ERJ Express. Published on November 21, 2018 as doi: 10.1183/13993003.01225-2018

EUROPEAN RESPIRATORY *journal*

FLAGSHIP SCIENTIFIC JOURNAL OF ERS

Early View

Original article

Airway Microbiome in Adult Survivors of Extremely Preterm Birth (The EPICure Study)

Sylvia A. D. Rofael, Timothy D. McHugh, Rachael Troughton, Joanne Beckmann, David Spratt, Neil Marlow, John R. Hurst

Please cite this article as: Rofael SAD, McHugh TD, Troughton R, *et al*. Airway Microbiome in Adult Survivors of Extremely Preterm Birth (The EPICure Study). *Eur Respir J* 2018; in press (https://doi.org/10.1183/13993003.01225-2018).

Airway Microbiome in Adult Survivors of Extremely Preterm Birth (The EPICure Study)

Sylvia A D Rofael[1,5], Timothy D McHugh[1], Rachael Troughton[1], Joanne Beckmann[2], David Spratt[3], Neil Marlow[2], John R Hurst[4*]

1. Centre for Clinical Microbiology, Division of Infection and Immunity, UCL, London, UK.
2. Academic Neonatology, UCL Elizabeth Garret Anderson Institute for Women's Health, UCL, London, UK.
3. Department of Microbial Diseases, UCL Eastman Dental Institute, UCL, London, UK.
4. UCL Respiratory, University College London, London, UK.
5. Faculty of Pharmacy, University of Alexandria, Egypt.
   *Corresponding author

Correspondence should be addressed to John Hurst, PhD. FRCP
Address: UCL Respiratory, University College London, 1st floor Royal Free Campus, Pond Street, London. NW4 2PF
Tel: +44 207 472 6260
Fax: +44 207 472 6141
E-mail: j.hurst@ucl.ac.uk

Authors contributions: TM, DS, JH conceived the idea, SR and RT processed the samples and did the lab work, SR performed the bioinformatics and data analyses, SR, TM and DS interpreted the data, JB collected the samples and clinical data, NM is the principal investigator of EPICure study. SR wrote the manuscript with contributions from all other authors.

The EPICure Study was supported by a Programme Grant made by the Medical Research Council, UK.

Take Home Message:

For the first time we demonstrated that extremely preterm birth results in significant dysbiosis in the airway microbiota in early adulthood which correlated with FEV1 as a clinical parameter of obstructive lung disease.

Take Home Message

Extremely preterm birth results in significant early-adult dysbiosis in the airway microbiota.

This article has an online data supplement.

Abstract:

Bronchopulmonary Dysplasia(BPD) is a major complication of preterm birth that leads to lifelong respiratory morbidity. The EPICure study has investigated the longitudinal health outcomes of infants born extremely preterm (<26 weeks-gestation). Our aim was to characterise the airway microbiome in young adults born extremely preterm (EP), with and without neonatal BPD, in comparison to matched term-born controls.

Induced sputum was collected from 92 young adults age 19 years (51 EP and 41 controls). Typical respiratory pathogens were detected using quantitative-PCR. 16S-rRNA gene sequencing was completed on 74 samples (29 EP with BPD, 9 EP without BPD and 36 controls).

The preterm group with BPD had the least diverse bacterial communities. The relative-abundance of *Bacteriodetes*, particularly *Prevotella melaninogenica* was significantly lower in the preterm group compared to controls. This decline was balanced by a nonsignificant increase in *Firmicutes.* Total *Prevotella* relative-abundance correlated with FEV_1 z-score (ρ=0.272; P<0.05). Typical respiratory pathogens loads and prevalence were similar between groups.

In conclusion**,** extremely preterm birth is associated with a significant dysbiosis in airway microbiome in young adulthood regardless of neonatal BPD status. This is characterised by a shift in the community composition away from *Bacteriodetes* as manifested in a significant drop in *Prevotella* relative-abundance.

Key words: Bacteria/Classification, 16S rRNA sequencing, Bronchopulmonary Dysplasia, Microbiota, *Prevotella*.

Introduction

Global and national survival rates of extremely preterm (EP) birth have steadily increased over the past decades with advances in perinatal and neonatal care; however, the prevalence of bronchopulmonary dysplasia (BPD) remains high [1, 2]. BPD is a multifactorial lung disease that develops primarily in preterm infants and may lead to lifelong respiratory morbidity. Initially, the aetiology was thought to be mechanical lung injury, but more recently in more immature infants a 'new' form of BPD has been defined, characterised by disrupted distal lung development, arrest of alveolarization and interference with normal vascularization [3]. The long-term sequelae of BPD include obstructive lung disease and reduced aerobic capacity in adulthood which can be confused with, and may meet spirometric criteria for asthma and chronic obstructive pulmonary disease (COPD) [4].

The EPICure study is a national cohort study investigating the health outcomes of babies born at less than 26 weeks of gestation in the United Kingdom and Ireland between March and December 1995. This unique cohort has been followed up and assessed at 2.5, 6, 9, 11 and now 19 years of age [2, 5, 6]. At 11 years of age, children who were born extremely preterm had significantly more chest deformities, respiratory symptoms, lung function abnormalities with evidence of airway obstruction, ventilation inhomogeneity, gas trapping and airway hyper-responsiveness, and twice the prevalence of asthma, compared to their classmates who were born full term [5, 7].

Recently, some studies have provided evidence that bacteria may play a role in the development of BPD in preterm infants [8-11]. One longitudinal study described a characteristic pattern of airway microbial dysbiosis prior to the development of BPD. This was characterised by a remarkable decrease in the richness and alpha diversity with time; in addition to a shift in the bacterial community composition, in contrast to a relatively diverse and stable community in the preterm infants who did not develop BPD [12]. However, the extent to which microbial succession sustains the dysbiosis of the airway microbiome described in infancy into later life stages has not been investigated. Also, how the airway microbiome of BPD survivors differs from the healthy microbiome in adulthood is not known. The aim of this study is to characterize the airway microbiome in the EPICure cohort at the age of 19 years and compare it with the healthy microbiome of their counterparts who were born full term.

Materials and Methods

The EPICure study was approved by the Southampton and South West Hampshire National Research Ethics Committee [5]. All participants gave informed consent for tests to be performed on their biological samples in relation to the EPICure@19 Study.

Details on sputum sample collection and processing are described in the online data supplement. Briefly, induced sputum was collected from 92 participants in the EPICure@19 study during their 19 years-old follow up visit. The average percentage of squamous epithelial cells in one aliquot of each sputum sample was measured by microscopy for quality control.

Samples were allocated to 3 groups based on the medical history of participants; those who were born extremely preterm without BPD history (EP no BPD); those who were born extremely preterm with neonatal BPD (EP+BPD); and full term-born participants of the same age, who had been evaluated as classmates in earlier studies (EPICure Control group). Microbiome investigators were blinded to the group assignment until analysis was complete.

Metagenomic DNA was extracted from 500 µL of Sputasol® treated sputum samples using Qiagen DNeasy® Blood and Tissue kit (Qiagen, UK) as per the manufacturer's protocol. Extraction negative controls of the saline used for sputum induction and diluted Sputasol® were also performed. A mock community composed of equal proportions of *Streptococcus pneumoniae* ATCC 49619, *Haemophilus influenzae* ATCC 8468 and *Moraxella catarrhalis* ATCC 25240 was run as a positive control.

The bacterial loads of *H. influenzae*, *M. catarrhalis*, *S. pneumoniae* were quantified using multiplex qPCR as previously described [13]. The bacterial load of each tested organism was calculated in colony forming unit (CFU) per mL of sputum and a mean of technical triplicates was taken for each sample. Additional details are provided in the online data supplement.

A sequence library was created by amplification of V5-V7 regions of the bacterial 16S *rrna* gene from metagenomic DNA using 785 forward primer and 1175 reverse primer. Each sample was assigned a unique pair combination of standard Illumina® dual indexed primers, 74 samples produced an amplicon at the expected size (504 bp). The PCR products were cleaned up using Agencourt AMPure XP beads (Beckman Coulter, UK) to remove amplicons <200 bp, DNA was quantified using Qubit™ dsDNA HS kit (Thermo Fisher, UK). The Samples were pooled in equimolar ratio at 10 nM. Sequencing was performed using Illumina MiSeq Platform using costume sequencing primers, MiSeq® Reagent Kit v2 (500 cycles) and PhiX control V3 KIT as internal control for sequencing (Illumina Ltd, UK). The extraction negative controls and a no-template PCR control (water) were run throughout the amplification and sequencing process to reveal any potential contamination. Additional details on the PCR and primers sequences are provided in the online data supplement. Sequence data are deposited in the European Nucleotide Archive (ENA), study accession number is PRJEB27216.

In bioinformatics analysis, we adopted the workflow established by Microbiome helper for stitching paired reads, quality filtering reads with Q<30 over 10% of bases and length <350 bp, and chimera screening [14]. The subsequent steps were done through QIIME pipeline (v.1.9.1) [15], the sequences were clustered based on 97% similarity into Operational Taxonomic Units (OTU) and taxonomic classification was assigned to OTUs using open-reference OTU picking against Greengenes database (v.13_8). The OTU table was then rarefied per sample to 4000 reads removing four samples with <1000 reads (1 EP with neonatal BPD and 3 controls). Alpha and beta diversity indices were calculated on the rarefied OTU table. The appropriate statistical significance tests were calculated using SPSS (v.23) or QIIME wrapper scripts. STAMP (v2.1.3) [16] was used to visualize the data. Whenever applicable the *P*-values were corrected using Benjamini-Hochberd False Discovery Rate (FDR) method for multiple comparisons.

Results

General Characteristics of the participants

Induced sputum samples were collected from 92 young adult participants; 51 were born extremely preterm (EP) <26 weeks' gestation and 41 were born full-term (control group). Mean age (SD) was 19 years (0.5). Microbiome analysis was completed on 74 samples with an amplified 16S rRNA gene: 36 controls and 38 EP; 29 of whom had neonatal BPD, defined as receiving supplemental oxygen or respiratory support at 36 weeks postmenstrual age. The demographic clinical and medical data of the whole cohort and the 74 participants whose samples were sequenced recorded at the 19 years follow up visit are reported in Table 1.

Within the sequenced cohort, forced expiratory volume in 1 second (FEV_1) was significantly lower in the EP group with BPD compared to controls (mean difference: -0.91 L, 95% CI: -1.24 L to -0.59 L) and the EP group without BPD (mean difference: -0.81 L, 95% CI: -1.31 L to -0.32 L). After adjustment for age, sex, and body size using z-scores, the mean FEV_1 z-score of the EP group with BPD was also significantly lower compared to the control group and the EP group without BPD, mean difference (95% CI) were -1.55 (-2.05 to -1.05) and -1.13 (-1.88 to -0.37) respectively.

The prevalence of self-reported asthma was relatively higher in the EP group with BPD (59%) and controls (38%), compared to the EP group without BPD (22%), although this did not reach statistical significance. As discussed further below it is likely that there is significant over-diagnosis of 'asthma' here, but we did not consider it ethical to stop asthma treatment for the purposes of research. At the time of sample collection, the mean (SD) fractional exhaled nitric oxide (FeNO) concentration was 16.59 (14.10) and 25.57 (27.71) ppb in the pre-term and control groups respectively (P>0.05). The mean (SD) eosinophil counts in blood were 190 (136) and 232 (156) cells/μL in the preterm group and controls respectively (P>0.05). Forty-seven percent of the EP group and 50% of the control group were prescribed inhalers. No statistically significant differences were found across the three groups with respect to the number of patients who were prescribed bronchodilator inhalers, inhaled corticosteroids (ICS) or those who had been treated with antibiotics for respiratory problems in the year prior to sample collection. The numbers of males and females, smokers and those who were exposed to passive smoking (>30 minutes/week) were also similar across the three groups.

Table 1: Demographic and clinical characteristics of the whole EPICure participants and those with sequenced samples

Characteristics	Whole Cohort (n=92)				Sequenced			
	EP (n=51)		Controls	*P*	EP		Cont	*P*
	BPD	No BPD			BPD	No		
N	37	14	41		29	9	36	
Age† years	19.02 (0.54)	19.02 (0.41)	19.09 (0.52)	0.803[1]	18.93 (0.66)	18.82 (0.48)	19.02 (0.55)	0.474[4]
Males‡	35%	43%	39%	0.866[2]	34%	56%	42%	0.521[2]
Females	65%	57%	61%		66%	44%	58%	
Asthma diagnosis‡	57%	21%*	37%	0.047[2]	59%	22%	38%	0.094[2]
Current Smoker‡	22%	29%	27%	0.821[2]	28%	33%	33%	0.873[2]
Passive Smoke exposure (>30 mins/week) ‡	22%	29%	35%	0.471[2]	32%	22%	60%	0.348[2]
Squamous epithelial cell %†	17.4 (17.7)	17.9 (11.9)	13.8 (11.1)	0.573[4]	20.9 (19.2)	19.5 (14.4%)	14.6 (11.3)	0.566[4]
Prescribed Inhalers§‡	60%	38%	64%	0.557[3]	62%	17%	50%	0.239[3]
Prescribed Bronchodilator inhalers§§‡	53%	25%	55%	0.417[3]	54%	17%	50%	0.298[3]
Prescribed ICS‡	27%	13%	36%	0.476[3]	31%	17%	33%	0.883[3]
Antibiotic treatment in past year ‡	7%	13%	30%	0.397[3]	8%	0	27%	0.395[3]
Treated for Respiratory problem in the past year‡	20%	21%	25%	0.393[3]	22%	11%	28%	0.556[2]
FEV_1 (L)†	2.66** (0.54)	3.22 (0.76)	3.57 (0.65)	0.000[1]	2.63** (0.66)	3.45 (0.76)	3.55 (0.61)	0.000[1]
FEV1 z score †	-1.66** (1.09)	-0.911 (0.04)	-0.37 (0.87)	0.000[1]	-1.87** (1.17)	-0.75 (0.69)	-0.32 (0.89)	0.000[1]

Percent change in FEV_1 with bronchodilator†	7.93% (6.25)	7.75% (7.81)	5.26% (5.60)	0.077[1]	9.52% (7.77)	7.19% (4.91)	5.50% (5.85)	0.058[1]
FeNO (ppb) †	16.33 (12.63)	18.00 (14.72)	26.47 (26.30)	0.683[4]	16.50 (13.98)	16.89 (15.34)	25.57 (27.71)	0.830[4]
Eosinophils (cells/µL) †	181 (127)	164 (104)	229 (148)	0.184[4]	194 (141)	179 (128)	232 (156)	0.399[4]

†Continuous data are expressed as mean (SD)
‡Categorical data are expressed as percentages
* *P*<0.05 ** *P*<0.01
1. *P* value calculated by ANOVA 2. *P* value calculated by Chi Square
3. *P* value calculated by Fisher's exact test 4. *P* value calculated by Kruskal Wallis Test
EP: Extreme Preterm birth BPD: Broncho-pulmonary dysplasia
FEV1: Forced Expiratory volume in 1 sec FeNO: Fractional Exhaled Nitric Oxide
§ Inhalers: β2 adrenoreceptor agonists: salbutamol (Ventolin®), terbutaline (Bricanyl®), salmeterol (Servent®, Seretide®), Muscarinic receptors antagonists: Ibratropium (Atrovent®), leukotriene receptor antagonist: Montelukast (Singulair®), Inhaled Corticosteroids (ICS): beclomethasone (Becotide®), budesonide (Pulmicort®), fluticasone (Flixotide®, Seretide®).
§§ Bronchodilator Inhalers: β2 agonist and Muscarinic receptors antagonists

Microbial Community Composition

The bacterial communities were significantly less diverse and less rich in the sputum samples from the whole EP group compared to controls; the mean difference (±SEM) in Chao 1 and Fisher alpha indices between the whole EP group and the controls were -39 (±13, $P<0.05$) and -4.8 (±2.3, $P<0.05$) respectively (online data supplement Figures E1). Other richness and α-diversity indices including the number of observed OTUs and PD whole tree also showed significantly less diverse and less rich microbiota in the EP group (online data supplement Figure E2). Within the EP group, the trend observed in all previously mentioned α-diversity indices suggests that those with neonatal BPD had the least diverse and rich microbial communities, while controls had the highest values (Figures 1A, 1B and online data supplement Figure E2). This trend was statistically significant only in Chao 1 when tested by ANOVA ($P<0.05$). Nevertheless, the post hoc comparisons of differences between the EP group with neonatal BPD and controls were significant (Figure 1).

In principal coordinate analysis of weighted Unifrac β diversity index, the samples from EP participants significantly clustered, regardless of neonatal BPD status; whereas the samples from controls were scattered, $P<0.01$ by ANOSIM and $P<0.05$ by PERMANOVA (Figure 1C and online data supplement Figures E1).

The bacterial community at phylum-level was dominated by *Firmicutes,* followed by *Bacteriodetes,* then *Proteobacteria* and *Actinobacteria*. The samples from both EP groups, with and without BPD, had a significantly lower relative abundance (RA) of phylum *Bacteriodetes* compared to the control group; $P<0.05$ by Kruskal Wallis test. Differences were compensated by a non-significant increase in the relative abundance of *Firmicutes* (Figure 2A).

Looking more closely at the composition of the microbial communities at genus level, although natural inter-individual differences were obvious in microbiome profiles of sputum from different participants, the relative abundance of *Prevotella*, was significantly lower in both EP groups, with and without BPD, in comparison to the control group ($P<0.05$ by Kruskal Wallis test) (Figures 2B and 3A). This was compensated for by a non-significant and inconsistent increase in relative abundance of other genera such as *Streptococcus*, *Veillonella*, *Rothia* and *Neisseria* which are normal microbiota in airways (Figure 2B).

Prevotella was completely absent in two negative controls and present at 0.4% relative abundance in the extraction negative control of the sputum induction matrix (NCm) which had 5198 reads classified into three main taxa: *Propionibacterium* (RA 62%), *Staphylococcus* (RA 14%) and *Streptococcus* (6%) (Figure E6 online data supplement). The sputasol extraction negative control (NCr) had 17 reads and the no-template PCR negative control had 2 reads; in contrast, the mean (SD) from the study samples was 24,089 reads (5751) (Figure 2B). This gives confidence that the impact of environmental contamination was minimal. The relative abundance of *Prevotella* did not correlate with the percentage of squamous epithelial cells in the sputum samples (Spearman $\rho=0.07$, $P=0.61$). Nevertheless, *Prevotella* did significantly correlate with the FEV_1 z score (Spearman $\rho = 0.272$, $P=0.02$) (Figure 3B), but not with a self-reported diagnosis of asthma or use of asthma inhalers (online data supplement).

Mining the sequencing data, of 121 OTUs belonging to genus *Prevotella*, the relative abundance of OTU 4458304 accounted for most of the observed difference in the total *Prevotella* relative abundance between the study groups (Figure 3C). By extracting and BLAST searching the 350 bp representative sequence of this OTU against the NCBI 16S ribosomal RNA database using the Nucleotide BLAST tool [17], the sequence was identified as 100% identical to *Prevotella melaninogenica* strains (Figure 3D).

Load and Prevalence of Airway Bacteria using Multiplex q-PCR

The loads and prevalence of the three bacteria: *S. pneumoniae* (Spn), *H. influenzae* (Hi) and *M. catarrhalis* (Mc) were similar within the three study groups (Figure 4). In the EP group with neonatal BPD, 28% had *S. pneumoniae*, 14% had *H. influenzae* and 11% had both organisms together; a similar pattern was observed in controls. The mean bacterial load (SD) of *S. pneumoniae* and *H. influenzae* were similar in these two groups; 4.44 (0.1).$\log_{10}$ CFU/mL. In the EP group without BPD history; *H. influenzae* was the most prevalent organism being detected in 57% of the sputum samples from this group with a mean load (SD) of 4.51 (0.88) $\log_{10}$ CFU/mL of sputum. It was found in 36% of the samples alone and in 21% of the samples with *S. pneumoniae*. *S. pneumoniae* was the most populous with a mean bacterial load (SD) of 4.96 (1.08) $\log_{10}$ CFU/mL (Figure 4).

M. catarrhalis was the least prevalent and the least populous organism. The mean load of *M. catarrhalis* was 0.7 $\log_{10}$ higher in BPD group compared to the other two groups (P>0.05, ANOVA). None of the differences in prevalence and load of the three organisms across the study groups were statistically significant.

There was a significant correlation between the sequencing relative abundance of genus *Moraxella* and the corresponding qPCR bacterial loads (rho= 0.472, P<0.01, Figure 5C). However, the correlation was not statistically significant for *Haemophilus* and *Streptococcus* (Figures 5A and 5D). This can be attributed to the presence of commensal members of these two genera in sputum as normal respiratory microbiota. Within the genus *Haemophilus,* 5 OTUs were classified as *H. influenzae*, and the correlation between the relative abundance of *H. influenzae* OTUs in sequencing and *H. influenzae* load in qPCR was statistically significant (rho= 0.43, P<0.01, Figure 5B). Comparing the relative abundance by sequencing to the qPCR loads and bacterial count by quantitative cultures in the mock community, *Streptococcu*s was over represented by 15% and *Haemophilus* was underrepresented by 21% (Figure 5D). Further data on the sensitivity and specificity of the methods is provided in the online data supplement.

Discussion

To the best of our knowledge, this is the first study to investigate the airway microbiome in adult survivors of preterm birth in comparison to matched full term-born controls. Thus far, previous studies investigating the association between bacteria and BPD involved preterm born infants [10-12, 18, 19]. In the current study, we have compared the airway microbiome in 38 young adults who were born at less than 26 weeks of gestation; 29 of them had neonatal BPD and 9 did not, and 36 full term-born controls, recruited from classmates at 6 and 11 years of age.

The airway microbiome profile in the three study groups was consistent with previous studies [20, 21]. It was dominated by *Firmicutes, Bacteriodetes, Proteobacteria* and *Actinobacteria*. The extremely preterm group had significantly less diverse and less rich microbial communities in comparison with the control group. We did not find significant differences between the airway microbiome profiles of the extremely preterm groups with and without neonatal BPD. The samples from both groups clustered together in principal coordinate analysis of weighted Unifrac β diversity index regardless of neonatal BPD status (Figure 1). No significant differences were observed in the richness and alpha diversities between the two groups; however, a trend was observed in which the alpha diversity and richness of the microbial communities in the BPD group was slightly but not significantly lower than the group without BPD, and significantly lower than the control group. This trend may be important, but our study did not have sufficient statistical power to be able to confirm a difference in all microbial diversity indices. We demonstrated a significant shift away from *Bacteriodetes*, driven particularly by reduction in the relative abundance of genus *Prevotella* in both extremely preterm groups, with and without BPD, relative to the control group. Moreover, *Prevotella* relative abundance significantly correlated with FEV_1 z-score but had no association with other clinical parameters such as smoking status, exposure to passive smoking, self-reported diagnosis of asthma, fractional exhaled nitric oxide concentration, blood eosinophils count, or use of inhalers.

Recent observational studies that have investigated the lung microbiome have commonly detected *Prevotella* in lung tissues and broncho-alveolar lavage of healthy subjects [20, 22, 23]. It is suggested that a high abundance of *Bacteriodetes* reflects a healthy lung microbiome [24]. An association between reduced relative abundance of *Bacteriodetes* and chronic lung disease has also been frequently reported [23-26]. A shift in community composition away from *Bacteriodetes* towards *Proteobacteria* has been observed in people with COPD [27, 28] and asthmatics [28]. A similar trend in which there is a shift from *Bacteriodetes* to *Firmicutes*, mainly streptococci, has been reported by Zhang *et al.* in severe asthma [29]. Hilty reported a shift in the community membership away from *Bacteriodetes*, mainly *Prevotella* towards *Proteobacteria* (mainly *Haemophilus),* as well as *Firmicutes* in asthmatic children [28]. In Lal's *et al.* study which compared the airway microbiome at birth in both preterm and full term born infants, *Prevotella* was detected in tracheal aspirates at day one of life. In their results, *Prevotella* relative abundance was lower in extremely low birth-weight preterm infants compared to controls. Among the preterm infants, those who were predisposed to BPD had the least *Prevotella* abundance, however this trend was not statistically significant [30].

We analysed the sequencing data to go beyond the genus level and found that *Prevotella melaninogenica* was the species that contributed most to the observed reduction in total *Prevotella* abundance in the preterm groups in comparison to the control group. *Prevotella* species are obligate anaerobes that have been regarded as opportunistic members of the oral microbiota as well as other body sites and have been isolated with other bacteria in mixed anaerobic infections and lower respiratory tract infections [31]. Increased abundance of *Prevotella* has been associated with some diseases such as periodontitis, rheumatoid arthritis, bacterial vaginosis and inflammatory bowel disease [32]. Nevertheless, other *Prevotella* species e.g. *P. intermedia* and *P. nigrescens* are usually associated with disease [33]. Very little has been published on *Prevotella melaninogenica,* but this species was the most frequently isolated from the sputum in cystic fibrosis in previous studies and it was noticed that the presence of anaerobes was associated with clinical stability [34, 35].

Currently, it is not clear what role *Prevotella* plays in lower respiratory microbial homeostasis. Many studies that compared the lung microbiome in health and disease suggest that *Prevotella* is associated with health, and is quickly replaced by a members of *Proteobacteria* or *Firmicutes* in various chronic lung conditions [24, 26-29].On the other hand, some studies have proposed that *Prevotella* might be contributing to the pathogenesis of lung diseases [32, 36-39]. However, caution must be exercised in extrapolating observations to the whole genus. The role of *Prevotella* species in pathogenesis has received little attention possibly because *Prevotella* is difficult to culture and is not usually isolated from specimens in routine microbiology laboratories. Consequently, further research is required to understand the interactions of *Prevotella* species with the host immune system and with other microbes within the lung microbial communities, and to characterise the strains that may be beneficial to respiratory health.

S. pneumoniae, H. influenzae, and *M. catarrhalis* conventionally have been classified as typical airway pathogens. These organisms normally reside harmlessly within the human nasopharynx [40]. Numerous studies have demonstrated the potential consequences of these organisms in COPD [13] and asthma [28] exacerbations. Therefore, it was important to investigate the prevalence and loads of these three bacteria as pathogens in our cohort to investigate the possibility that reduced abundance of normal flora may open niches for opportunistic pathogens. Unfortunately, one limitation of 16S sequencing methods is that 16S rRNA hypervariable regions exhibit different degrees of sequence diversity, and no single hypervariable region can discriminate between all bacterial groups. In many instances, the bioinformatic pipelines provide reliable identification down to the genus level for most of organisms [41]. For this reason, despite the high sensitivity of the 16S sequencing and good correlation with the qPCR loads, specificity was relatively low especially for *S. pneumoniae* and *H. influenzae* – likely due to the presence of commensal members of these genera. Therefore, qPCR may be a more appropriate method to determine the prevalence and load of specific pathogenic bacteria. We did not find significant differences in the bacterial load nor the prevalence of these pathogens between the three study groups.

We also compared the airway microbiome profiles of our young adult cohort with published microbiological data for preterm infants during infancy. Numerous studies have identified an important role for respiratory colonisation with *Ureaplasma* in the development of BPD [10, 11, 18]. In our results, *Ureaplasma* was rarely identified in the extremely preterm group (RA< 0.01% in 6 sputum samples) (data not shown). In Lohmann's study, notable changes were reported in airway microbiome of preterm infants who subsequently developed BPD. Immediately after birth, the preterm airway microbiome was mainly dominated by *Proteobacteria,* particularly *Acinetobacter* species and over time the relative abundance of *Firmicutes* increased, driven mainly by *Staphylococcus* in those infants who developed BPD, in contrast to the relatively diverse and stable community in the non-BPD group [12]. *Staphylococcus* was also associated with BPD development in another study [19]. In our results, *Acinetobacter* and *Staphylococcus* were rarely identified in our cohort (RA for both was <0.1%). Interestingly when detected, *Staphylococcus* relative abundance was slightly higher in the BPD group compared to the other two groups;

however, this observation was not statistically significant (Figure E5 online data supplement). *Corynebacterium* has also been associated with the development of severe BPD in premature infants [42]. In our results, Corynebacterium was present at similar RA in the three groups at around 1%.

The main limitations of this study relate to the small sample size and an inability to induce sputum in all subjects, which may not be random as it is more challenging to induce sputum in healthy subjects compared to those with respiratory pathology. For clinical safety reasons, lower hypertonic saline strength was used for sputum induction in those who were labelled as asthmatics which might have accounted for some of the variations between the groups. Although the microbial signature in the hypertonic saline matrix negative control was quite different from that in the sputum samples, it would have been useful to investigate the effect of the different saline strengths on the quality of sample, DNA extraction and PCR inhibition. Asthma may also be regarded as a potential confounder. The asthma prevalence was self-reported, we do not have objective testing to confirm or refute this, and many subjects were prescribed asthma inhalers. EPICure and other preterm cohort studies have previously reported that BPD survivors often have airflow obstruction later in life which can be mis-labelled as asthma [5, 43, 44]. Nevertheless, in covariate analysis, asthma status was not a significant covariate (P>0.05 by both adonis and ANOSIM). Moreover, none of the microbiome describing parameters (α and β indices) nor the *Prevotella* relative abundance were different between subjects labelled as asthmatics and non-asthmatics, either in the whole cohort or the control group (Figure E3 online data supplement). In all sputum-based studies, upper airway contamination is a concern. Induced sputum was the best available sample in a group of young adults with lung disease many of whom have additional developmental problems preventing research bronchoscopy. It would have been useful to directly compare the microbial signature in parallel upper and lower respiratory samples.

In conclusion, young adults born extremely preterm exhibit significant dysbiosis at 19 years of age. This is characterised by a shift in the microbial community structure away from *Bacteriodetes* and specifically manifest by a significant reduction in the relative abundance of genus *Prevotella*, as frequently described in other chronic lung diseases. *Prevotella melaninogenica* was the species showing most variation within this genus. The prevalence and loads of typical respiratory opportunistic pathogens were not affected by such dysbiosis, and we did not identify persistence of dysbiosis patterns related to the development of neonatal BPD reported in studies of BPD in infancy.

Acknowledgments
The first author would like to thank the Missions sector in the Egyptian Ministry of Higher Education and the British Council Egypt for funding her PhD as part of the Newton-Mosharafa scheme. NM receives part funding from the Department of Health's NIHR Biomedical Research Centre's funding scheme at UCLH/UCL. This study was carried out in the NIHR/Wellcome UCLH Clinical Research Facility.

References

1. Santhakumaran S, Statnikov Y, Gray D, Battersby C, Ashby D, Modi N. Survival of very preterm infants admitted to neonatal care in England 2008-2014: time trends and regional variation. Arch Dis Child Fetal Neonatal Ed 2018; 103: F208-F215.

2. Costeloe KL, Hennessy EM, Haider S, Stacey F, Marlow N, Draper ES. Short term outcomes after extreme preterm birth in England: comparison of two birth cohorts in 1995 and 2006 (the EPICure studies). BMJ 2012; 345: e7976.

3. Gien J, Kinsella JP. Pathogenesis and Treatment of Bronchopulmonary Dysplasia. Current opinion in pediatrics 2011; 23: 305-313.

4. Lovering AT, Elliott JE, Laurie SS, Beasley KM, Gust CE, Mangum TS, Gladstone IM, Duke JW. Ventilatory and sensory responses in adult survivors of preterm birth and bronchopulmonary dysplasia with reduced exercise capacity. Ann Am Thorac Soc 2014; 11: 1528-1537.

5. Fawke J, Lum S, Kirkby J, Hennessy E, Marlow N, Rowell V, Thomas S, Stocks J. Lung function and respiratory symptoms at 11 years in children born extremely preterm: the EPICure study. Am J Respir Crit Care Med 2010; 182: 237-245.

6. Hennessy EM, Bracewell MA, Wood N, Wolke D, Costeloe K, Gibson A, Marlow N. Respiratory health in pre-school and school age children following extremely preterm birth. Arch Dis Child 2008; 93: 1037-1043.

7. Lum S, Kirkby J, Welsh L, Marlow N, Hennessy E, Stocks J. Nature and severity of lung function abnormalities in extremely pre-term children at 11 years of age. Eur Respir J 2011; 37: 1199-1207.

8. Stressmann FA, Connett GJ, Goss K, Kollamparambil TG, Patel N, Payne MS, Puddy V, Legg J, Bruce KD, Rogers GB. The use of culture-independent tools to characterize bacteria in endo-tracheal aspirates from pre-term infants at risk of bronchopulmonary dysplasia. J Perinat Med 2010; 38: 333-337.

9. Mourani PM, Harris JK, Sontag MK, Robertson CE, Abman SH. Molecular identification of bacteria in tracheal aspirate fluid from mechanically ventilated preterm infants. PLOS ONE 2011; 6: e25959.

10. Kotecha S, Hodge R, Schaber JA, Miralles R, Silverman M, Grant WD. Pulmonary Ureaplasma urealyticum is associated with the development of acute lung inflammation and chronic lung disease in preterm infants. Pediatr Res 2004; 55: 61-68.

11. Beeton ML, Maxwell NC, Davies PL, Nuttall D, McGreal E, Chakraborty M, Spiller OB, Kotecha S. Role of pulmonary infection in the development of chronic lung disease of prematurity. Eur Respir J 2011; 37: 1424-1430.

12. Lohmann P, Luna RA, Hollister EB, Devaraj S, Mistretta TA, Welty SE, Versalovic J. The airway microbiome of intubated premature infants: characteristics and changes that predict the development of bronchopulmonary dysplasia. Pediatr Res 2014; 76: 294-301.

13. Garcha DS, Thurston SJ, Patel AR, Mackay AJ, Goldring JJ, Donaldson GC, McHugh TD, Wedzicha JA. Changes in prevalence and load of airway bacteria using quantitative PCR in stable and exacerbated COPD. Thorax 2012; 67: 1075-1080.

14. Comeau AM, Douglas GM, Langille MG. Microbiome Helper: a Custom and Streamlined Workflow for Microbiome Research. mSystems 2017; 2.

15. 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, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 2010; 7: 335-336.

16. Parks DH, Tyson GW, Hugenholtz P, Beiko RG. STAMP: statistical analysis of taxonomic and functional profiles. Bioinformatics 2014; 30: 3123-3124.

17. National Center for Biotechnology Information (NCBI)

[database on the Internet]. National Library of Medicine (US), National Center for Biotechnology Information. 1988 [cited 16 Feb 2018]. Available from: https://www.ncbi.nlm.nih.gov/.

18. Van Marter LJ, Dammann O, Allred EN, Leviton A, Pagano M, Moore M, Martin C. Chorioamnionitis, mechanical ventilation, and postnatal sepsis as modulators of chronic lung disease in preterm infants. J Pediatr 2002; 140: 171-176.

19. Wagner BD, Sontag MK, Harris JK, Miller JI, Morrow L, Robertson CE, Stephens M, Poindexter BB, Abman SH, Mourani PM. Airway Microbial Community Turnover Differs by BPD Severity in Ventilated Preterm Infants. PLOS ONE 2017; 12.

20. Charlson ES, Bittinger K, Haas AR, Fitzgerald AS, Frank I, Yadav A, Bushman FD, Collman RG. Topographical continuity of bacterial populations in the healthy human respiratory tract. Am J Respir Crit Care Med 2011; 184: 957-963.

21. Morris A, Beck JM, Schloss PD, Campbell TB, Crothers K, Curtis JL, Flores SC, Fontenot AP, Ghedin E, Huang L, Jablonski K, Kleerup E, Lynch SV, Sodergren E,

Twigg H, Young VB, Bassis CM, Venkataraman A, Schmidt TM, Weinstock GM. Comparison of the respiratory microbiome in healthy nonsmokers and smokers. Am J Respir Crit Care Med 2013; 187: 1067-1075.

22. Pragman AA, Lyu T, Baller JA, Gould TJ, Kelly RF, Reilly CS, Isaacson RE, Wendt CH. The lung tissue microbiota of mild and moderate chronic obstructive pulmonary disease. Microbiome 2018; 6: 7.

23. Erb-Downward JR, Thompson DL, Han MK, Freeman CM, McCloskey L, Schmidt LA, Young VB, Toews GB, Curtis JL, Sundaram B, Martinez FJ, Huffnagle GB. Analysis of the lung microbiome in the "healthy" smoker and in COPD. PLOS ONE 2011; 6: 0016384.

24. Dickson RP, Erb-Downward JR, Martinez FJ, Huffnagle GB. The Microbiome and the Respiratory Tract. Annu Rev Physiol 2016; 78: 481-504.

25. Garcia-Nunez M, Millares L, Pomares X, Ferrari R, Perez-Brocal V, Gallego M, Espasa M, Moya A, Monso E. Severity-related changes of bronchial microbiome in chronic obstructive pulmonary disease. J Clin Microbiol 2014; 52: 4217-4223.

26. Wu J, Liu W, He L, Huang F, Chen J, Cui P, Shen Y, Zhao J, Wang W, Zhang Y, Zhu M, Zhang W. Sputum microbiota associated with new, recurrent and treatment failure tuberculosis. PLoS One 2013; 8: e83445.

27. Einarsson GG, Comer DM, McIlreavey L, Parkhill J, Ennis M, Tunney MM, Elborn JS. Community dynamics and the lower airway microbiota in stable chronic obstructive pulmonary disease, smokers and healthy non-smokers. Thorax 2016; 71: 795-803.

28. Hilty M, Burke C, Pedro H, Cardenas P, Bush A, Bossley C, Davies J, Ervine A, Poulter L, Pachter L, Moffatt MF, Cookson WO. Disordered microbial communities in asthmatic airways. PLOS ONE 2010; 5: e8578.

29. Zhang Q, Cox M, Liang Z, Brinkmann F, Cardenas PA, Duff R, Bhavsar P, Cookson W, Moffatt M, Chung KF. Airway Microbiota in Severe Asthma and Relationship to Asthma Severity and Phenotypes. PLOS ONE 2016; 11.

30. Lal CV, Travers C, Aghai ZH, Eipers P, Jilling T, Halloran B, Carlo WA, Keeley J, Rezonzew G, Kumar R, Morrow C, Bhandari V, Ambalavanan N. The Airway Microbiome at Birth. Sci Rep 2016; 6.

31. Kedzia A, Kwapisz E, Wierzbowska M. Incidence of anaerobic bacteria in respiratory tract infections. Pneumonol Alergol Pol 2003; 71: 68-73.

32. Larsen JM. The immune response to Prevotella bacteria in chronic inflammatory disease. Immunology 2017; 151: 363-374.

33. Stingu CS, Schaumann R, Jentsch H, Eschrich K, Brosteanu O, Rodloff AC. Association of periodontitis with increased colonization by Prevotella nigrescens. J Investig Clin Dent 2013; 4: 20-25.

34. Field TR, Sibley CD, Parkins MD, Rabin HR, Surette MG. The genus Prevotella in cystic fibrosis airways. Anaerobe 2010; 16: 337-344.

35. Tunney MM, Field TR, Moriarty TF, Patrick S, Doering G, Muhlebach MS, Wolfgang MC, Boucher R, Gilpin DF, McDowell A, Elborn JS. Detection of anaerobic bacteria in high numbers in sputum from patients with cystic fibrosis. Am J Respir Crit Care Med 2008; 177: 995-1001.

36. Sherrard LJ, McGrath SJ, McIlreavey L, Hatch J, Wolfgang MC, Muhlebach MS, Gilpin DF, Elborn JS, Tunney MM. Production of extended-spectrum β-lactamases and the potential indirect pathogenic role of Prevotella isolates from the cystic fibrosis respiratory microbiota. International journal of antimicrobial agents 2016; 47: 140-145.

37. Flynn JM, Niccum D, Dunitz JM, Hunter RC. Evidence and Role for Bacterial Mucin Degradation in Cystic Fibrosis Airway Disease. PLoS Pathog 2016; 12.

38. Segal LN, Clemente JC, Tsay JC, Koralov SB, Keller BC, Wu BG, Li Y, Shen N, Ghedin E, Morris A, Diaz P, Huang L, Wikoff WR, Ubeda C, Artacho A, Rom WN, Sterman DH, Collman RG, Blaser MJ, Weiden MD. Enrichment of the lung microbiome with oral taxa is associated with lung inflammation of a Th17 phenotype. Nat Microbiol 2016; 1: 16031.

39. Twigg HL, 3rd, Knox KS, Zhou J, Crothers KA, Nelson DE, Toh E, Day RB, Lin H, Gao X, Dong Q, Mi D, Katz BP, Sodergren E, Weinstock GM. Effect of Advanced HIV Infection on the Respiratory Microbiome. Am J Respir Crit Care Med 2016; 194: 226-235.

40. Sulikowska A, Grzesiowski P, Sadowy E, Fiett J, Hryniewicz W. Characteristics of Streptococcus pneumoniae, Haemophilus influenzae, and Moraxella catarrhalis isolated from the nasopharynges of asymptomatic children and molecular analysis of S. pneumoniae and H. influenzae strain replacement in the nasopharynx. J Clin Microbiol 2004; 42: 3942-3949.

41. Chakravorty S, Helb D, Burday M, Connell N, Alland D. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. Journal of microbiological methods 2007; 69: 330-339.

42. Imamura T, Sato M, Go H, Ogasawara K, Kanai Y, Maeda H, Chishiki M, Shimizu H, Mashiyama F, Goto A, Momoi N, Hosoya M. The Microbiome of the Lower Respiratory Tract in Premature Infants with and without Severe Bronchopulmonary Dysplasia. Am J Perinatol 2017; 34: 80-87.

43. Northway WH, Jr., Moss RB, Carlisle KB, Parker BR, Popp RL, Pitlick PT, Eichler I, Lamm RL, Brown BW, Jr. Late pulmonary sequelae of bronchopulmonary dysplasia. N Engl J Med 1990; 323: 1793-1799.

44. Vrijlandt EJ, Gerritsen J, Boezen HM, Duiverman EJ. Gender differences in respiratory symptoms in 19-year-old adults born preterm. Respir Res 2005; 6: 1465-9921.

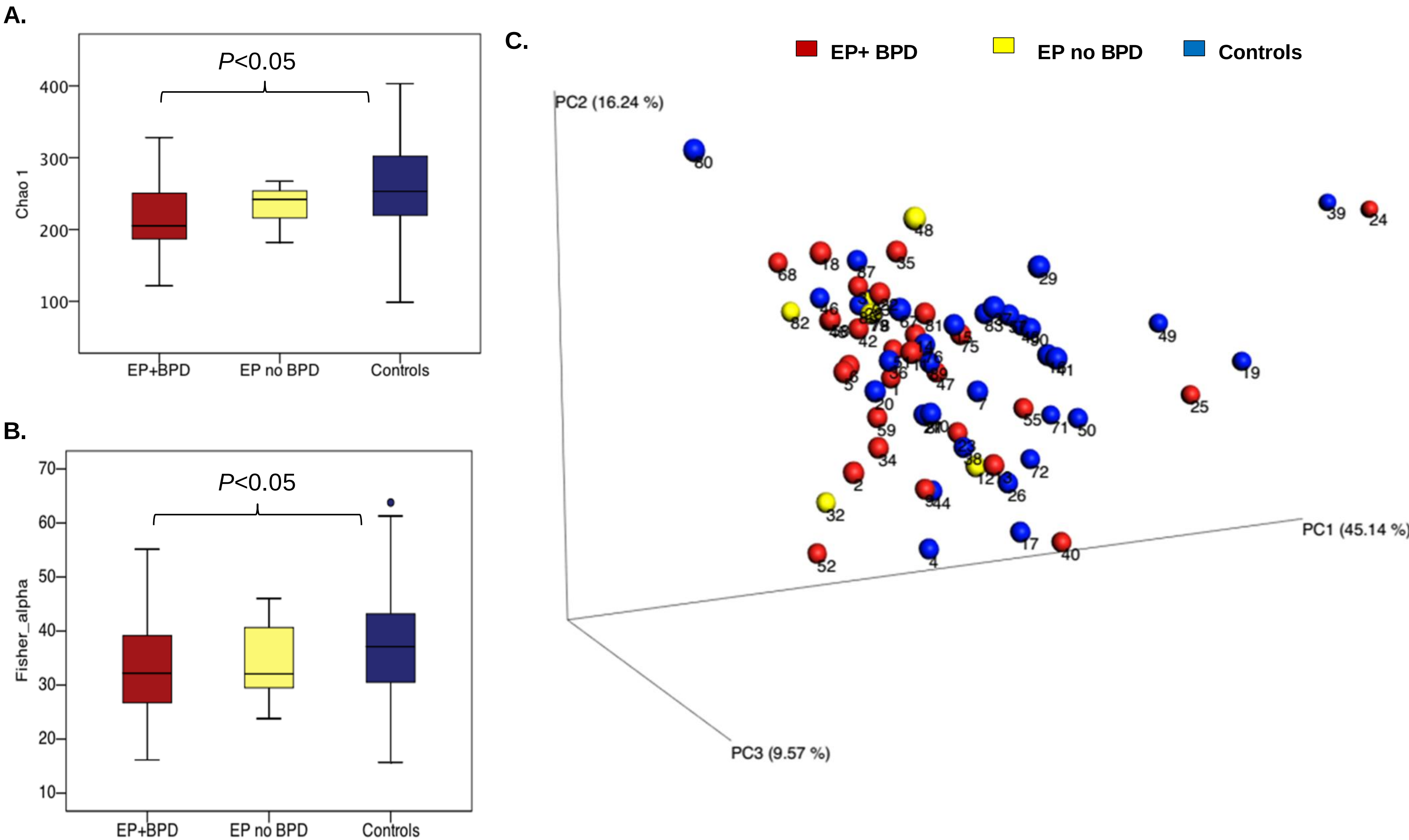


Figure1: **Comparison of the richness and alpha diversity of microbial communities in sputum between the pre-term birth survivors (EP), with and without Neonatal Bronchopulmonary Dysplasia (BPD), and controls** Richness and α diversity measured by **(A)** Chao 1 (*P*<0.05 by ANOVA), **(B)** Fisher-alpha diversity index (*P*=0.07 by ANOVA); nevertheless, Fisher alpha was significantly lower in the whole EP group compared to controls P<0.05 by T-test) **(C)** Principal Coordinate Analysis (PCoA) of weighted UniFrac β diversity index (*P*<0.01 by ANOSIM and *P*<0.05 by PERMANOVA comparing the whole EP group (n=37) and controls (n=33) (*P* >0.05 by ANOSIM and PERMANOVA comparing the 3 groups: EP+BPD (n= 28), EP no BPD (n=9), Controls (n=33).

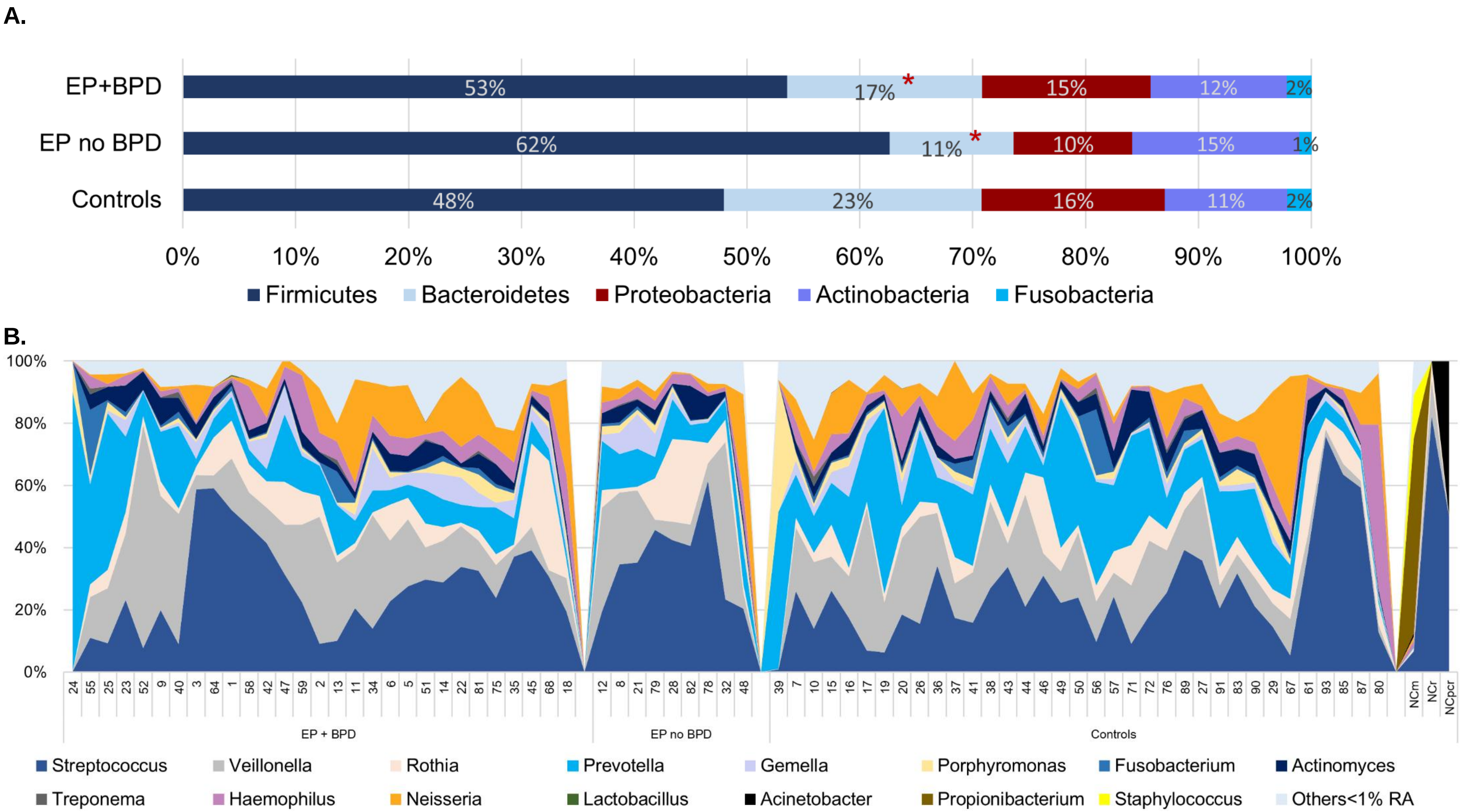


Figure 2: Comparison of the airway microbiome profile in the EPICure groups at (A) phylum level and (B) genus level
(A) At Phylum level, the *Bacteroidetes* relative abundance (RA) was significantly lower in both extremely preterm-born (EP) groups compared to controls (*P*<0.05, by Kruskal-Wallis (KW)), **P*<0.05, by Mann Whitney, **(B)** At the Genus level, the *Prevotella* RA was significantly lower in EP groups compared to controls (*P*<0.05 by KW). NCm: extraction negative control of the saline matrix used for sputum induction, NCr: extraction negative control of the diluted sputasol and reagents, NCpcr: PCR negative control. Sample size:29 EP+BPD, 9 EP no BPD and 36 Controls.

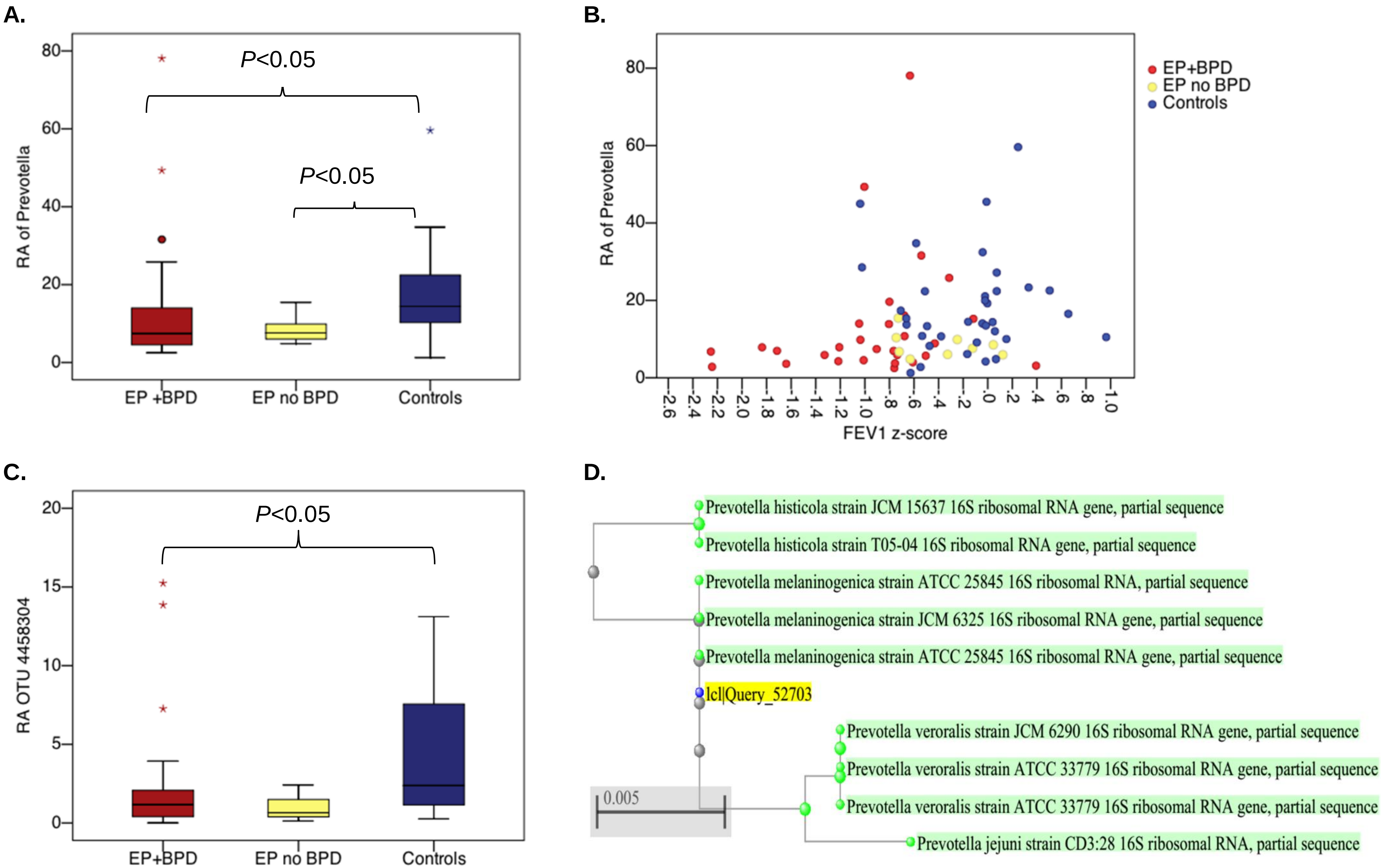
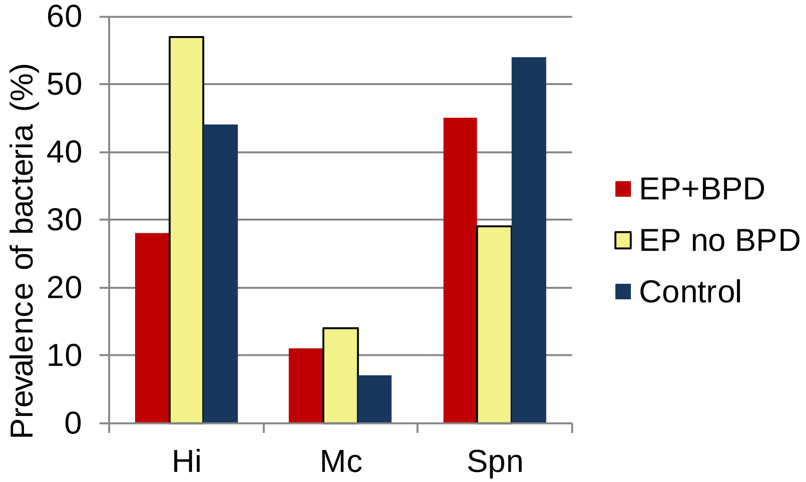
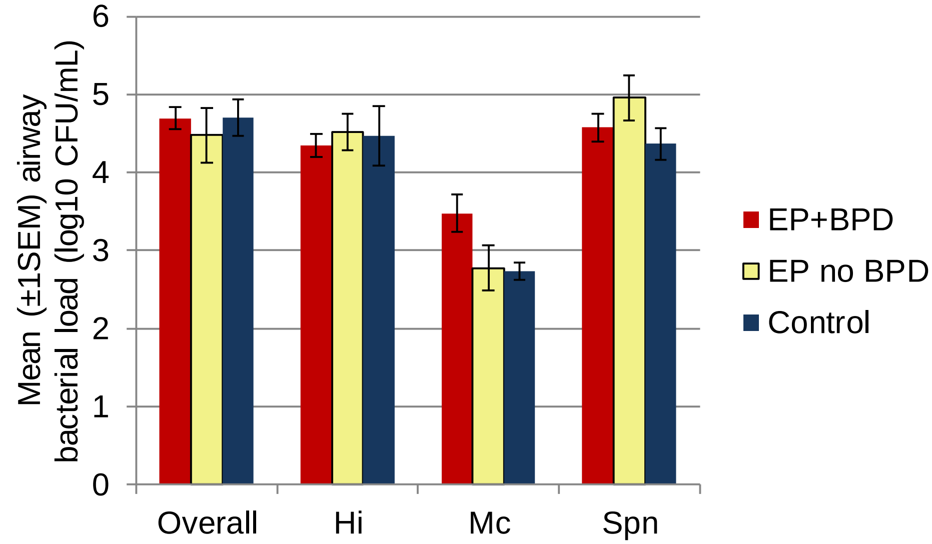


Figure 3: Comparison of *Prevotella* relative abundance (RA) across study groups (A)genus *Prevotella* RA was significantly lower in both extremely preterm (EP) groups regardless the neonatal Broncho-pulmonary Dysplasia (BPD) status compared to controls (P<0.05, by Kruskal-Wallis (KW)) **(B)** genus *Prevotella* RA correlated with the FEV1 z-score (Spearman ρ=0.272, P=0.02) **(C)** OTU 4458304 contributed most to the observed difference in genus *Prevotella* RA across the study groups (P<0.05, by KW) **(D)** Phylogenetic Tree of OTU 4458304 was 100% identical to *Prevotella melaninogenica* strains as obtained by BLAST Analysis of the representative sequence against NCBI 16S ribosomal RNA database. Sample size:29 EP+BPD, 9 EP no BPD and 36 Controls.

A.



C.



B.

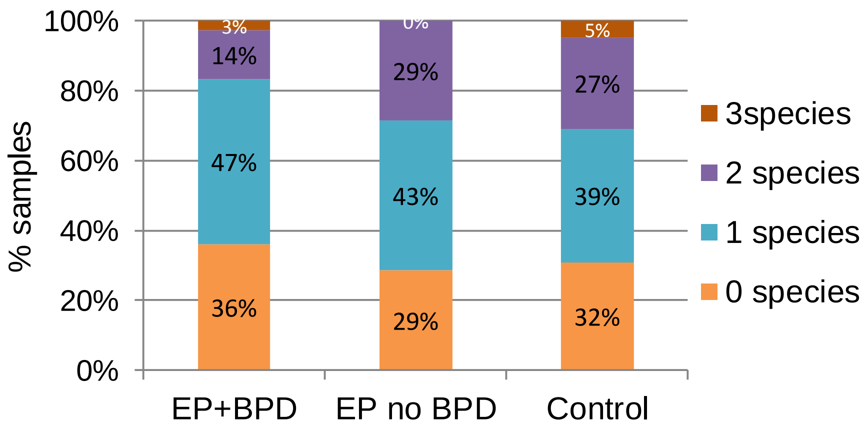
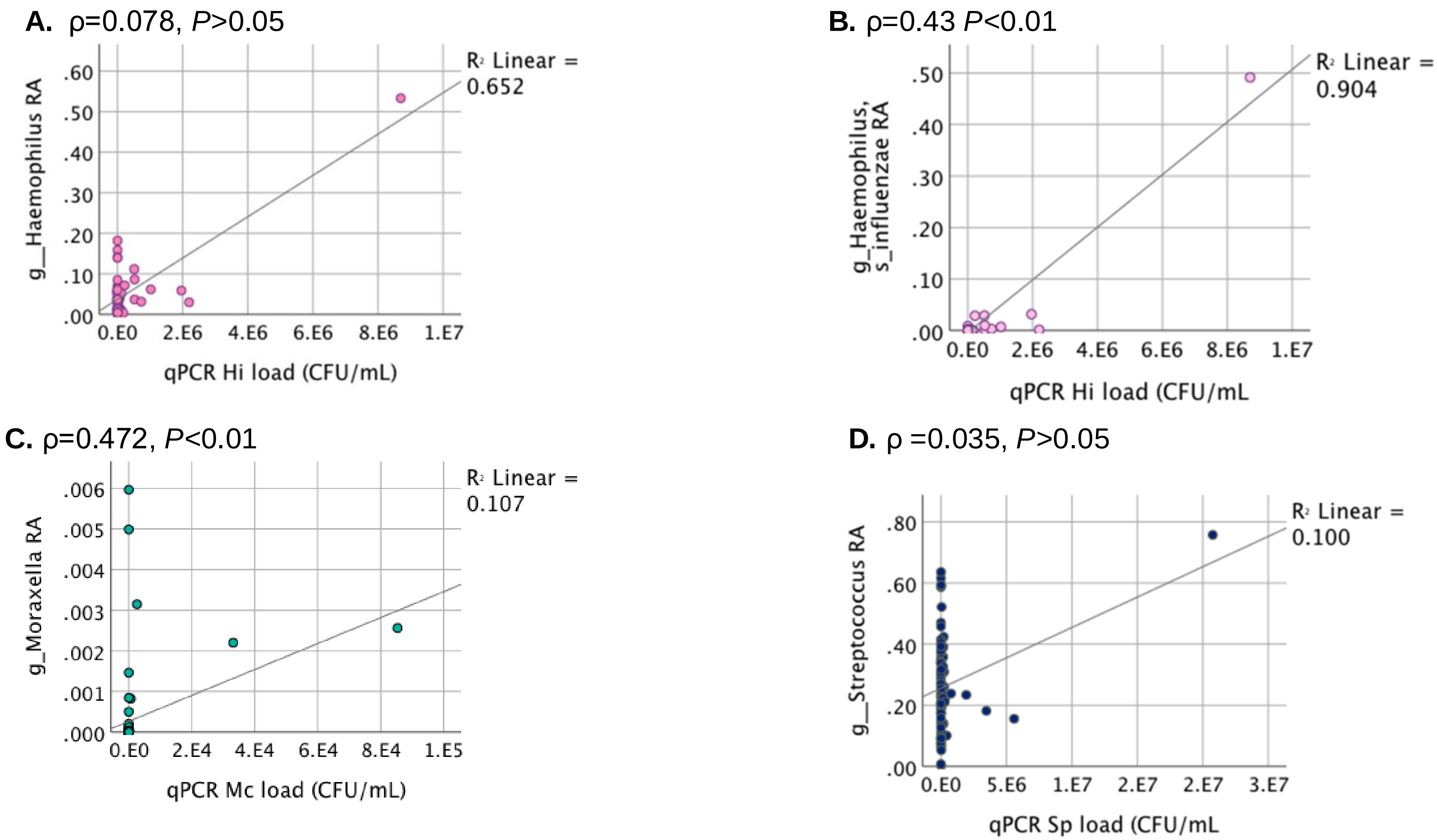


Figure 4: Prevalence and loads of pathogenic airway bacteria; *H. influenzae* (Hi), *M. catarrhalis* (Mc), *S. pneumoniae* (Spn) **(A)** Prevalence of the three airway bacteria within the three study groups (*P*>0.05, Chi squared test) **(B)** Co-existence of the three bacteria in the sputum samples in each group (*P*>0.05, Fisher Exact test) **(C)** Mean bacterial load of each of the three bacteria as determined by the multiplex q-PCR (*P*>0.05, MANOVA). Loads (CFU/ mL) of original sputum sample for each bacterium and the sum of the three (overall) were calculated for each sample, then means were calculated for each study group. Samples which gave negative results for a given bacteria were excluded from the analysis. Error bars show ±1 SEM; EP: extreme preterm birth, BPD: Broncho-pulmonary dysplasia; sample sizes (37, 14, 41) respectively.



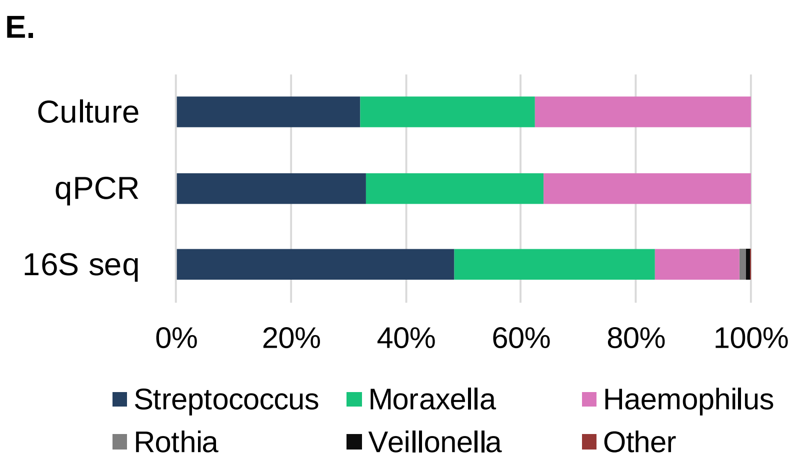


Figure 5: Correlation between the 16S Sequencing Relative Abundance (RA) results and the qPCR results (colony forming unit (CFU)/mL) in sputum samples and mock community

(A) Correlation between relative abundance of genus *Haemophilus* and load of *H. influenzae* (Hi) in sputum samples **(B)** Correlation between relative abundance of *H. influenzae* OTUs and load of *H. influenzae* in sputum samples **(C)** Correlation between relative abundance of genus *Moraxella* and load of *M. catarrhalis* (Mc) in sputum samples **(D)** Correlation between relative abundance of genus *Streptococcus* and load of *S. pneumoniae* (Sp) in sputum samples **(E)** Comparison of the relative abundance of the 3 bacteria in a lab prepared mock community by 16S sequencing, qPCR and quantitative culture. ρ: Spearman's Correlation, Sample size: 74

Online Data Supplement

Sample Collection

Samples were collected in the period between January 2014 and March 2015 at University College London Hospital (UCLH).

Sputum induction was carried out by inhalation of nebulised hypertonic saline: 4% hypertonic saline over 5 or 10 min increasing to 7% over a further 10 min as tolerated in non-asthmatics; or 0.9% hypertonic saline over 5 min increasing to 3% over another 5 min and then 4% for a further 10 min as tolerated in asthmatics. FEV1 and oxygen saturation (SPO2) was monitored every 5 min after baseline and the procedure was stopped if FEV1 dropped to <80% baseline level or if SPO2 dropped to <92% or if the participant requested to stop the procedure, and if required in the event of bronchospasm, 400 μg salbutamol was administered.

Participants were asked to take a sip of water and blow their nose before expectorating to minimise contamination from saliva and post-nasal drip.

As soon as a satisfactory sample was collected, the sample was placed immediately on ice and transported on dry ice to the UCL-RFH biobank lab for processing and freezing within an hour. The samples were split into aliquots and stored at -80°C in the UCL-RFH biobank. One aliquot of each sample was unprocessed and these were selected for the microbiome study. Samples were sorted into three groups by the clinical team in the EPICure study based on medical history of participants; those who were born prematurely with no history of bronchopulmonary dysplasia; those who were born premature with history of bronchopulmonary dysplasia; and full term born controls.

The samples were removed from -80°C freezers and allowed to thaw at room temperature before 500μL of each sample was aliquoted for testing. The samples were treated with an equal volume of freshly diluted Sputasol® (as per the manufacturer's instructions; Oxoid, UK). The samples were thoroughly mixed using a vortex (Clifton™ Cyclone vortex mixer, Nickel-Electro Ltd, UK) for 10 seconds and incubated at room temperature (24-25°C) for 15 min with vortexing 2 or 3 times before they were heated at 95°C for 30 min.

Multiplex qPCR for respiratory pathogens

The master-mix was prepared using Platinum® quantitative PCR Supermix-UDG (Thermo-Fisher Scientific, UK) and additional magnesium chloride at final concentration of 3 mM. The thermo-cycles of 95°C for 3 min followed by 40 cycles of 95°C for 10 sec and 60°C for 45 sec were carried out on Qiagen Rotor– gene® 6000 real-time PCR machine (Corbett Research UK, Cambridgeshire, UK). An internal amplification control, Spud A, was used at final concentration of 0.04 pM to test for PCR inhibition [1].

16S rRNA gene sequencing

A sequence library was created by amplification of V5-V7 regions of the bacterial 16S *rrna* gene through conventional PCR on the extracted metagenomic DNA using 785 forward primer (785F: GGATTAGATACCCBRGTAGTC) and 1175 reverse primer (1175R: ACGTCRTCCCCDCCTTCCTC). Each sample was assigned a unique pair combination of standard Illumina® dual indexed primers (with adaptors attached: P5 and

P7 in the forward and reverse primers respectively). The PCR master-mix per reaction was composed of; 0.4 µM for each of the forward and reverse primers, 0.625 units Mol Taq 16S/18S basic Master-mix (Molzym, VH Bio Limited, UK) with additional 0.5 mM magnesium chloride and 800 µM deoxynucleosides triphosphate (dNTPs) mixture. The amount of DNA template added was adjusted such that the final DNA input per reaction was around 300ng. The thermo-cycling conditions were 95°C for 5 min, followed by 30 cycles of 95°C for 30 sec, 55°C for 40 sec and 72°C for 1 min, in addition to a final extension phase at 72°C for 10 min. Seventy-four samples produced an amplicon at the expected size of 504 bp. The PCR products were cleaned up using Agencourt AMPure XP beads (Beckman Coulter, UK) with a binding buffer of 2.5 M sodium chloride and 20 g% PEG-8000, 80% ethanol and EB Buffer® (Qiagen, UK) to remove amplicons <200bp and primer dimers. DNA in the cleaned products was then quantified using Qubit™ dsDNA HS kit and Qubit® 2.0 Fluorometer (Thermo Fisher Scientific, UK). The Samples were pooled in an equimolar ratio at 10 nM into one library. The library was checked on bioanalyzer. Sequencing was performed using Illumina MiSeq Platform using costume sequencing primers for read 1: ACGTACGTACGTGGATTAGATACCCBRGTAGTC, read 2: AGTCAGTCAGCCACGTCRTCCCCDCCTTCCTC and index i7: GAGGAAGGDGGGGARGACGTGGCTGACTGACT, MiSeq® Reagent Kit v2 (500 cycles) (cat no. MS-102-2003) and PhiX control V3 KIT (cat no. FC-110-3001) as internal control for the sequencing run (Illumina Cambridge, Ltd,UK). The extraction negative control and a no-template PCR control (water) were run throughout the amplification and sequencing process as negative controls to allow for the evaluation of potential contamination.

Bioinformatics and Statistical Analyses

In bioinformatic analysis we adopted the workflow established by Microbiome helper [2]. Briefly, the sequencing reads were primary analysed and demultiplexed and exported to Illumina cloud-based BaseSpace. The paired end reads were stitched together using PEAR.[3]. The low-quality reads with quality score <30 over 10% of its bases and length less than 350 bp were filtered out using FASTX-toolkit (v.0.0.14) [4]. The reads were then screened for possible chimeras that may have resulted from PCR using VSEARCH (v1.11.1) [5]. The subsequent steps were through QIIME pipeline v1.9.1 [6] where the sequences were clustered based on 97% similarity into Operational Taxonomic Units (OTU) and taxonomic classification was assigned to OTUs using open reference OTU picking against Greengenes database version 13_8. The OTU table was then rarefied per sample to 4000 reads removing all samples having number of reads less than 1000 reads (4 samples). Alpha and beta diversity indices were calculated on the rarefied OTU table using QIIME. The appropriate statistical significance tests were calculated using SPSS v. 23 or QIIME wrapper scripts after checking the normal distribution assumption of the continuous variables through Shapiro-Wilk test, Skewness and Kurtosis z-scores, Normal Q-Q Plot and Levene test for homogeneity of variance. Both PERMANOVA and ANOSIM tests were performed on weighted Unifrac distance matrix through QIIME. STAMP (v2.1.3) [7] was used to visualize the results and explore the OTUs showing significant differences across the groups. Whenever applicable the *P*-values were corrected using Benjamini-Hochberd False Discovery Rate (FDR) method for multiple comparisons on filtered OTU tables to compare highly abundant taxa (RA>5%).

Results

Sensitivity and Specificity of methods

Conventional bacteriology has a diagnostic cut-off of 10^6 CFU/mL [8], the qPCR had a sensitivity of 3700 CFU/mL for *S. pneumoniae,* 1000 CFU/mL for *H. influenzae* and 500 CFU/mL for *M. catarrhalis* [9]. The specificity of the multiplex qPCR was previously determined by in-silico analysis of the primers specificity and validated by screening against a range of airway bacteria and viruses (data not showed). 16S rRNA sequencing is a semi-quantitative method in which the results are expressed as the relative abundance. Currently, the resolution of taxonomic classification cannot go beyond the genus level for most OTUs. Comparing the sequencing results with the qPCR results the sensitivity and specificity of 16S rRNA v5-v7 sequencing were 82% and 35% respectively for *H. influenzae*, 100% and 74% respectively for *M. catarrhalis* and 100% and 0% respectively for *S. pneumoniae*. Sensitivity in this context is defined as true positive rate and specificity as the true negative rate [10].

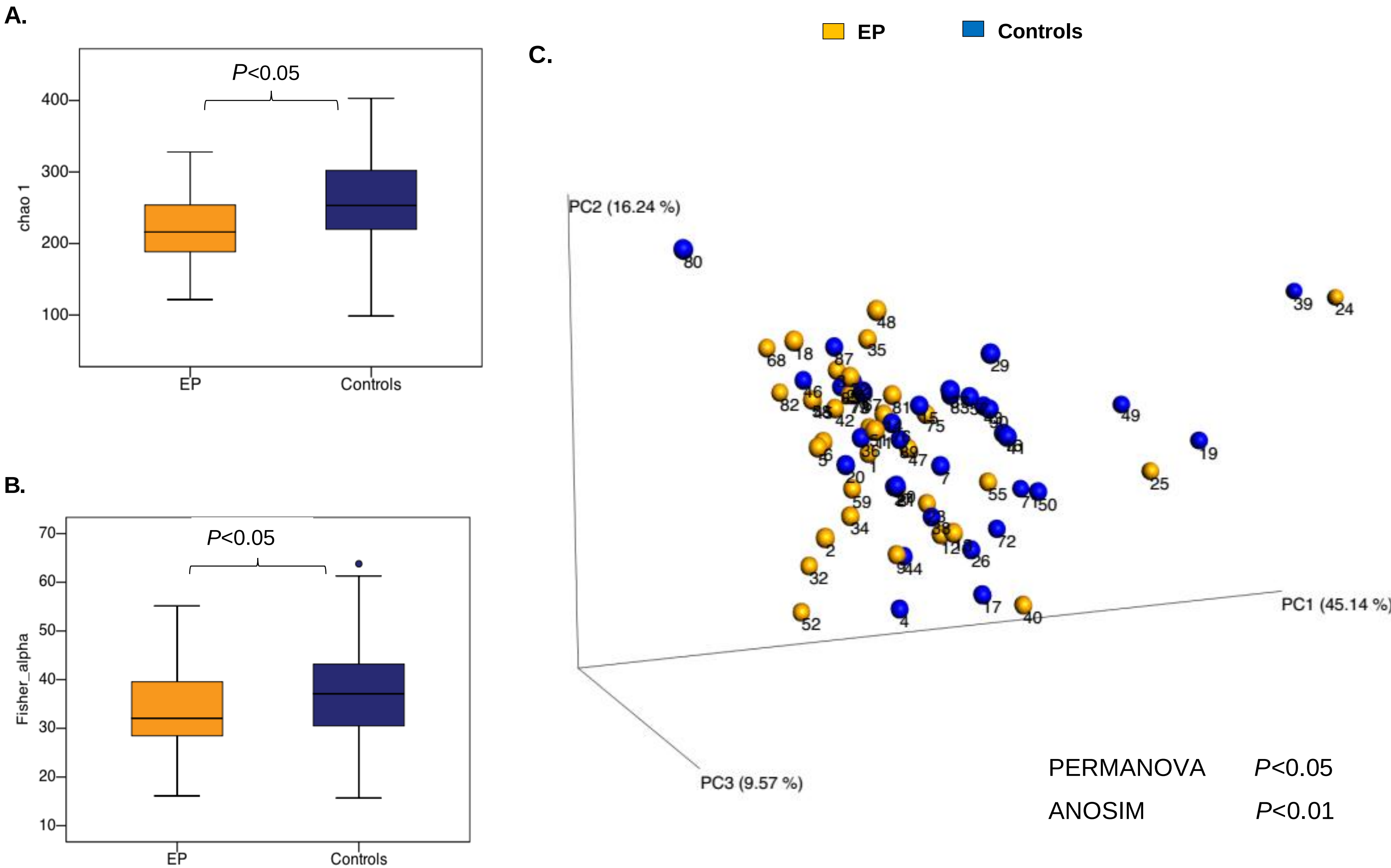


Figure E1: **Comparison of the richness and diversity of microbial communities in sputum between the pre-term birth survivors (EP) (orange) and controls (blue)** Richness and α diversity measured by **(A)** Chao 1 and **(B)** Fisher-alpha diversity index, **(C)** Principal Coordinate Analysis (PCoA) of weighted UniFrac β-diversity index. Sample size: 37 EP and 33 Controls

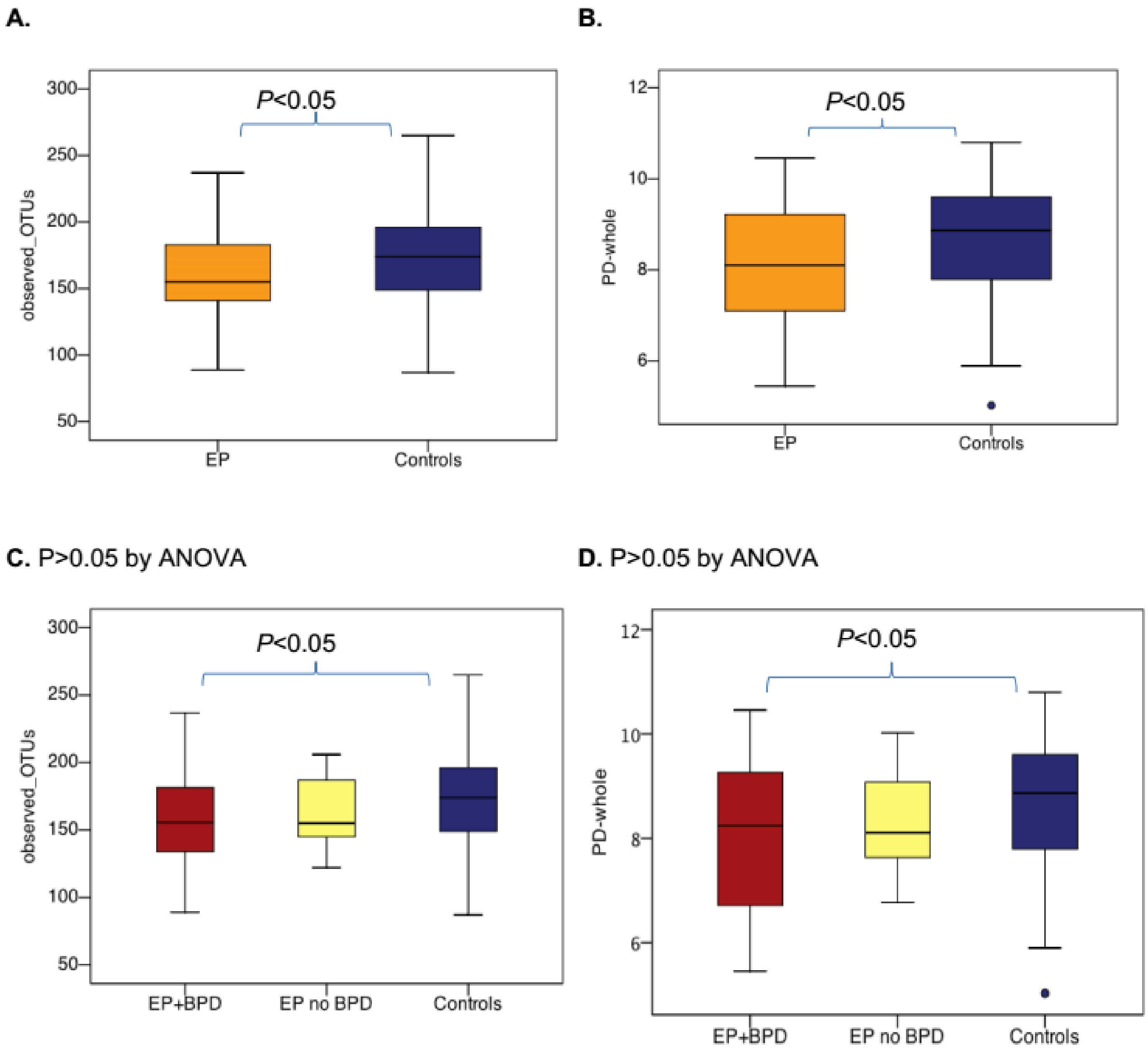


Figure E2: Richness and alpha diversity of the airway microbial communities measured by total number of observed OTU **(A &C)** and PD whole tree **(B&D)** respectively, both were significantly lower in the whole extremely pre-term born (EP)(orange) group (*P*<0.05, T-test) **(A&B)**, the BPD group had significantly less diverse microbial community compared to controls (*P*<0.05, T-test) **(C&D)**, EP+BPD extremely preterm born group with Bronchopulmonary Dysplasia history (n=28), EP no BPD: extremely preterm born group without BPD history (n=9), Controls (blue) (n=33)

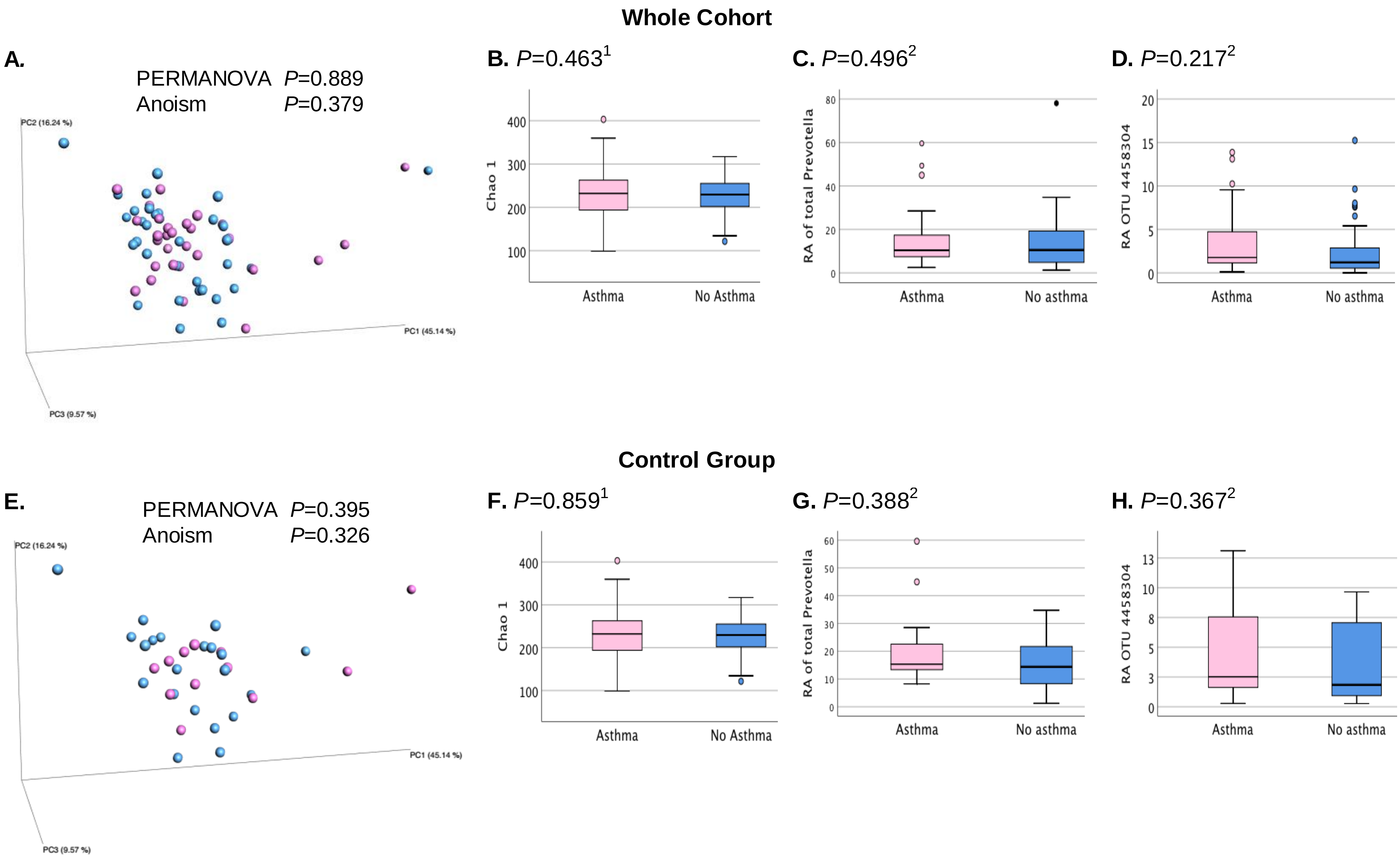


Figure E3: Comparison of weighted Unifrac β diversity index, Chao 1 α diversity index, genus *Prevotella* relative abundance (RA) and RA of OTU 4458304 identified as *Prevotella melaninogenica* in sputum samples from participants who were labelled with asthma and those who were not within the whole cohort (A, B, C and D respectively) and within our control group (E, F, G and H respectively), No significant differences were detected. 1: *P*-values by T test, 2: *P*-values by Mann-Whitney Test

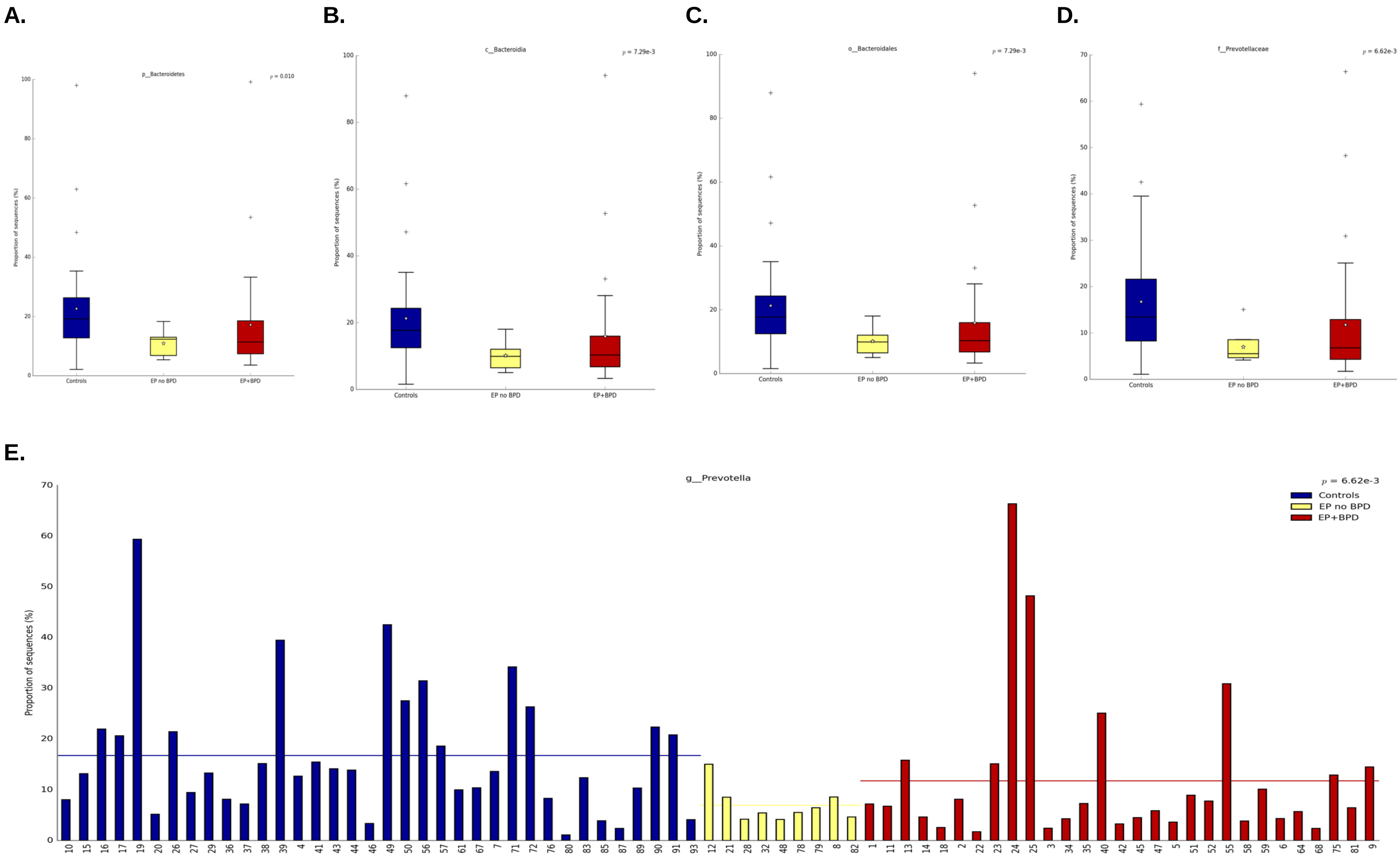


Figure E4: Comparison of relative abundancies of **A**. Phylum *Bacteroidetes*, **B.** Class: *Bacteroidia*, **C.** Order *Bacteroidales*, **D.** Family: *Prevotellaceae*, **E.** Genus*: Prevotella* across the three study groups the Extremely Preterm (EP) group with neonatal Bronchopulmonary Dyspalsia (BPD) (red), EP group without BPD (yellow) and the control group (blue); sample size:29, 9 and 36 respectively.

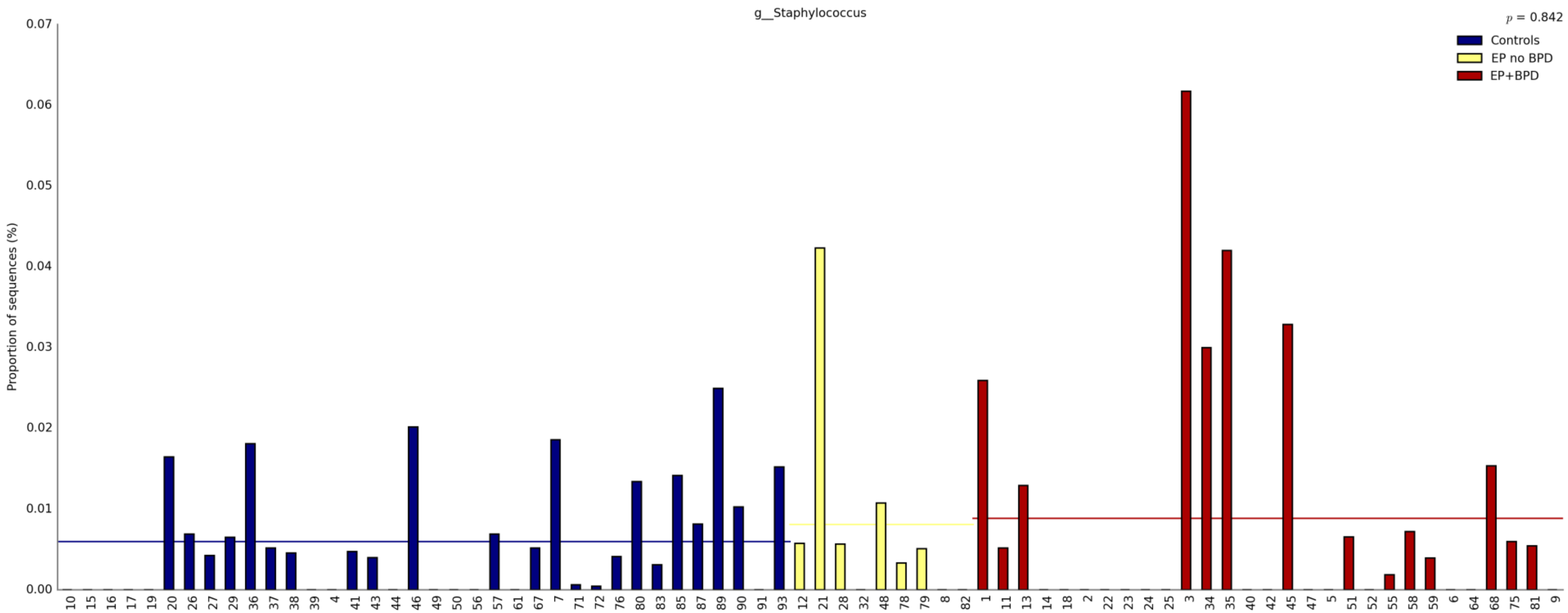


Figure E5: Comparison of the relative abundance of *Staphylococcus* species between study groups EP+BPD: extremely preterm born group with Bronchopulmonary Dysplasia (BPD) history (n=29), EP no BPD: extremely preterm born group without BPD history (n=9) (yellow), Controls (n=36) (blue).

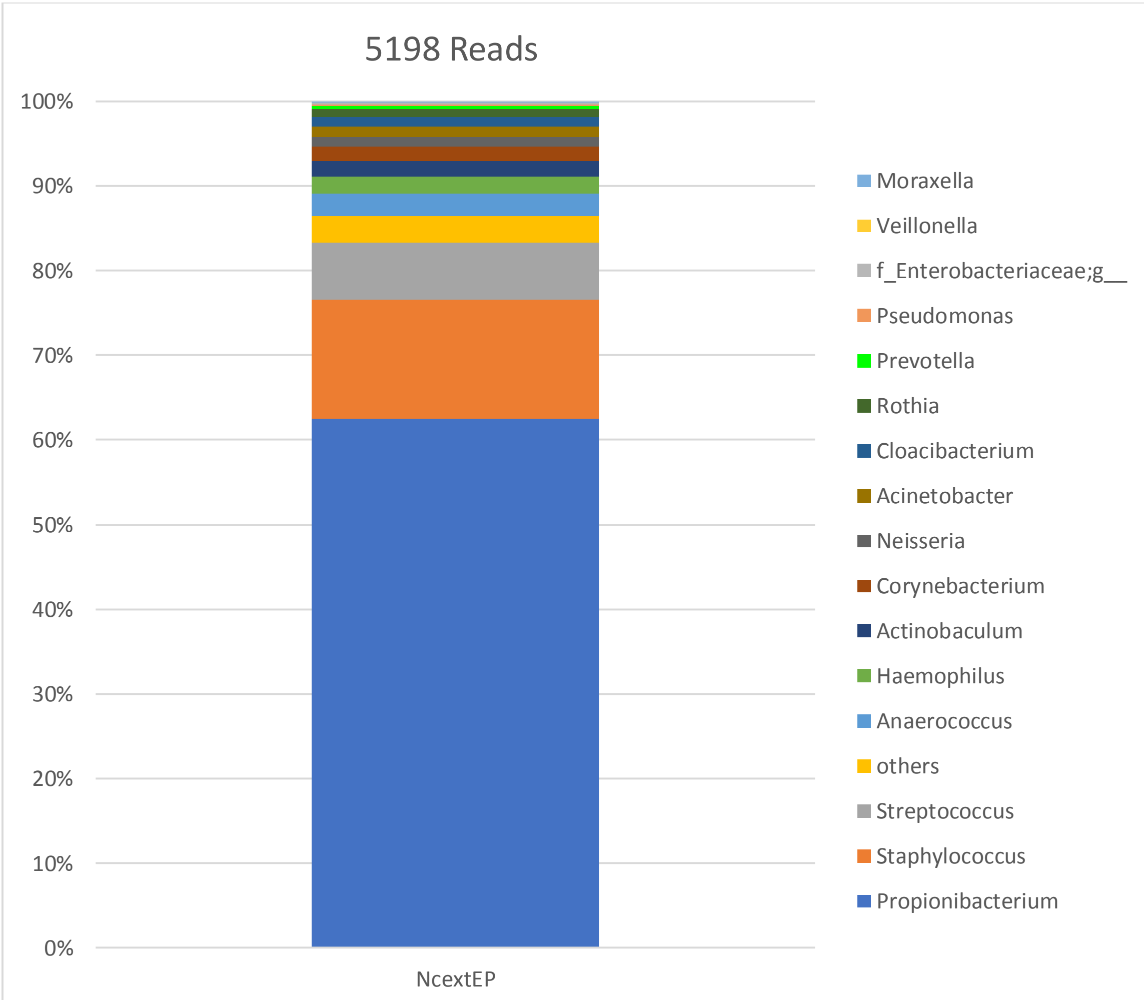


Figure E6: Microbiome profile of the extraction negative control of the saline used for sputum induction (composed of Nebusal 7% by Forest©, and Sodium chloride 0.9% w/v BP by B. Braun© and water of injection BP by B. Braun©)

References

1. Nolan T, Hands RE, Ogunkolade W, Bustin SA. SPUD: a quantitative PCR assay for the detection of inhibitors in nucleic acid preparations. Anal Biochem 2006; 351: 308-310.
2. Comeau AM, Douglas GM, Langille MG. Microbiome Helper: a Custom and Streamlined Workflow for Microbiome Research. mSystems 2017; 2.
3. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 2014; 30: 614-620.
4. A. G. FASTX-Toolkit: FASTQ/A short-reads pre-processing tools. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory 2009.
5. Rognes T, Flouri T, Nichols B, Quince C, Mahe F. VSEARCH: a versatile open source tool for metagenomics. PeerJ 2016; 18.
6. 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, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 2010; 7: 335-336.
7. Parks DH, Tyson GW, Hugenholtz P, Beiko RG. STAMP: statistical analysis of taxonomic and functional profiles. Bioinformatics 2014; 30: 3123-3124.
8. da Silva RM, Teixeira PJ, Moreira Jda S. The clinical utility of induced sputum for the diagnosis of bacterial community-acquired pneumonia in HIV-infected patients: a prospective cross-sectional study. Braz J Infect Dis 2006; 10: 89-93.
9. Kralik P, Ricchi M. A Basic Guide to Real Time PCR in Microbial Diagnostics: Definitions, Parameters, and Everything. Front Microbiol 2017; 8: 108.
10. Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity. Continuing Education in Anaesthesia Critical Care & Pain 2008; 8: 221-223.