driscoll KE, Costa DL, Hatch G, Henderson R, Oberdorster G, Salem H, et al. Intratracheal instillation as an exposure technique for the evaluation of respiratory tract toxicity: uses and limitations. Toxicol Sci. 2000;55(1):24-35.
16. Dray SaD, A.B. The ade4 package: implementing the duality diagram for ecologists. Journal of Statistical Software. 2007;22(4):1-20.
17. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. UniFrac: an effective distance metric for microbial community comparison. The ISME journal. 2011;5(2):169-72.
18. Holmes I, Harris K, Quince C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. PLoS One. 2012;7(2):e30126.
19. Morgan M. DirichletMultinomial: Dirichlet-Multinomial Mixture Model Machine Learning for Microbiome
Data. R package version 1.20.0 ed2017.
20. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
21. Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics. 2003;19(3):368-75.
22. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545-50.
23. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome. 2018;6(1):226.