Online Supplement

Lung Microbiome Analysis and Stochastic Modelling of COPD Exacerbations in the AERIS Study

David Mayhew, Nathalie Devos, Christophe Lambert, James R. Brown, Stuart Clarke, Victoria Kim, Michal Magid-Slav, Bruce E. Miller, Kristoffer Ostridge, Ruchi Patel, Ganesh Sathe, Daniel F. Simola, Karl J. Staples, Ruby Sung, Ruth Tal-Singer, Andrew C. Tuck, Stephanie Van Horn, Vincent Weynants, Nicholas Williams, Jeanne-Marie Devaster, Tom M. Wilkinson; on behalf of the AERIS Study Group

Supplementary methodology

DNA extraction

Sputum samples were homogenized by adding one volume of 0.1% dithiothreitol (DTT) solution. Phosphate-buffered saline (400 µl) was added to the homogenized sample (200 µl), which was subjected to mechanical lysis. Samples were organized in a 96-well rack containing Lysing Matrix B (MP Biomedicals) and submitted to two disruption cycles (1 minute at 16000 rpm) in a FastPrep Homogenizer (MP Biomedicals). After bead sedimentation, DNA was extracted from supernatant (400 µl) using the MagNA Pure 96 system (Roche Life Science) and MagNA Pure 96 DNA and Viral NA Large Volume Kit (Roche Life Science) and the Pathogen Universal Protocol recommended by the manufacturer (elution volume, 50 µl).

Processing of sputum samples

Sputum samples were obtained by spontaneous expectoration or induced and were processed according to standard methods. Potential bacterial respiratory pathogens, including *Haemophilus influenzae*, *Moraxella catarrhalis*, *Streptococcus pneumoniae*, *Pseudomonas aeruginosa*, and *Staphylococcus aureus* were identified using conventional culture techniques and by qPCR. A qualitative nucleic acid multiplex test (xTAG® Respiratory Viral Panel Fast v2; Luminex, Austin, TX, USA) was used for the detection of viruses, including human rhinovirus (HRV), respiratory syncytial virus (RSV), influenza virus, parainfluenza virus, human metapneumovirus, adenovirus, human bocavirus, and coronavirus.

16S rRNA gene amplification and sequencing

A segment of the hypervariable 16S rRNA gene was amplified using conserved V4 region-specific primers (forward primer 515F 5'-GTGCCAGCMGCCGCGGTAA-3' and the reverse primer 806R 5'-GGACTACHVGGGTWTCTAAT-3'), including Illumina sequencing adapters.[1] The reverse amplification primer contained a 12 bp error-correcting Golay barcode sequence allowing for pooling of multiple samples in the same flowcell.[2] The primers also included nine extra bases in the adapter region of both

forward and reverse amplification primers and a pad region to avoid primer-dimer formation.

Aseptic technique and DNA-free reagents were used in a biological containment hood to avoid bacterial DNA contamination during processing. In addition, negative controls for extraction (no sputum material) and PCR amplification (no template, Qiagen Elution Buffer only) were included in each experiment. The extraction negative control for each experiment was subsequently sequenced to identify any potential contaminating bacterial species.

The amplification mix (25 µl) contained 4 µl sputum DNA, 2 µl (0.2 µM) each of forward and reverse primers (Integrated DNA Technologies, Coralville, IA), 12.5 µl of 2x KAPA HiFi HotStart Ready Mix (KK2602, Kapabiosystems, Boston MA), and 4.5 µl RNase free water. PCR amplification was performed on an ABI 9700 thermocycler using the following cycling protocol: initial denaturation at 95°C for 3 min, followed by 35 cycles of 98°C for 20 sec, 66°C for 15 sec, and 72°C for 15 sec, with a final hold of 72°C for 1 min. Aliquots of reaction mixture (3 µl each) were analyzed by 2% agarose gel (2% Egel, Invitrogen) with samples containing a band of approximately 385 bp considered ‘PCR positive’. Samples with no visible amplified product were considered ‘PCR negative’. Unincorporated nucleotides and remaining primers were removed using Agencourt AMPure XP-PCR clean up (A63882, Beckman Coulter, Pasadena, CA), according to the manufacturer’s protocol. The DNA concentration of the eluted product was quantified using the KAPA Library Quantification Kit for Ilumina platform (KK4835, Kapabiosystems, Boston MA). PCR products were normalized to 10 nM and quantified again using the KAPA Library Quantification kit and pooled into equimolar 4 nM pools.

DNA extracted from sputum samples was analyzed in four runs on an Illumina MiSeq desktop sequencer (Illumina, San Diego, CA). Following cluster formation on the MiSeq instrument, the amplicons were sequenced using primers complimentary to the V4 region and designed for paired-ends sequencing. A third sequencing primer was used for reading the barcodes. To check for proper cluster density and sample normalization, a MiSeq single-end 26 bp +12 bp index sequencing run was performed using the MiSeq instrument. The pool was mixed with a PhiX library (Illumina, San Diego CA) at a ratio of

1:9 in order to increase the entropy of the library. A final MiSeq 2 x 150 bp + 12 bp index sequencing run was performed on the pooled samples.

Sequence Availability

Sequence data are deposited in NCBI's Sequence Read Archive.
Short Read Archive accession: SRP102629
https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?study= SRP102629
Bioproject accession: PRJNA377739
https://www.ncbi.nlm.nih.gov/bioproject/?term= PRJNA377739

16S rRNA gene sequence processing

First, reads were filtered to remove PhiXsequences. All reads mapping to Enterobacteria phage PhiX 174 reference genome (GenBank: NC_001422.1) using the software Bowtie[3] v1.0.1 were removed from the analysis. Remaining reads were merged using PEAR[4] v0.9.5-64, discarding all reads containing ambiguous bases (option '-u 0'). A paired-end read was discarded if one of the following conditions was met: overlap below 10 bp, assembly length <50 bp or p-value of alignment >0.01. Sequences were then processed with the QIIME[5] pipeline version 1.8.0. Sequences were assigned to samples on the basis of their MID tag, allowing for no base error. Chimeric sequences were identified and removed from the dataset using Usearch[6] version 6.1. The "closed-reference" QIIME protocol was used with the UCLUST method to select operational taxonomic units[7] (OTUs). Sequences with at least 97% identity were clustered together. A representative sequence from each cluster was used to identify bacterial taxa from the May 2013 edition of the Greengenes 16S rRNA sequence database[8-10] (v13_8).

Overview of the sequence processing pipeline commands

Illumina sequence files:

Sample_R1.fastq (150 bp)
Sample_R2.fastq (150 bp)
IndexRead.fastq (12 bp)

Metadata file:

Mapping_file.txt

```
# Use bowtie to identify the sequences corresponding to PhiX spike-in
$ bowtie --best --tryhard --suppress 2,3,4,5,6,7,8../phiX_ NC_001422.1_index
        Sample_R1.fastq R1_bowtie_phiX_alignment.txt

# Remove PhiX sequences from files
$ filter_fasta.py
        -s R1_bowtie_phiX_alignment.txt
        -f Sample_R1.fastq
        -o phiX_Removed_Sample_R1.fastq
        -n

# Use pear to merge the paired-end runs using overlapping sequence
$ pear -u 0 -m 50 -p 0.01
        -f phiX_Removed_Sample_R1.fastq
        -r phiX_Removed_Sample_R2.fastq
        -o Merged_phiX_Removed

# Convert fastq files into fasta file with barcode information from metadata file
$ split_libraries_fastq.py
        -i MergedPhiXremoved.assembled.fastq
        -b IndexRead.fastq
        -m Mapping_file.txt
        -o Library_Output/

# Identify chimeric sequences using the usearch61 method
$ identify_chimeric_seqs.py
        -m usearch61
        -i seqs.fna
        -r ../gg_13_8_otus/rep_set/97_otus.fasta
        -o usearch61_chimera_check/

# Remove sequences identified as chimeras
$ filter_fasta.py
        -s usearch61_chimera_check/chimeras.txt
        -f seqs.fna
        -o seqs_no_chimeras.fna
        -n

# Perform closed-reference alignment and OTU picking at 97% similarity
$ pick_closed_reference_otus.py
```

-i seqs_no_chimeras.fna
-r ../gg_13_8_otus/rep_set/97_otus.fasta
-t ../gg_13_8_otus/taxonomy/97_otu_taxonomy.txt
-o ClosedRef_13_8_97otus_v18/

Statistical analyses of 16S rRNA gene data

Rarefactions, alpha diversity (within sample evenness – Shannon diversity index), and beta diversity (differences in taxa between samples – UniFrac distance) calculations were all performed with the same QIIME pipeline. Samples were rarefied to 30,419 reads, which corresponded to the minimum number of aligned reads to a sample passing quality standards. The Shannon diversity index or weighted and unweighted UniFrac distances[11] were computed at this rarefaction level. For alpha diversity, 100 rarefactions were simulated and alpha diversity indices computed. For each sample, the reported alpha diversity values were obtained by computed the average value of the 100 simulations.

Statistical analyses were performed using QIIME[5] or the 'R' language and environment (version 3.3.2). Comparisons of bacterial relative abundances were performed at the phylum and genus level. Only taxa with at least a 1% average abundance across all samples were compared in any statistical test. Corrections for on multiple testing were performed by the Benjamini-Hochberg false discovery rate procedure. Normality was tested with each dataset with the Shapiro-Wilk test. While UniFrac distances could be assumed to normal, relative abundance and Shannon diversity index could not.

Comparisons of Shannon diversity and relative abundance between non-longitudinal groups were performed with Mann-Whitney or Kruskal-Wallis tests (after averaging repeated measures within a subject to a single value) and comparisons of UniFrac distance were performed with a two-way ANOVA followed by the Tukey honestly significant difference (HSD) method to correct for multiple comparisons. Longitudinal comparisons between stable and exacerbation samples were performed with a linear mixed-effects model (using the "nlme" package in R), where the subject was included as a random effect, or a paired Student's t-test for matched samples within a subject where a previous stable sample was available at minimum of 1 week and a maximum of 6

months prior to the exacerbation sample. We assume that multiple exacerbations within an individual are independent.

Classifying Exacerbation and COPD subtypes

COPD exacerbation subtypes were classified using previously defined criteria[12] of (a) Bacterial, if at least one positive potentially-pathogenic bacteria culture from sputum (*Haemophilus influenzae*, *Moraxella catarrhalis*, *Streptococcus pneumoniae*, *Pseudomonas aeruginosa*, or *Staphylococcus aureus)* (b) Viral, if at least one positive viral PCR (HRV, RSV, influenza virus, parainfluenza virus, human metapneumovirus, adenovirus, human bocavirus, or coronavirus) from sputum (c) Eosinophilic, if eosinophils are greater than 3% of nonsquamous cells from sputum, or mixed states (when appropriate).

COPD severity was classified by airflow obstruction by FEV1 as a percentage of predicted at enrollment with 1) Moderate = 50-79%, 2) Severe = 30-49%, and 3) Very Severe < 30%.

Bronchiectasis status was determined by clinician's diagnosis via CT scan.

Markov chain analysis

In the Markov chain analysis, exacerbation states were defined using the same criteria (inclusive of mixed state, i.e. an exacerbation with bacterial culture and viral PCR would appear in both the bacterial-positive and viral-positive Markov states) or modifications as listed. Exacerbations with missing data were excluded from the relevant model and transitions were included only if the temporally adjacent individual exacerbation could be classified for that model type. Transition probabilities were calculated by counting the relative frequency of observed transitions between temporally adjacent exacerbations within an individual for all possible transitions to other states. Overall (expected) frequencies were determined by the proportion of all exacerbations classified in that state (independent of whether they could be paired with another exacerbation in the model). Differences in Markov chain transition frequencies were tested between observed frequencies and expected independent frequencies from incidence of each

phenotype with a chi-square test and comparisons of frequencies between nodes were tested with a Fisher's exact test.

Supplementary Table Legends

Supplementary Table S1- Basic information about individuals in the microbiome cohort

Supplementary Table S2- Genus-level bacterial relative abundances for all samples

Supplementary Table S3- Association of UniFrac distance vs exacerbation frequency

Supplementary Table S4- Frequency of exacerbation events with Markov chain model

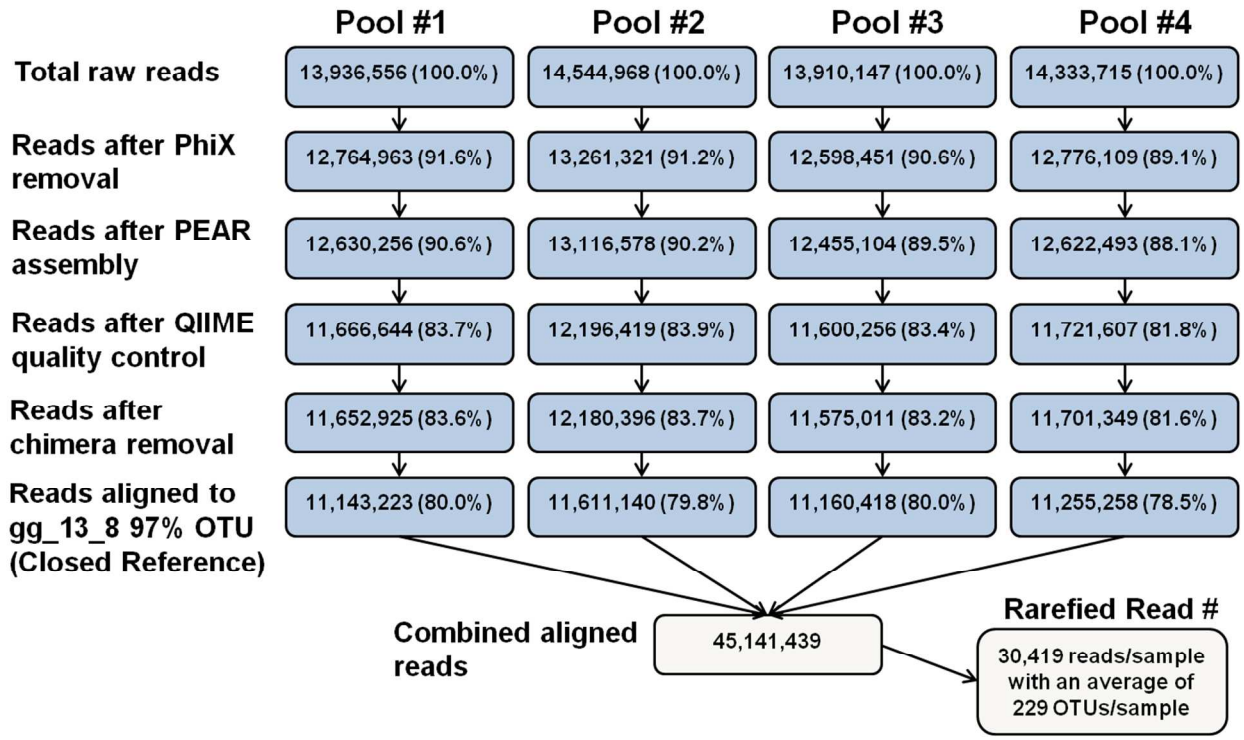

Figure S1: Flowchart for the analysis of 16S rRNA gene sequencing reads Paired-end 2x150 bp reads (+12 bp indexing read) for this study were sequenced across four MiSeq flow cells, represented by columns. The number of paired reads after each filtering, quality control, or alignment steps are shown in each box and the percentage of remaining reads relative to the starting read count are shown in parentheses. Reads were filtered to remove PhiX spike-in material, assembled into a single read spanning the V4 region with PEAR, processed with the QIIME quality control metrics, filtered for chimeric sequences, clustered and aligned to the Greengenes 16S rRNA reference.

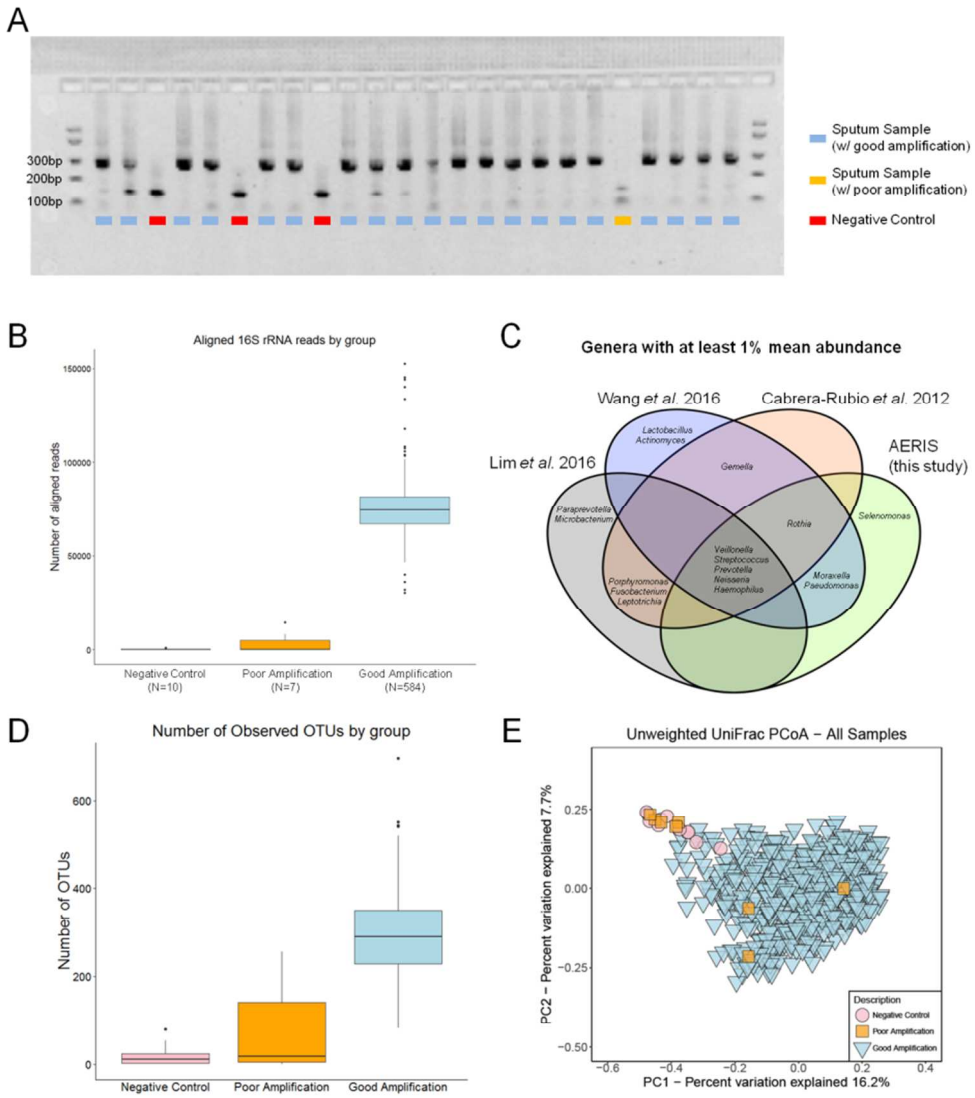

Figure S2: Experimental controls and checks for contaminating sequences (A) Representative PCR of the extracted DNA from sputum samples included negative controls (shown in red) which gave overwhelmingly primer-dimer products relative to the expected V4 product at 285 bp. Several sputum samples with lower DNA concentrations did not give robust PCR products (shown in orange) compared to higher quality samples (shown in blue). (B) After processing in the QIIME pipeline negative controls and poorly amplified samples showed demonstrably fewer aligned reads than the well-amplified samples. (C) The bacterial genera with a minimum of 1% average abundance across the study shows concordance with genera identified by other lung microbiome studies[13-15] examining sputum as a source of bacterial DNA. (D) Negative control samples showed lower diversity with fewer OTUs per sample. (E) Principal Coordinate Analysis (PCoA) of Unweighted UniFrac distances of the OTU table before rarefaction shows the negative controls have similar compositions distinct from the sputum samples with good amplification.

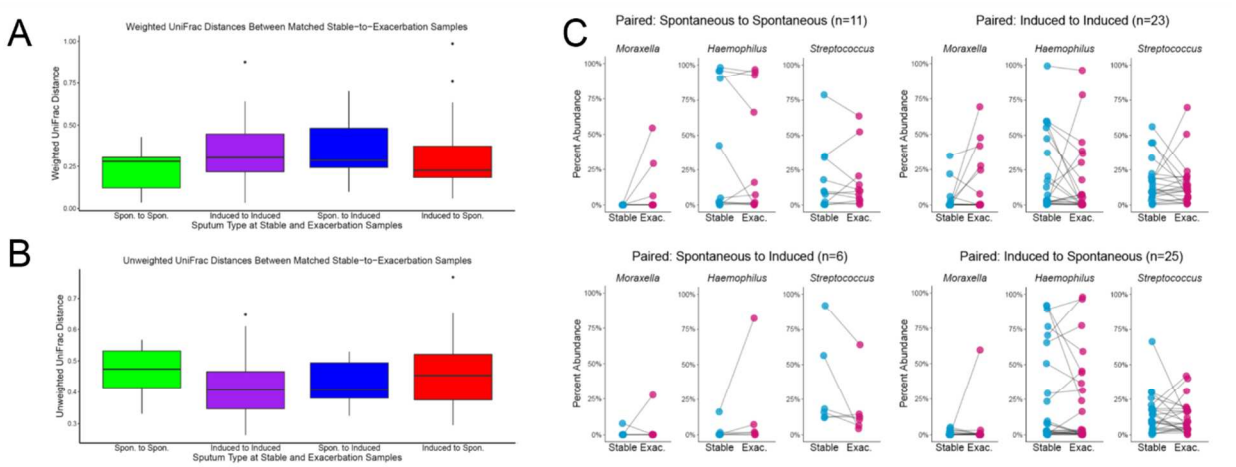

Figure S3: Different sputum types do not show systematic bacterial composition differences within individuals (A) The weighted UniFrac distances between longitudinal stable-exacerbation pairs did not show a significant difference between samples with different sputum types (four possible sputum combinations: spontaneous-to-spontaneous, induced-to-induced, spontaneous-to-induced, and induced-to-spontaneous); p=0.40 (ANOVA). (B) The unweighted UniFrac distances were also not significantly different for the same comparisons; p=0.63 (ANOVA). (C) Longitudinal changes in the relative abundances of the individual genera of Moraxella, Haemophilus, and Streptococcus are not unique to one sputum type pair.

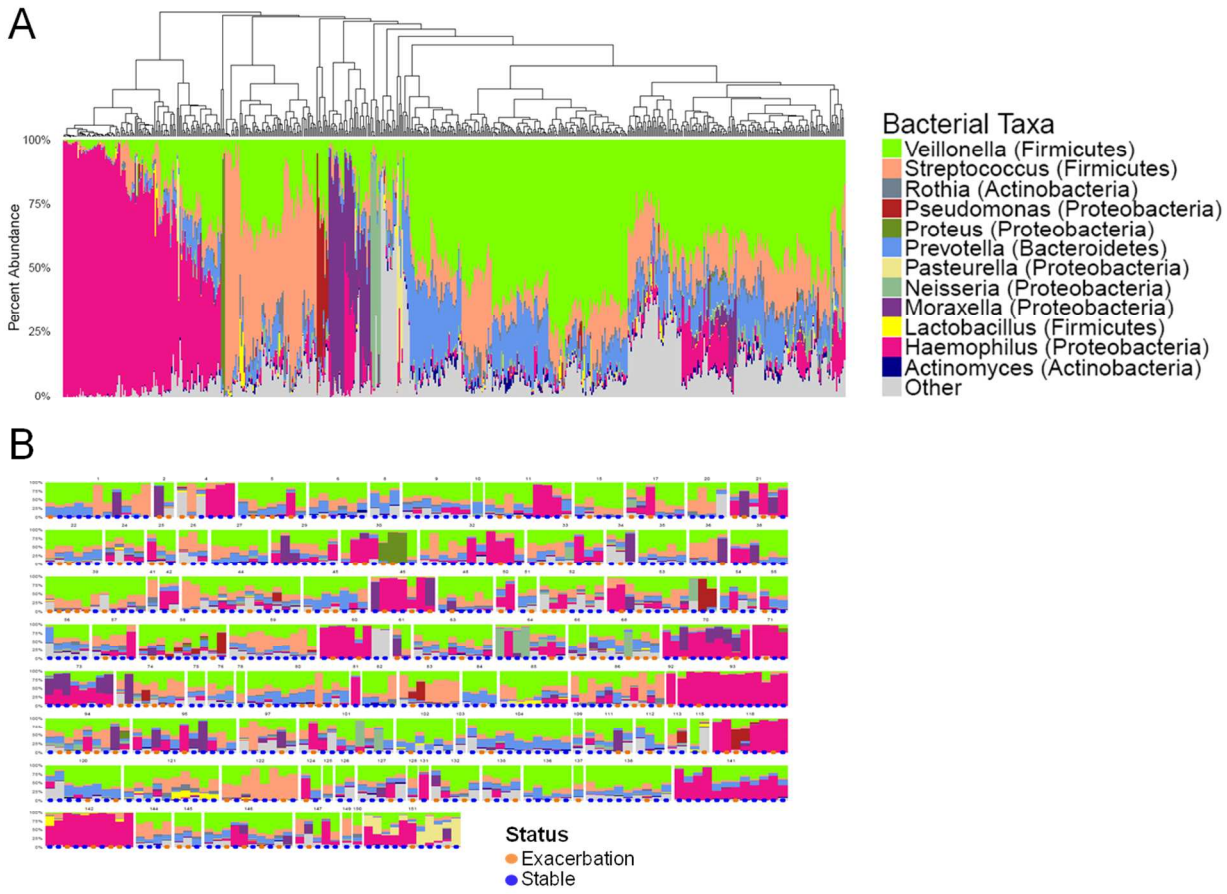

Figure S4: Abundance and correlations of lung bacterial taxa (A) The taxonomic overview of the relative abundance of taxa at the genus level for all samples identifies commonly reported lung microbiota. (B) The genus-level abundances are grouped by individual within each block and are ordered chronologically from left to right by collection date. The status of a sample is designated by the blue (stable) or orange (exacerbation) circle at the bottom of each bar.

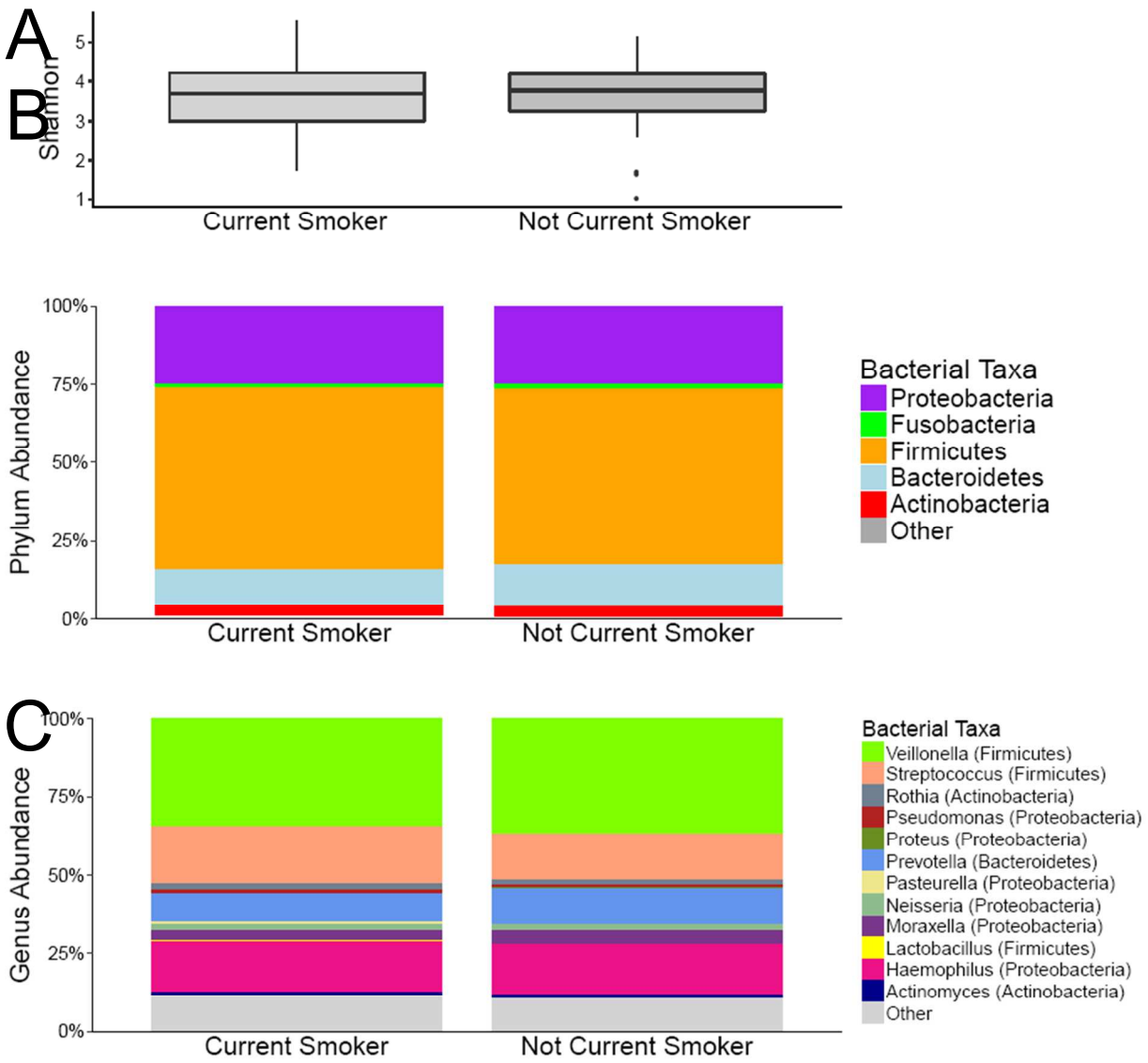

Figure S5: Diversity and composition of lung microbiome between smokers and non-smokers (A) The Shannon diversity index did not show a significant difference between smokers and non-smokers; $p>0.05$ (Mann-Whitney). (B) The phylum-level abundances and (C) genus-level abundances both showed no significant differences between the two groups.

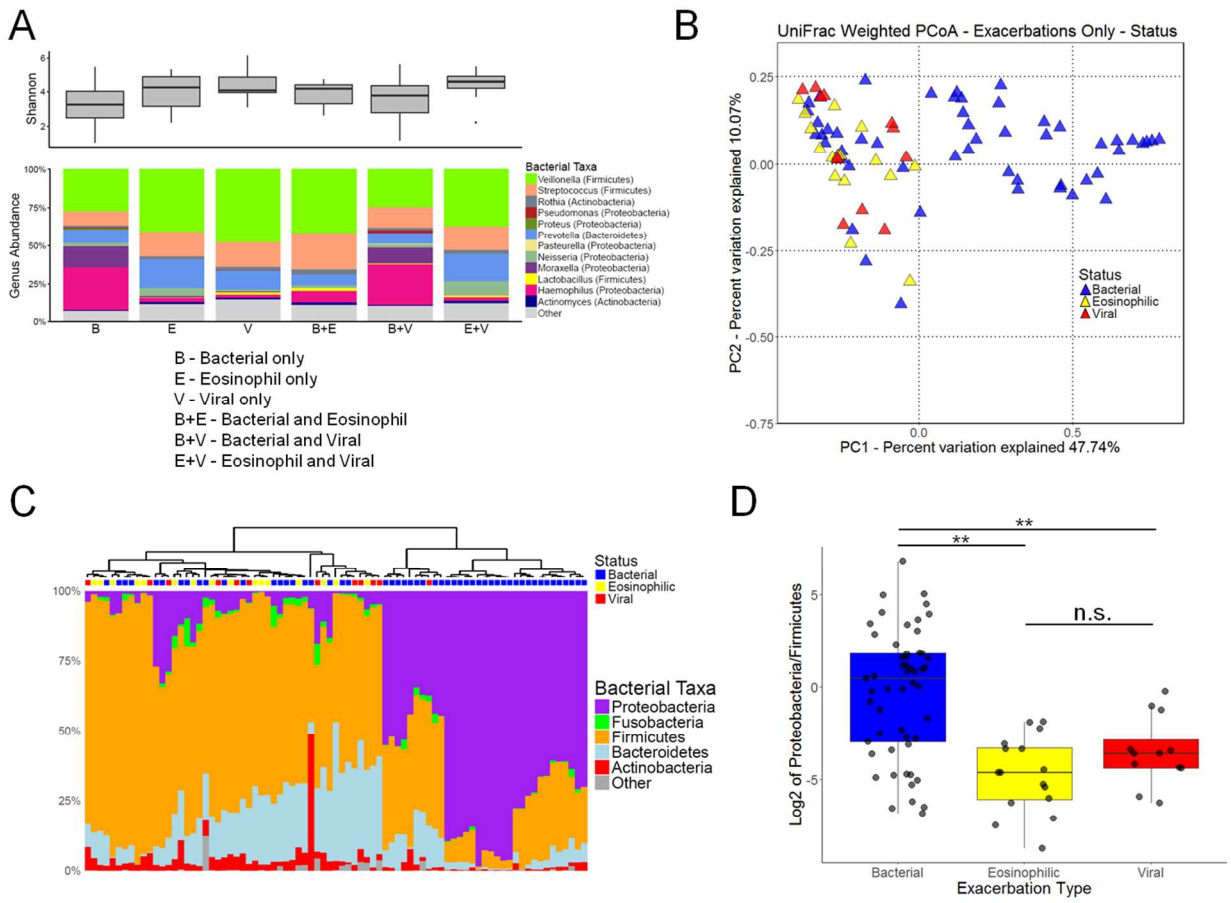

Figure S6: Bacterial exacerbations have different lung microbiome compositions compared to viral and eosinophilic exacerbations (A) The relative abundances of key genera in lung microbiome in exacerbation states grouped by the classification of the exacerbation type. Because bacterial exacerbations are defined by culture detection of potentially-pathogenic bacteria from sputum, this classification is not independent of the microbiome composition. (B) Principal Coordinate Analysis (PCoA) of weighted UniFrac distances shows exacerbations with only high bacteria (blue) are distinct microbiome compositions compared with only high eosinophils (yellow) or only viral (red). (C) When grouped by phylum-level relative abundances these three types of exacerbations differ in their relative abundances of Proteobacteria and Firmicutes. (D) Exacerbations classified as bacterial-dominant have higher relative abundances of Proteobacteria, while eosinophilic-dominant and viral exacerbations show increased relative abundances of Firmicutes; **p<0.001 (two-tailed Student's t-test), n.s. not significant.

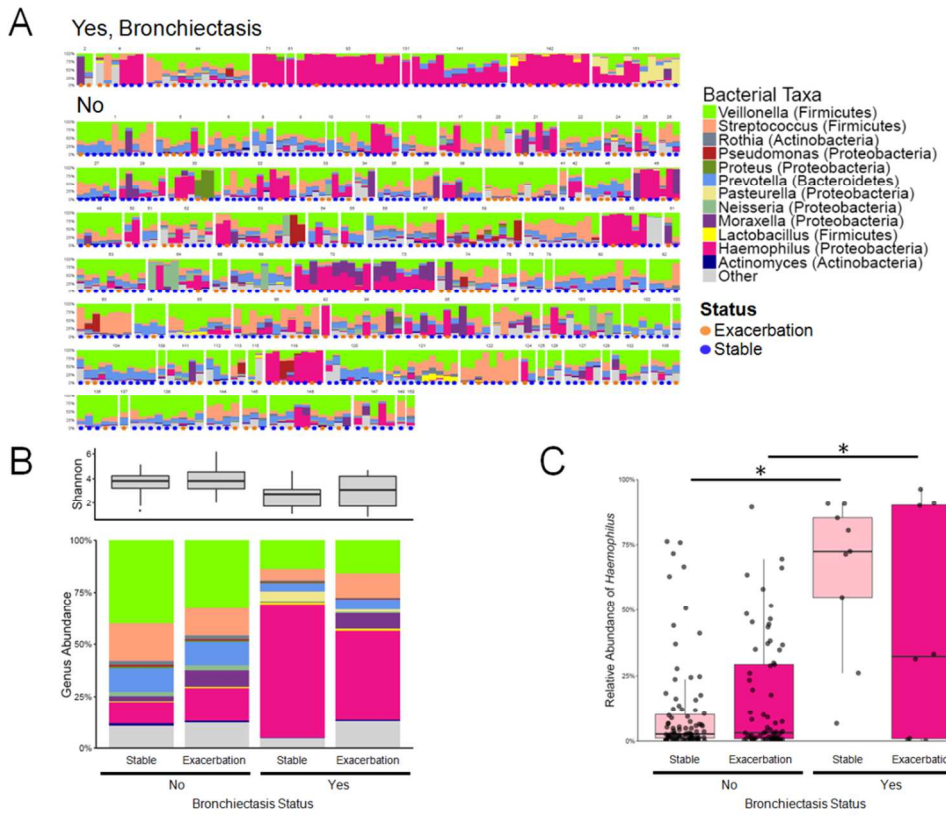

Figure S7: COPD subjects with bronchiectasis have high levels of *Haemophilus* in stable and exacerbation events (A) The genus-level taxonomic summary of lung microbiome samples separated by bronchiectasis status (based on chest CT examination). Relative abundance bars are grouped by samples from each individual within each block and are ordered chronologically from left to right. The status of a sample is designated by the blue (stable) or orange (exacerbation) circle at the bottom of each bar. (B) Individuals with bronchiectasis show different lung microbiome profiles. (C) Specifically, individuals with bronchiectasis show significant increases in the relative abundance of the genus *Haemophilus* in both stable and exacerbation states compared to individuals without bronchiectasis; *p<0.01 (Kruskal-Wallis test).

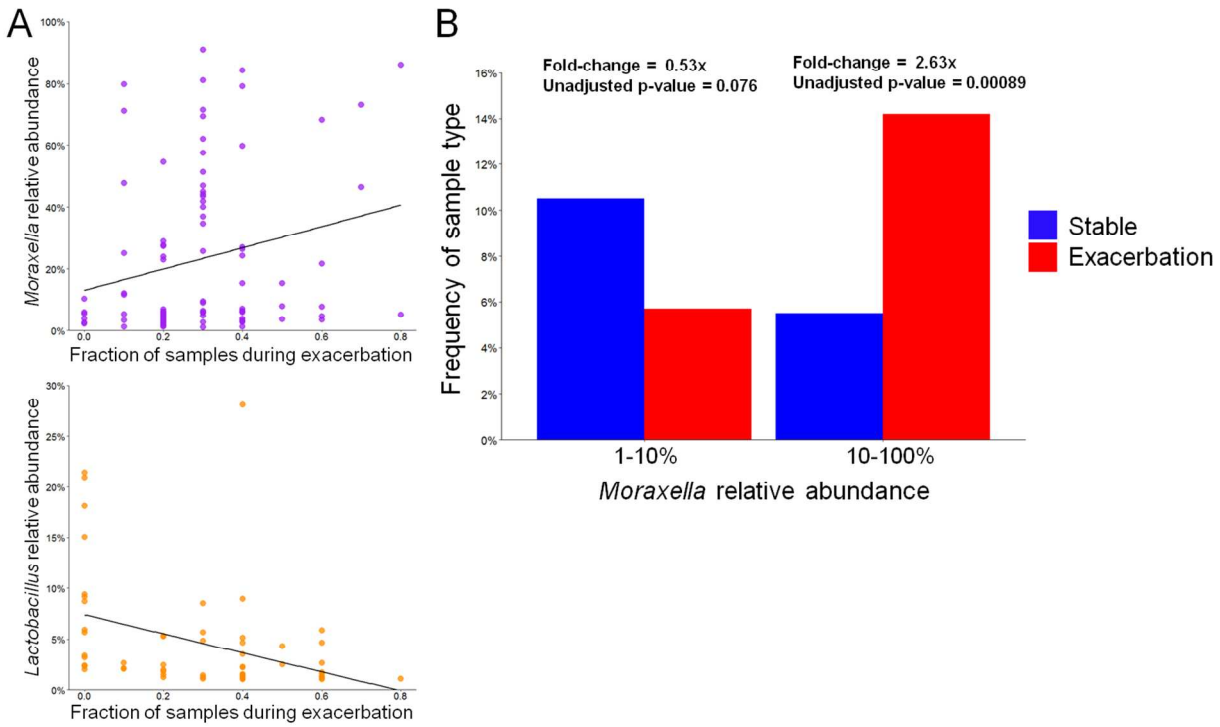

Figure S8: *Moraxella* and *Lactobacillus* abundances have opposite correlations with exacerbation frequencies (A) Linear regression of the relative abundances of genus-level taxonomies relative to the proportion exacerbation samples measured for that subject identified *Moraxella* as the genus with the highest positive correlation with exacerbation frequency (R=0.23, p=0.016, Pearson) and *Lactobacillus* as the genus with most negative correlation with exacerbation frequency (R=-0.37, p=0.02, Pearson). (B) A relative abundance of *Moraxella* of more than 10% increased the chance of exacerbation by a factor of 2.6 (95% CI 2.1 to 38.1) (*Moraxella* abundance >10% in 5.4% of stable samples versus 14.3% of exacerbation samples).

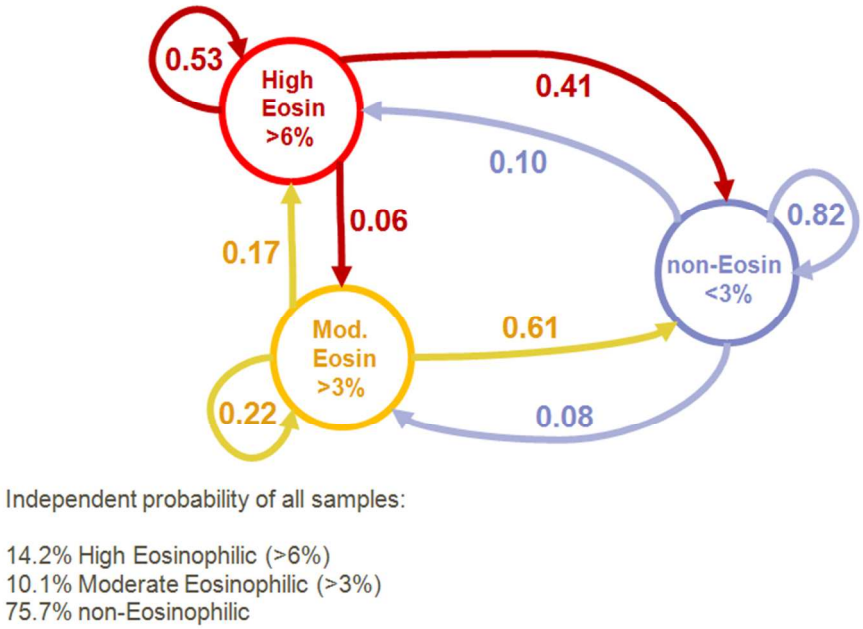

Figure S9: Extended Markov chain analysis of eosinophilic exacerbations The Markov chain analysis of eosinophilic exacerbations was extended to included two eosinophilic-positive states of high-eosinophil (>6% from sputum) and moderate-eosinophil (>3% and <6% from sputum), while non-eosinophilic exacerbations were defined as those with <3% from sputum. High-eosinophil exacerbations show a significantly increased probability of repeating the same phenotype in their next exacerbation and not transitioning to other states (p=0.02, Fisher's exact test).

References for online supplement

1. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci USA* 2011;**108** Suppl 1:4516-22.
2. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* 2012;**6**:1621-4.
3. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 2009;**10**:R25.
4. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 2014;**30**:614-20.
5. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walter WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. QIIME allows analysis of high-throughput community sequencing data. *Nat Meth*, 2010;**7**:335-6.
6. Edgar, R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 2010;**26**:2460-1.
7. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 2011;27:2194-2200.
8. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microb* 2007;**73**:5261-7.
9. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Anderson GL. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microb* 2006;**72**:5069-72.
10. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* 2012;**6**:610-8.
11. Lozupone C, Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 2005;**71**:8228-35.
12. Bafadhel M, McKenna S, Terry S, Mistry V, Reid C, Haldar P, Kebadze T, Duvoix A, Lindblad K, Patel H, Rugman P, Dodson P, Jenkins M, Saunders M, Newbold P, Green RH, Venge P, Lomas DA, Barer MR, Johnston SL, Pavord ID, Brightling CE. Acute exacerbations of chronic obstructive pulmonary disease: identification of biological clusters and their biomarkers. *Am J Respir Crit Care Med* 2011;**184**:662-71.
13. Wang Z, Bafadhel M, Haldar K, Spivak A, Mayhew D, Miller BE, Tal-Singer R,

Johnston SL, Ramsheh MY, Barer MR, Brightling CE, Brown JR. Lung microbiome dynamics in COPD exacerbations. *Eur Resp J* 2016;**47**:1082-1092.
14. Cabrera-Rubio R, Garcia-Núñez M, Setó L, Antó JM, Moya A, Monso E, Mira, A. Microbiome diversity in the bronchial tracts of patients with chronic obstructive pulmonary disease. *J Clin Microbiol* 2012;**50**:3562-8.
15. Lim MY, Yoon HS, Rho M, Sung J, Song YM, Lee K, Ko G. Analysis of the association between host genetics, smoking, and sputum microbiota in healthy humans. *Sci Rep* 2016;**6**:2374.