

A Refined View of Airway Microbiome in Chronic Obstructive Pulmonary Disease at Species and Strain-levels

Zhang Wang^{1*}, Haiyue Liu^{2*}, Fengyan Wang^{3*}, Yuqiong Yang^{3*}, Xiaojuan Wang¹, Boxuan Chen¹, Martin R. Stampfli⁴, Hongwei Zhou², Wensheng Shu¹, Christopher E. Brightling⁵, Zhenyu Liang^{3,7}, Rongchang Chen^{6,3,7}

¹ Institute of Ecological Science, School of Life Science, South China Normal University, Guangzhou, Guangdong Province, China

² State Key Laboratory of Organ Failure Research, Microbiome Medicine Center, Division of Laboratory Medicine, Zhujiang Hospital, Southern Medical University, Guangzhou, Guangdong Province, China

³ State Key Laboratory of Respiratory Disease, National Clinical Research Center for Respiratory Disease, Guangzhou Institute of Respiratory Health, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, Guangdong Province, China

⁴ Department of Medicine, Firestone Institute of Respiratory Health at St. Joseph's Healthcare, McMaster University, Hamilton, ON, Canada

⁵ Institute for Lung Health, Leicester NIHR Biomedical Research Centre, Department of Respiratory Sciences, University of Leicester, Leicester, UK

⁶ Pulmonary and Critical Care Department, Shenzhen Institute of Respiratory Diseases, Shenzhen People's Hospital, Shenzhen, Guangdong Province, China

⁷ Co-corresponding authors

* These authors contributed equally to this work.

Corresponding authors:

Zhenyu Liang (490458234@qq.com) and Rongchang Chen (chenrc@vip.163.com)

Word counts:

Abstract: 250

Main text: 3,000

“Take-home” message (193 characters with spaces):

The species-level analysis using the ‘third-generation’ sequencing enabled a refined view of the airway microbiome and its relationship with clinical outcome and inflammatory phenotype in COPD.

Abstract

Little is known about the species and strain-level diversity of the airway microbiome, and its implication in chronic obstructive pulmonary disease (COPD).

Here we report the first comprehensive analysis of the COPD airway microbiome at species and strain-levels. The full-length 16S rRNA gene was sequenced from sputum in 98 stable COPD patients and 27 age-matched healthy controls, using the 'third-generation' Pacific Biosciences sequencing platform.

Individual species within the same genus exhibited reciprocal relationships with COPD and disease severity. Species dominant in health can be taken over by another species within the same genus in GOLD IV patients. Such turnover was also related to enhanced symptoms and exacerbation frequency. *Ralstonia mannitolilytica*, an opportunistic pathogen, was significantly increased in COPD frequent exacerbators. There were inflammatory phenotype-specific associations of microbiome at the species-level. One group of four pathogens including *Haemophilus influenzae* and *Moraxella catarrhalis*, were specifically associated with sputum mediators for neutrophilic inflammation. Another group of seven species, including *Tropheryma whippelii*, showed specific associations with mediators for eosinophilic inflammation. Strain-level detection uncovered three non-typeable *H. influenzae* strains PittEE, PittGG and 86-028NP in the airway microbiome, where PittGG and 86-028NP abundances may inversely predict eosinophilic inflammation. The full-length 16S data augmented the power of functional inference and led to the unique identification of butyrate-producing and nitrate reduction pathways as significantly depleted in COPD.

Our analysis uncovered substantial intra-genus heterogeneity in the airway microbiome associated with inflammatory phenotypes and could be of clinical importance, thus enabled a refined view of the airway microbiome in COPD.

67 Introduction

68 The airway microbiome in chronic obstructive pulmonary disease (COPD) has been
 69 well studied in the last decade. The airway microbiome differs between health and
 70 COPD[1-3], shifts during exacerbations[4-6], associates with airway inflammation[5, 7]
 71 and predicts 1-year mortality of hospitalized exacerbation patients[8], all suggesting
 72 the implication of airway microbiome in COPD pathogenesis. Despite advances, the
 73 precise role of airway microbiome in COPD remains incompletely understood. An
 74 important knowledge gap is that our current view of airway microbiome is limited at
 75 most to its composition at the genus-level, due to insufficient resolution of one or few
 76 hypervariable regions of 16S rRNA gene being sequenced in essentially all previous
 77 studies. In these studies, certain bacterial genera were often reported to be altered as
 78 a whole in disease and in relation to airway inflammation[5, 6]. However, from an
 79 ecological perspective, members of microbial community do not necessarily function
 80 according to their taxonomic groups, instead diversified species can act in ecological
 81 “guilds” that co-adapt to altered environment[9, 10]. Therefore, the aggregated
 82 genus-level associations can be spurious or even misleading due to violation of basic
 83 ecological concepts. The inadequate depth of taxonomic profiling limits not only the
 84 accuracy of ecological inferences but also the ability to identify key bacterial species
 85 to use in follow-up experimental studies.

86
 87 The recently advanced ‘third-generation’ sequencing technologies such as Pacific
 88 Biosciences (PacBio) and Nanopore is increasingly applied to microbiome studies[11,
 89 12]. By generating long reads that extend tens of thousands of nucleotides, they offer
 90 the promise of increased taxonomic resolution by sequencing the full-length of 16S
 91 rRNA gene[13]. In these applications, the 16S amplicon is circularized and read
 92 through multiple passes before circular consensus sequences (CCS) is reported,
 93 which greatly reduced the initial high error rate (~10%) of the long-read sequencing to
 94 that comparable to short-read sequencing (~0.5%)[14, 15]. Recent development of
 95 sophisticated denoising algorithms further enable accurate bacterial species
 96 identification at single-nucleotide resolution with near-zero error rate[16]. In some
 97 situations, strain-level identity can be further resolved utilizing information on the full
 98 complement of 16S rRNA gene alleles in bacterial genomes[16, 17].

99
 100 Here we report the first comprehensive analysis of airway microbiome in COPD at
 101 species-level using PacBio sequencing. We also attempted to resolve strain-level
 102 identity when possible. Our results showed that there was substantial intra-genus
 103 diversity and heterogeneity in the airway microbiome that was previously
 104 underappreciated, which was associated with patient clinical features and airway

105 inflammatory phenotypes.

106

107 **Methods**

108 **Subjects and samples**

109 Sputum samples of 98 stable COPD patients and 27 age-matched healthy controls
110 were collected in the First Affiliated Hospital of Guangzhou Medical University. The
111 study was approved by the ethics committee of the First Affiliated Hospital of
112 Guangzhou Medical University (No. 2017-22) and was registered in
113 www.clinicaltrials.gov (NCT 03240315). All COPD patients met the diagnostic criteria
114 according to GOLD guideline and were assessed for symptoms and exacerbation
115 frequency (Table 1). Patients with antibiotic usage within 4 weeks were excluded.
116 Induced sputum were obtained for all subjects and quality-controlled. A panel of 47
117 sputum mediators were measured in a subset of 59 patients using custom antibody
118 microarray[18]. Additional information on sequencing, reagent controls, qPCR, and
119 statistical analyses are provided in the supplementary document.

120

121 **PacBio sequencing and analysis**

122 Bacterial genomic DNA was extracted from selected sputum plugs using Qiagen DNA
123 Mini kit. Negative controls for extraction and PCR were sequenced with all samples.
124 The full-length 16S rRNA gene was amplified using barcoded 27F and 1492R primers
125 and sequenced using PacBio Sequel. Circular consensus sequences (CCS) were
126 generated using the ccs application in SMRTLink 5.1 with minPasses=5 and
127 minPredictedAccuracy=0.90. The demultiplexed CCS were analyzed using DADA2
128 v1.12.1 recently customized for the PacBio full-length 16S sequencing data[16, 19].
129 Amplicon sequence variants (ASVs) were assigned to species only if they had unique,
130 100% identity match to a single species. Sequences were rarefied to 3,119 reads
131 (Figure S1).

132

133 **Strain-level identification**

134 Callahan et al. described a method for strain-level identification using full-length 16S
135 data leveraging the full complement of 16S rRNA alleles in bacterial genomes[16]. In
136 principle, a strain can be confidently assigned if all intra-genomic 16S sequence
137 variants of that strain are recovered in integral ratios according to its genuine allelic
138 variants. In extension to this approach, we designed a pipeline to assign strain-level
139 bins in three steps. 1) All species-level ASVs were BLASTn-searched against NCBI-nt
140 database. ASVs with 100% identity to the same bacterial genome were assigned to
141 the same initial bins. 2) The ASVs within each initial bin were subject to pairwise
142 Pearson correlation, to generate refined bins by identifying ASVs with co-occurrence

143 pattern (Pearson's $R > 0.7$). 3) For each refined bins, the copy number ratio of ASVs
144 were determined based on linear regression coefficient, and reconciled with the
145 genuine copy number ratio of the 16S alleles in the corresponding bacterial genomes.
146 The ASVs in integral copy number ratio with the genuine ratio were retained in the
147 final bins and assigned with strain-level taxonomy.

148

149 **Statistical analysis**

150 Differential microbiome features between COPD and controls were identified using
151 linear discriminant analysis (LDA) effect size (LEfSe) method with $LDA > 2.0$ [20].
152 Random forest analysis was performed using Weka with 7-fold cross-validation[21].
153 To identify microbiome-mediator associations independent of patient demographic
154 factors, all microbiome features and the 47 sputum mediators were first residualized
155 using a general linear model adjusting for covariates such as age, gender and
156 smoking history. An all-against-all correlation analysis was performed on the residues
157 of microbiome features and mediators using HALLA[22], and was subject to
158 unsupervised clustering. Co-occurrence analysis of microbiome was performed using
159 SparCC[23]. Functional inference of microbiome was performed using PICRUST2[24].
160 The false discovery rate (FDR) method was used to adjust P -values.

161

162 **Results**

163 **Overview of the species-level airway microbiome profile**

164 A total of 2,635,140 high-quality CCS reads were obtained for 98 stable COPD
165 patients and 27 controls (Table 1). The average number of passes on the 16S gene
166 was 34.9 for all CCS, equivalent to a low error rate of ~0.48% based on previous
167 sequencing runs on a mock community[17]. A total of 2,868 non-singleton ASVs were
168 identified, of which 795 ASVs (27.7%) were putatively assigned to 228 bacterial
169 species from 92 genera. Twenty species had an average relative abundance greater
170 than 0.005 (Table 2). The number of species capable of being detected increased by
171 3.26 folds compared to a re-analysis of all previous COPD airway microbiome studies
172 using the same pipeline (Table S1). *Streptococcus*, *Prevotella* and *Neisseria* had the
173 highest number species identified (Figure S2a). There was significant community shift
174 in COPD versus controls (Figure 1a-b, Adonis, $P = 0.004$). LEfSe analysis identified 11
175 discriminatory species between COPD and controls (Figure 1c, $LDA > 2.0$). Random
176 forest analysis using these 11 species yielded significantly increased precision in
177 classifying patients, compared to that using 9 discriminatory genera with the same
178 criteria ($LDA > 2.0$) (Figure 1d, Figure S2b, AUC: 0.787 versus 0.706, $P = 0.026$). Figure
179 1e showed an overview of species-level airway microbiome profile.

180

Overall there were no significant microbial community shifts between smokers and non-smokers within COPD patients or healthy controls, between patients with and without inhaled corticosteroid usage, and between frequent and non-frequent exacerbators (defined as exacerbation events ≥ 2 /last year, Figure S3a-b). Among species with relative abundance > 0.001 , *Haemophilus parahaemolyticus* was significantly increased in COPD smokers (Fold-change=6.40, FDR $P=0.02$, Figure S3c). *Ralstonia mannitolilytica*, an opportunistic pathogen, was significantly increased in frequent exacerbators (Fold-change=4.94, FDR $P=0.005$, Figure S3c). The increase of *R. mannitolilytica* was further confirmed by qPCR (Figure S3d).

Substantial intra-genus heterogeneity in the airway microbiome

Inspection of individual species revealed substantial intra-genus heterogeneity in their relationships with COPD. For example, while *Neisseria mucosa* was increased in COPD versus controls, its counterpart *Neisseria subflava* was significantly depleted (Figure 1a). The reciprocal relationships with COPD were also observed between *Haemophilus influenzae* and *Haemophilus parainfluenzae*, and between *Prevotella oris* and other *Prevotella* species (Figure S4). The species also altered differently with enhanced disease severity. For example, *H. parainfluenzae* and *N. subflava* were the most predominant species within the respective genera in healthy subjects, while *H. influenzae* and *N. meningitidis* took over and became over-dominant in GOLD IV patients (Figure 2a). Within *Streptococcus*, *Streptococcus salivarius* and *Streptococcus thermophilus* were most highly abundant in GOLD I patients, whereas *Streptococcus pseudopneumoniae* and *Streptococcus pneumoniae* became dominant in GOLD II and IV patients respectively (Figure 2a). Similar turnovers were also observed in patients classified using new GOLD classification scheme based on mMRC, CAT score and exacerbation frequency[25] (Figure S5). Opposite relationships with patient sputum neutrophilic levels were further observed between *H. influenzae* and *H. parainfluenzae* (Figure 2b), and between *Prevotella melaninogenica* and *Prevotella denticola* (Figure S6a). Individual species within the same genus exhibited disproportionately more co-exclusive than co-occurrence relationships (Figure 2c, Figure S6b), indicating ecological competition. qPCR using primers designed on species-specific genes showed concordance between the absolute count and relative abundance of *H. influenzae* and *H. parainfluenzae* (Figure 2d), as well as two other paired species within *Streptococcus* and *Prevotella* (Figure S6c), indicating accuracy of our approach in species quantification. These results suggested that there were substantial intra-genus heterogeneity resulting from interspecific competition in the airways.

Specific bacterial species were associated with neutrophilic or eosinophilic inflammation

To investigate how the intra-genus heterogeneity was related to airway inflammation, we performed an all-against-all correlation analysis between the species-level microbiome features and a panel of 47 sputum inflammatory mediators measured in a subset of 59 COPD patients. We used residualized correlation to identify microbiome-mediator correlations independent of patient demographic co-factors[22]. Unsupervised clustering based on the correlation profile revealed four clusters of bacterial species that each had distinct association patterns with three groups of mediators (Group 1-3, Figure 3). Four pathogens, *Moraxella catarrhalis*, *Pseudomonas aeruginosa*, *N. meningitidis* and *H. influenzae*, exhibited negative associations with a group of 11 mediators mostly Th2-related (i.e. IL-5, IL-13, CCL17), while they were positively correlated with a group of 21 mediators mostly Th1, Th17-related or pro-inflammatory (i.e. IL-8, IL-17, MMP-8), and had mixed relationships with the remaining mediators. By contrast, another seven species, *Prevotella aurantiaca*, *Fusobacterium nucleatum*, *Leptotrichia buccalis*, *Prevotella histicola*, *Porphyromonas gingivalis*, *N. mucosa* and *Tropheryma whippelii*, were specifically associated with increased Th2 mediators. Members of the two groups of mediators further showed specific correlations with increased sputum neutrophil or eosinophil percentages respectively (FDR $P < 0.05$), in agreement with their roles in neutrophilic or eosinophilic inflammation. Correspondingly, all seven species were increased in the eosinophilic COPD patients (eosinophil $> 3\%$, Figure S7a). The increase of *T. whippelii* was further confirmed by qPCR (Figure S7b). Such clustering patterns were however not observed at the genus-level (Figure S8), indicating microbiome associates with airway inflammatory phenotypes in a species-specific manner.

Strain-level identification of the airway microbiome

We further explored possible strain-level diversity in the airway microbiome. Recent studies showed that it is possible to resolve strain-level identity using full-length 16S sequences by leveraging the power of the full complement of 16S rRNA alleles within bacterial genomes[16, 17]. Using a set of stringent criteria (see methods, Figure S9), we were able to identify ASV bins corresponding to 10 bacterial strains (Table S2). For the first time, we identified three non-typeable *H. influenzae* (NTHi) strains PittEE, PittGG and 86-028NP in the airway microbiome, although the major allele of 86-028NP was not detected (Figure 4a-b). All three strains increased in COPD versus controls, and were associated with distinct groups of mediators (Figure 4c). Notably, 86-028NP and PittGG exhibited inverse associations with Th2 chemokines such as

257 CCL17 and CCL13 related to eosinophilic inflammation. qPCR using strain-specific
258 primers validated our results in PittEE and PittGG (Figure 4d), although the strain
259 detection rate by sequencing was lower than that using qPCR. qPCR for 86-028NP
260 yield positive but non-significant correlation (Figure 4d).

261

262 **A systematic evaluation of 16S sub-regions for airway microbiome profiling**

263 The full-length 16S sequences can serve as a benchmark for a systematic evaluation
264 on the performance of individual hypervariable regions for airway microbiome studies.
265 We created partitions of 16S sequences from the full-length data according to nine
266 hypervariable regions used in previous COPD microbiome studies, and analyzed
267 each partition separately. Among all sub-regions, V1V3 and V3V4 were the highest in
268 the number of species assigned as well as the proportion of sequences assigned to
269 species (Figure S10a). In addition, the V1V3 and V3V4 regions captured the greatest
270 microbial beta diversity measured using pairwise Bray-Curtis dissimilarity, whereas
271 the diversity was the lowest for V4 (Figure S10a). The V4 region was particularly poor
272 in classifying Proteobacteria and Actinobacteria, with 79.8% and 90.9% of sequences
273 from these two phyla unable to be assigned to species (Figure S10b). The V1V3
274 region also bear the highest similarity with the full-length data in microbial community
275 composition (Mantel test, Figure S10c).

276

277 **Full-length 16S sequences enhanced the power of functional inference**

278 PICRUSt is a useful tool to infer functional capacity of microbiome based on 16S
279 sequences[26]. PICRUSt analysis using the full-length 16S data enhanced the power
280 of functional inference by increasing the predicted pathway abundances by an
281 average 1.83 fold compared to individual sub-regions. Again, V1V3 were next best in
282 terms of the predicted pathway abundances (Figure S10a).

283

284 The augmented power of functional prediction led to the unique identification of 9
285 pathways as disease-associated using the full-length 16S data (Table S3). Of interest
286 are two pathways 'acetyl-CoA fermentation to butyrate' and 'nitrate reduction', both
287 inferred as significantly depleted in COPD (FDR $P < 0.05$, Figure 4e, Figure S11).
288 qPCR using validated broad-spectrum primers on
289 butyryl-CoA:acetate-CoA-transferase gene[27] in the butyrate pathway confirmed our
290 findings by showing 4.32 fold decrease of the gene in COPD versus controls (Table
291 S4). Furthermore, the two pathways showed inverse correlations with IL-17, which
292 were more pronounced when inferred from full-length 16S data than from sub-regions
293 (Figure 4f).

294

295 Discussion

296 Here we provided the first comprehensive insights on the COPD airway microbiome at
297 the species and strain-levels. By applying the 'third-generation' sequencing to the
298 full-length 16S rRNA gene, we uncovered diversity and complexity in the airway
299 microbiome at in-depth taxonomic levels that were previously underappreciated. In
300 light of our results, many aspects of our understanding of the COPD airway
301 microbiome need to be refined.

302

303 Our results showed that there were substantial intra-genus heterogeneity in the
304 airway microbiome in relation to patient clinical outcomes. Individual species within
305 the same genus often altered differentially in COPD and with enhanced clinical
306 severity and exacerbation frequency. The species predominant in healthy state can be
307 taken over by another species within the same genus in severe COPD patients.
308 Therefore, the genus-level associations reported in all previous airway microbiome
309 studies likely represent a weakened signal confounded by the mixed effects of
310 individual species within and should therefore be interpreted with caution.

311 Unsupervised cluster analysis identified two groups of bacterial species showing
312 specific associations with mediators related to neutrophilic or eosinophilic
313 inflammation respectively. The neutrophil-associated species included respiratory
314 pathogens like *H. influenzae* and *Moraxella catarrhalis*[28]. The eosinophil-associated
315 species included *T. whipplei*, a clinically important species reported to be implicated in
316 pneumonia[29], HIV infection[30] and eosinophilic, corticosteroid-resistant asthma[31,
317 32]. Such clustering pattern was non-existent at the genus-level. Hence the
318 species-level delineation enabled a more ecologically coherent view of airway
319 microbiome according to inflammatory phenotypes.

320

321 In extension to a previous approach[16], we detected three NTHi clinical strains
322 PittEE, PittGG and 86-028NP in the airway microbiome with reasonably high
323 confidence, based on which the strain-level heterogeneity was also observed in the
324 airways. All three strains were initially isolated from otitis media patients[33-35]. It has
325 been shown that the PittGG strain, by possessing an extra cluster of 339 genes and a
326 Hif-type pili structure, conveyed greater virulence than PittEE[34]. qPCR assays
327 based on *alpA* gene on this extra locus confirmed our results in PittGG quantification.
328 While all three strains were related to increased Th1/Th17 mediators, 86-028NP and
329 PittGG were further associated with decreased Th2-related CCL13 and CCL17,
330 indicating their abundances may negatively predict eosinophilic inflammation. We
331 realize that the strain-level diversity and detection rate remained relatively low, which
332 is a caveat due to inherently limited power of 16S sequences in strain-level resolution

333 and its sensitivity to potential sequencing errors.

334

335 We identified *Ralstonia mannitolilytica* as significantly increased in COPD patients
336 with frequent exacerbator phenotype. *R. mannitolilytica* is an opportunistic pathogen
337 that has been recovered from cystic fibrosis airways[36]. In a previous report, the
338 same species was isolated from one COPD exacerbation patient in western China
339 with extreme symptoms and acute respiratory failure[37]. *Ralstonia* spp. rarely cause
340 infection in healthy individuals but can be a severe pathogen especially in
341 immunosuppressed patients[38]. Therefore, the presence of *R. mannitolilytica* in
342 stable COPD patients may be an important contributing factor in predisposing patients
343 to recurrent infection and exacerbations.

344

345 The systematic comparison of 16S sub-regions indicated that V1V3 performed the
346 best in terms of microbial diversity and the power of functional inference. Our results
347 are consistent with the analysis by Johnston et al.[17], and should guide future studies
348 that sequencing V1V3 region may be a surrogate for the full-length 16S data.
349 Conversely, sequencing V4 alone, despite its wide usage in airway microbiome
350 studies, might not provide sufficient resolution for in-depth taxonomic profiling.

351

352 With augmented power in functional inference, we identified butyrate-producing and
353 nitrate reduction pathways as uniquely depleted in COPD using full-length 16S data.
354 Butyrate is a well-characterized microbial metabolite with anti-inflammatory effects[39],
355 and nitric oxide, the end product of nitrate reduction, may also have
356 disease-ameliorating role via suppressing NLRP3 inflammasome activation[40].
357 Functional validations are warranted to explore these microbial metabolites as novel
358 therapies for COPD.

359

360 The limitations of this study include its single-centered, cross-sectional design, the
361 relatively small group of healthy subjects, and the absence of sufficient data to explore
362 species-specific relationships with other etiological factors such as viral infections.
363 The species-level characterization on larger, longitudinal cohorts is necessary to
364 understand how species alter differentially during exacerbations and to treatment, the
365 temporal dynamics of ecological heterogeneity, and its underlying relationships with
366 airway inflammation and disease progression.

367

368 In summary, we reported the comprehensive landscape of COPD airway microbiome
369 at species and strain-levels. We showed there was substantial intra-genus
370 heterogeneity associated with patient clinical outcome and inflammatory phenotypes.

371 Sequencing the full-length 16S rRNA gene enabled a refined, ecologically coherent
372 view on the composition and function of the COPD airway microbiome, and should
373 see a wider applicability in airway microbiome studies in future.

374

375 **Acknowledgement**

376 This work was supported by the National Key R&D Program of China
377 (2017YFC1310600) funded to HZ and RC, and the National Natural Science
378 Foundation of China (31970112) funded to ZW.

379 **Figure legends**

380 **Figure 1. The overview of species-level profile of the airway microbiome in**
 381 **COPD patients and healthy controls. a)** Principal coordinate analysis based on
 382 weighted UniFrac distance on sputum samples from 98 COPD patients and 27
 383 healthy controls. **b)** The Shannon diversity and relative abundances of major genera
 384 (relative abundance>0.005) in COPD patients and healthy controls. **c)** The 11 top
 385 discriminatory species-level taxa between COPD and controls as identified from
 386 LEfSe analysis (LDA>2.0). **d)** The receiver operating characteristic curves for the
 387 Random Forest analyses using the 11 species-level and 9 genus-level discriminatory
 388 taxa (LDA>2.0) to segregate COPD patients from controls. **e)** The heatmap for the
 389 species-level microbiome profile. The major species-level taxa (relative
 390 abundance>0.001) within each genus in panel b) were shown. The fold change of
 391 each species (Sp) and its corresponding genus (Gn) in COPD patients versus
 392 controls were shown beside the taxonomy.

393
 394 **Figure 2. The intra-genus heterogeneity of the airway microbiome. a)** The
 395 alternation of major species in *Haemophilus*, *Neisseria*, *Streptococcus* and *Prevotella*
 396 between healthy controls and COPD patients with increasing disease severity based
 397 on GOLD classification (spirometry-based). The number of subjects in each subgroup
 398 was indicated in the parenthesis. **b)** The reciprocal relationship between *H. influenzae*
 399 and *H. parainfluenzae* with sputum neutrophilic percentage. **c)** More pervasive
 400 co-exclusive than co-occurrence relationships between major species in *Prevotella*,
 401 *Streptococcus*, *Neisseria* and *Haemophilus*. Only significant correlations were shown
 402 in the networks (SparCC, $P<0.05$). Co-exclusion relationships were in red, whereas
 403 co-occurrence relationships were in grey. **d)** qPCR assays using species-specific
 404 primers showed concordance between absolute counts and relative abundances of *H.*
 405 *influenzae* and *H. parainfluenzae*.

406
 407 **Figure 3. Species-specific association of airway microbiome with inflammatory**
 408 **phenotypes.** Unsupervised hierarchical cluster analysis on an all-against-all
 409 correlation profile between species-level microbiome features and 47 sputum
 410 mediators from a subset of 59 COPD patients (Ward's method). The species were
 411 shown if they had relative abundance>0.001 and were significantly associated with at
 412 least one of the 47 sputum mediators (HALLA, FDR $P<0.05$). The mediators were
 413 clustered into three groups and termed based on their classes and associations with
 414 airway eosinophils or neutrophils (Group 1: Th2-related, Group 2:
 415 Th1/Th17/Pro-inflammatory-related, Group 3: Others). The microbiome features were
 416 clustered into four groups based on their association patterns with the three groups of

417 mediators (termed “Pro-inflammatory”, “Neutrophilic”, “Eosinophilic” and
418 “Anti-inflammatory”). The significant associations were indicated in asterisks.
419 Significant positive and negative associations between sputum mediators and
420 neutrophilic and eosinophilic percentages were shown on bottom of the heatmap
421 (FDR $P < 0.05$).
422

423 **Figure 4. Strain-level identification and functional inference of the airway**
424 **microbiome. a)** The strong correlation pattern between pairs of ASVs assigned to the
425 strains PittEE, PittGG and 86-028NP (Pearson's $R > 0.93$). **b)** The copy number of the
426 highly-correlated ASVs are in integral ratio with the genuine allelic frequency of the
427 16S rRNA genes within the genome (PittEE 3:3, PittGG 5:1, 86-028NP 4:1:1),
428 supporting the assignment of the ASVs to the corresponding strains. The major 16S
429 allele of the 86-028NP strain was not detected. **c)** Significant associations between
430 the three *H. influenzae* strains with sputum mediators (Spearman, FDR $P < 0.05$). **d)**
431 qPCR results using strain-specific primers for the three *H. influenzae* strains in
432 relation to their relative abundances in the sequencing data. **e)** The abundances of
433 the two pathways ‘PWY-5676: acetyl-CoA fermentation to butyrate’ and ‘PWY-490-3:
434 nitrate reduction’ in COPD and healthy subjects as inferred from the full-length (V1V9)
435 and V1V3 data using PICRUST2 (** FDR $P < 0.01$, * FDR $P < 0.05$). **f)** The two pathways
436 showed negative correlations with IL-17, which was more pronounced when inferred
437 from full-length 16S sequences than V1V3 sequences.
438

Table 1. Major demographic and clinical characteristics of subjects.

Demographic and clinical features	Healthy (n=27)	COPD (n=98)	P-value
Age	65.4 (10.8)	66.2 (8.9)	0.67
Gender, n(M/F)	23/4	89/9	0.48
Current smoking, n(Y/N)	9/18	85/13	1.0e-4***
GOLD (1/2/3/4)	NA	24/33/32/9	NA
New GOLD (a/b/c/d) ^{\$}	NA	39/38/4/17	NA
Frequent exacerbator (Y/N) ^{\$\$}	NA	20/78	NA
ICS usage (Y/N)	NA	58/40	NA
pre-FEV ₁ (L)	2.8±0.1	1.5±0.1	5.1e-10***
pre-FVC (L)	3.4±0.2	2.9±0.1	0.01**
pre-FEV ₁ (%)	100.0±2.6	56.1±2.7	1.0e-10***
pre-FEV ₁ /FVC	0.81±0.01	0.49±0.02	2.5e-13***
post-FEV ₁ (L)	NA	1.6±0.7	NA
post-FVC (L)	NA	3.1±0.8	NA
post-FEV ₁ (%)	NA	59.6±2.6	NA
post-FEV ₁ /FVC	NA	0.5±0.1	NA
CAT score	NA	4.0±0.6	NA
mMRC	NA	0.4±0.1	NA
Total sputum cells (cells×10 ⁹ /L)	NA	20.4±2.8	NA
Sputum neutrophils (%)	NA	86.2±1.2	NA
Sputum eosinophils (%)	NA	5.3±0.7	NA
Sputum lymphocyte (%)	NA	0.7±0.1	NA
Sputum monocyte (%)	NA	7.9±1.1	NA

Continuous data are present as mean (range) or mean±SEM.

P-value was calculated using Fisher exact test for categorical variables and using Wilcoxon rank-sum test for continuous variables. *** $P < 0.001$; ** $P < 0.01$; * $P < 0.05$

ICS: inhaled corticosteroids; FEV₁: forced expiratory volume in one second; FVC: forced vital capacity, CAT: COPD Assessment Test, mMRC: modified Medical Research Council.

^{\$} The new GOLD classification based on mMRC, CAT and exacerbation frequency[25].

^{\$\$} The frequent exacerbator was defined as exacerbation event ≥ 2 /last year.

449 **Table 2.** The major species-level taxa identified in this study (average relative
450 abundance>0.005).

Species	All average	COPD average	Healthy average	Fold change	FDR P-value
<i>Prevotella intermedia</i>	0.005	0.006	0.005	1.191	0.319
<i>Prevotella melaninogenica</i>	0.060	0.057	0.071	0.801	0.342
<i>Prevotella pallens</i>	0.013	0.012	0.017	0.700	0.0140*
<i>Streptococcus pseudopneumoniae</i>	0.017	0.020	0.004	5.667	0.222
<i>Streptococcus salivarius</i>	0.013	0.015	0.004	3.664	0.277
<i>Streptococcus thermophilus</i>	0.027	0.030	0.014	2.228	0.809
<i>Haemophilus influenzae</i>	0.043	0.052	0.008	6.779	0.617
<i>Haemophilus parahaemolyticus</i>	0.009	0.010	0.004	2.564	0.204
<i>Haemophilus parainfluenzae</i>	0.012	0.009	0.020	0.474	0.0074**
<i>Neisseria meningitidis</i>	0.009	0.009	0.008	1.044	0.128
<i>Neisseria mucosa</i>	0.007	0.008	0.002	4.452	0.785
<i>Neisseria perflava</i>	0.011	0.012	0.005	2.391	0.742
<i>Neisseria subflava</i>	0.017	0.012	0.036	0.330	0.0415*
<i>Fusobacterium nucleatum</i>	0.016	0.013	0.027	0.495	0.139
<i>Fusobacterium periodonticum</i>	0.014	0.012	0.022	0.531	0.0199*
<i>Pseudomonas aeruginosa</i>	0.016	0.020	0.000	NA	0.105
<i>Moraxella catarrhalis</i>	0.030	0.038	0.001	28.947	0.275
<i>Veillonella parvula</i>	0.005	0.004	0.010	0.439	0.208
<i>Porphyromonas gingivalis</i>	0.010	0.009	0.015	0.570	0.282
<i>Campylobacter concisus</i>	0.006	0.006	0.007	0.902	0.233

451 P-value was calculated using Wilcoxon rank-sum test. *** FDR $P < 0.001$; ** $P < 0.01$; * $P < 0.05$

452

453

454 **References**

- 455 1. Pragman AA, Kim HB, Reilly CS, Wendt C, Isaacson RE. The lung microbiome in
456 moderate and severe chronic obstructive pulmonary disease. *PLoS One* 2012; 7(10): e47305.
- 457 2. Einarsson GG, Comer DM, McIlreavey L, Parkhill J, Ennis M, Tunney MM, Elborn JS.
458 Community dynamics and the lower airway microbiota in stable chronic obstructive pulmonary
459 disease, smokers and healthy non-smokers. *Thorax* 2016; 71(9): 795-803.
- 460 3. Zakharkina T, Heinzl E, Koczulla RA, Greulich T, Rentz K, Pauling JK, Baumbach J,
461 Hermann M, Grunewald C, Dienemann H, von Muller L, Bals R. Analysis of the airway
462 microbiota of healthy individuals and patients with chronic obstructive pulmonary disease by
463 T-RFLP and clone sequencing. *PLoS One* 2013; 8(7): e68302.
- 464 4. Huang YJ, Sethi S, Murphy T, Nariya S, Boushey HA, Lynch SV. Airway microbiome
465 dynamics in exacerbations of chronic obstructive pulmonary disease. *J Clin Microbiol* 2014;
466 52(8): 2813-2823.
- 467 5. Wang Z, Bafadhel M, Halder K, Spivak A, Mayhew D, Miller BE, Tal-Singer R, Johnston
468 SL, Ramsheh MY, Barer MR, Brightling CE, Brown JR. Lung microbiome dynamics in COPD
469 exacerbations. *Eur Respir J* 2016; 47(4): 1082-1092.
- 470 6. Wang Z, Singh R, Miller BE, Tal-Singer R, Van Horn S, Tomsho L, Mackay A, Allinson JP,
471 Webb AJ, Brookes AJ, George LM, Barker B, Kolsum U, Donnelly LE, Belchamber K, Barnes
472 PJ, Singh D, Brightling CE, Donaldson GC, Wedzicha JA, Brown JR, Copdmap. Sputum
473 microbiome temporal variability and dysbiosis in chronic obstructive pulmonary disease
474 exacerbations: an analysis of the COPDMAP study. *Thorax* 2018; 73(4): 331-338.
- 475 7. Wang Z, Maschera B, Lea S, Kolsum U, Michalovich D, Van Horn S, Traini C, Brown JR,

476 Hessel EM, Singh D. Airway host-microbiome interactions in chronic obstructive pulmonary
477 disease. *Respiratory research* 2019; 20(1): 113.

478 8. Leitaio Filho FS, Alotaibi NM, Ngan D, Tam S, Yang J, Hollander Z, Chen V, FitzGerald
479 JM, Nislow C, Leung JM, Man SFP, Sin DD. Sputum Microbiome is Associated with 1-Year
480 Mortality Following COPD Hospitalizations. *Am J Respir Crit Care Med* 2018.

481 9. Zander CD. The guild as a concept and a means in ecological parasitology. *Parasitol Res*
482 2001; 87(6): 484-488.

483 10. Zhao L, Zhang F, Ding X, Wu G, Lam YY, Wang X, Fu H, Xue X, Lu C, Ma J, Yu L, Xu C,
484 Ren Z, Xu Y, Xu S, Shen H, Zhu X, Shi Y, Shen Q, Dong W, Liu R, Ling Y, Zeng Y, Wang X,
485 Zhang Q, Wang J, Wang L, Wu Y, Zeng B, Wei H, Zhang M, Peng Y, Zhang C. Gut bacteria
486 selectively promoted by dietary fibers alleviate type 2 diabetes. *Science* 2018; 359(6380):
487 1151-1156.

488 11. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation
489 sequencing technologies. *Nat Rev Genet* 2016; 17(6): 333-351.

490 12. Levy SE, Myers RM. Advancements in Next-Generation Sequencing. *Annu Rev*
491 *Genomics Hum Genet* 2016; 17: 95-115.

492 13. Wagner J, Coupland P, Browne HP, Lawley TD, Francis SC, Parkhill J. Evaluation of
493 PacBio sequencing for full-length bacterial 16S rRNA gene classification. *BMC Microbiol* 2016;
494 16(1): 274.

495 14. Hebert PDN, Braukmann TWA, Prosser SWJ, Ratnasingham S, deWaard JR, Ivanova
496 NV, Janzen DH, Hallwachs W, Naik S, Sones JE, Zakharov EV. A Sequel to Sanger: amplicon
497 sequencing that scales. *BMC Genomics* 2018; 19(1): 219.

498 15. Jiao X, Zheng X, Ma L, Kutty G, Gogineni E, Sun Q, Sherman BT, Hu X, Jones K, Raley C,
499 Tran B, Munroe DJ, Stephens R, Liang D, Imamichi T, Kovacs JA, Lempicki RA, Huang DW. A
500 Benchmark Study on Error Assessment and Quality Control of CCS Reads Derived from the
501 PacBio RS. *J Data Mining Genomics Proteomics* 2013: 4(3).

502 16. Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS, McGill SK, Dougherty MK.
503 High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide
504 resolution. *Nucleic Acids Res* 2019: 47(18): e103.

505 17. Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, Leopold SR,
506 Hanson BM, Agresta HO, Gerstein M, Sodergren E, Weinstock GM. Evaluation of 16S rRNA
507 gene sequencing for species and strain-level microbiome analysis. *Nat Commun* 2019: 10(1):
508 5029.

509 18. Wang F, Liang Z, Yang Y, Zhou L, Guan L, Wu W, Jiang M, Shi W, Deng K, Chen J, Chen
510 R. Reproducibility of fluid-phase measurements in PBS-treated sputum supernatant of healthy
511 and stable COPD subjects. *Int J Chron Obstruct Pulmon Dis* 2019: 14: 835-852.

512 19. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2:
513 High-resolution sample inference from Illumina amplicon data. *Nat Methods* 2016: 13(7):
514 581-583.

515 20. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C.
516 Metagenomic biomarker discovery and explanation. *Genome Biol* 2011: 12(6): R60.

517 21. Frank E, Hall M, Witten I. The WEKA Workbench. Online Appendix for "Data Mining:
518 Practical Machine Learning Tools and Techniques". *Morgan Kaufmann, Fourth Edition* 2016.

519 22. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW,

520 Andrews E, Ajami NJ, Bonham KS, Brislawn CJ, Casero D, Courtney H, Gonzalez A, Graeber
521 TG, Hall AB, Lake K, Landers CJ, Mallick H, Plichta DR, Prasad M, Rahnavard G, Sauk J,
522 Shungin D, Vazquez-Baeza Y, White RA, 3rd, Investigators I, Braun J, Denson LA, Jansson
523 JK, Knight R, Kugathasan S, McGovern DPB, Petrosino JF, Stappenbeck TS, Winter HS,
524 Clish CB, Franzosa EA, Vlamakis H, Xavier RJ, Huttenhower C. Multi-omics of the gut
525 microbial ecosystem in inflammatory bowel diseases. *Nature* 2019; 569(7758): 655-662.

526 23. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. *PLoS*
527 *Comput Biol* 2012; 8(9): e1002687.

528 24. Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC,
529 Burkepile DE, Vega Thurber RL, Knight R, Beiko RG, Huttenhower C. Predictive functional
530 profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol*
531 2013; 31(9): 814-821.

532 25. Global Initiative for Chronic Obstructive Lung Disease. Global Strategy For the Diagnosis,
533 Management, and Prevention of Chronic Obstructive Pulmonary Disease. 2019 Report. 2019.

534 26. Douglas GM, Maffei VJ, Zaneveld J, Yurgel SN, Brown JR, Taylor CM, Huttenhower C,
535 Langille MG. PICRUSt2: An improved and extensible approach for metagenome inference.
536 *bioRxiv* 2019; <https://doi.org/10.1101/672295>.

537 27. Vital M, Penton CR, Wang Q, Young VB, Antonopoulos DA, Sogin ML, Morrison HG,
538 Raffals L, Chang EB, Huffnagle GB, Schmidt TM, Cole JR, Tiedje JM. A gene-targeted
539 approach to investigate the intestinal butyrate-producing bacterial community. *Microbiome*
540 2013; 1(1): 8.

541 28. Sethi S, Murphy TF. Bacterial infection in chronic obstructive pulmonary disease in 2000:

542 a state-of-the-art review. *Clin Microbiol Rev* 2001; 14(2): 336-363.

543 29. Bousbia S, Papazian L, Auffray JP, Fenollar F, Martin C, Li W, Chiche L, La Scola B,
544 Raoult D. Tropheryma whipplei in patients with pneumonia. *Emerg Infect Dis* 2010; 16(2):
545 258-263.

546 30. Lozupone C, Cota-Gomez A, Palmer BE, Linderman DJ, Charlson ES, Sodergren E,
547 Mitreva M, Abubucker S, Martin J, Yao G, Campbell TB, Flores SC, Ackerman G, Stombaugh
548 J, Ursell L, Beck JM, Curtis JL, Young VB, Lynch SV, Huang L, Weinstock GM, Knox KS,
549 Twigg H, Morris A, Ghedin E, Bushman FD, Collman RG, Knight R, Fontenot AP, Lung HIVMP.
550 Widespread colonization of the lung by Tropheryma whipplei in HIV infection. *Am J Respir Crit*
551 *Care Med* 2013; 187(10): 1110-1117.

552 31. Simpson JL, Daly J, Baines KJ, Yang IA, Upham JW, Reynolds PN, Hodge S, James AL,
553 Hugenholtz P, Willner D, Gibson PG. Airway dysbiosis: Haemophilus influenzae and
554 Tropheryma in poorly controlled asthma. *Eur Respir J* 2016; 47(3): 792-800.

555 32. Goleva E, Jackson LP, Harris JK, Robertson CE, Sutherland ER, Hall CF, Good JT, Jr.,
556 Gelfand EW, Martin RJ, Leung DY. The effects of airway microbiome on corticosteroid
557 responsiveness in asthma. *Am J Respir Crit Care Med* 2013; 188(10): 1193-1201.

558 33. Hogg JS, Hu FZ, Janto B, Boissy R, Hayes J, Keefe R, Post JC, Ehrlich GD.
559 Characterization and modeling of the Haemophilus influenzae core and supragenomes based
560 on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol*
561 2007; 8(6): R103.

562 34. Arce FT, Carlson R, Monds J, Veeh R, Hu FZ, Stewart PS, Lal R, Ehrlich GD, Avci R.
563 Nanoscale structural and mechanical properties of nontypeable Haemophilus influenzae

564 biofilms. *J Bacteriol* 2009; 191(8): 2512-2520.

565 35. Harrison A, Dyer DW, Gillaspay A, Ray WC, Mungur R, Carson MB, Zhong H, Gipson J,
566 Gipson M, Johnson LS, Lewis L, Bakaletz LO, Munson RS, Jr. Genomic sequence of an otitis
567 media isolate of nontypeable *Haemophilus influenzae*: comparative study with *H. influenzae*
568 serotype d, strain KW20. *J Bacteriol* 2005; 187(13): 4627-4636.

569 36. Coenye T, Vandamme P, LiPuma JJ. Infection by *Ralstonia* species in cystic fibrosis
570 patients: identification of *R. pickettii* and *R. mannitolilytica* by polymerase chain reaction.
571 *Emerg Infect Dis* 2002; 8(7): 692-696.

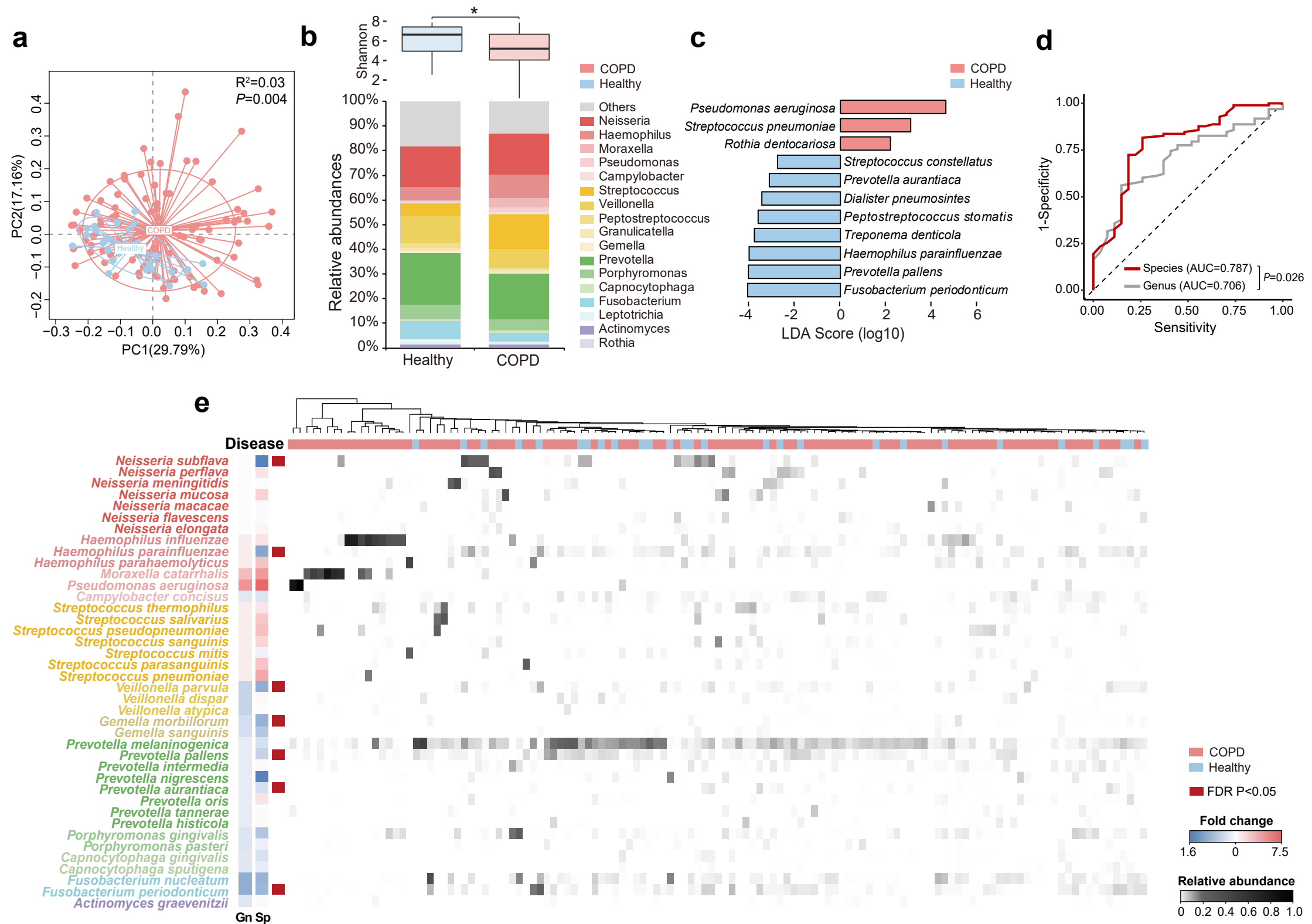
572 37. Zong ZY, Peng CH. *Ralstonia mannitolilytica* and COPD: a case report. *Eur Respir J* 2011;
573 38(6): 1482-1483.

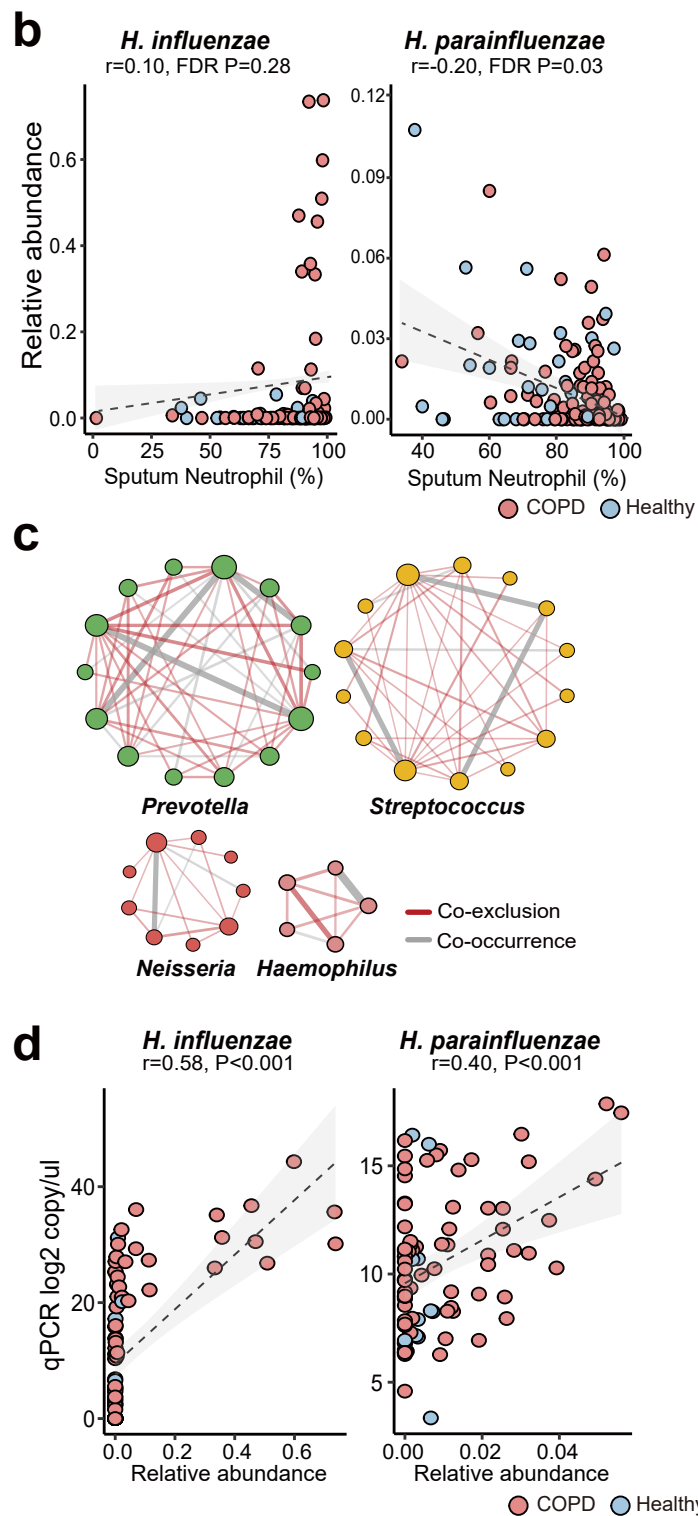
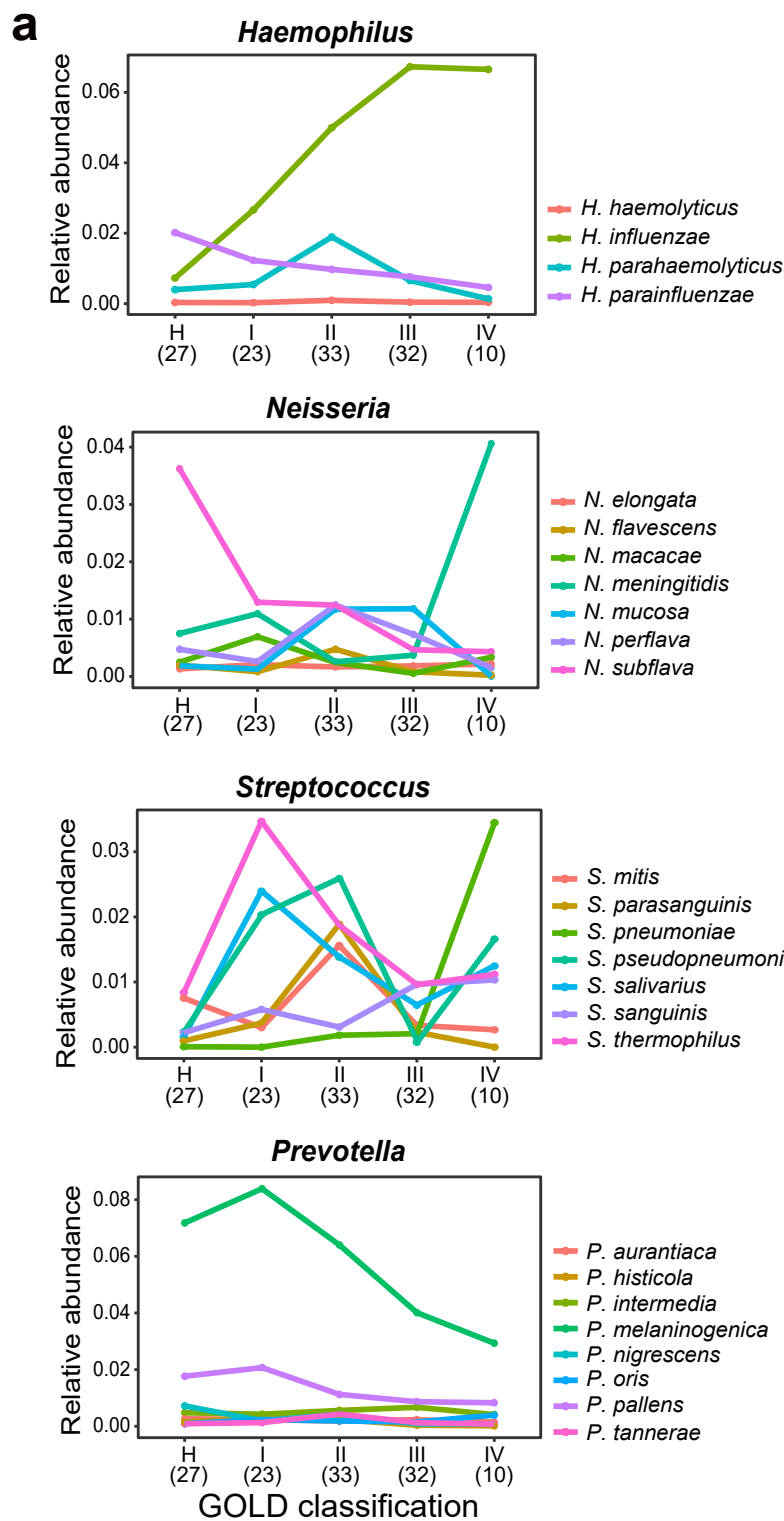
574 38. Basso M, Venditti C, Raponi G, Navazio AS, Alessandri F, Giombini E, Nisii C, Di Caro A,
575 Venditti M. A case of persistent bacteraemia by *Ralstonia mannitolilytica* and *Ralstonia*
576 *pickettii* in an intensive care unit. *Infect Drug Resist* 2019; 12: 2391-2395.

577 39. Cait A, Hughes MR, Antignano F, Cait J, Dimitriu PA, Maas KR, Reynolds LA, Hacker L,
578 Mohr J, Finlay BB, Zaph C, McNagny KM, Mohn WW. Microbiome-driven allergic lung
579 inflammation is ameliorated by short-chain fatty acids. *Mucosal Immunol* 2018; 11(3): 785-795.

580 40. Mao K, Chen S, Chen M, Ma Y, Wang Y, Huang B, He Z, Zeng Y, Hu Y, Sun S, Li J, Wu X,
581 Wang X, Strober W, Chen C, Meng G, Sun B. Nitric oxide suppresses NLRP3 inflammasome
582 activation and protects against LPS-induced septic shock. *Cell Res* 2013; 23(2): 201-212.

583





Group 1: Th2

Group 2: Th1/Th17/Pro-inflammatory

Group 3: Others

