

Methodological issues in airway microbiome studies

Christine Drengenes

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2020

UNIVERSITY OF BERGEN



Methodological issues in airway microbiome studies

Christine Drengenes



Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 04.12.2020

© Copyright Christine Drengenes

The material in this publication is covered by the provisions of the Copyright Act.

Year: 2020

Title: Methodological issues in airway microbiome studies

Name: Christine Drengenes

Print: Skipnes Kommunikasjon / University of Bergen

Table of Contents

1. Scientific environment	6
2. Acknowledgements	7
3. Abbreviations	8
4. Abstract	9
5. List of publications	12
6. Introduction	13
6.1 <i>Sampling the lower airway microbiome</i>	14
6.1.1 Sputum sampling.....	15
6.1.2 Bronchoscopy sampling	16
6.1.3 Bronchoscopy and upper airway contamination	18
6.2 <i>Amplicon-based marker gene analyses</i>	22
6.2.1 DNA extraction	24
6.2.2 Library preparation for sequencing.....	26
6.2.3 High-throughput sequencing	31
6.3 <i>Bioinformatic sequence processing</i>	33
6.3.1 Demultiplexing	33
6.3.2 Quality filtering.....	34
6.3.3 Clustering and denoising	37
6.3.4 Chimera removal	38
6.4 <i>Low biomass samples from the lungs</i>	39
6.4.1 Sample bacterial load and contamination	39
6.4.2 Sample bacterial load varies across protocols	41
7. Objectives	44
8. Materials and methods	45
8.1 <i>Study design and participants</i>	45
8.2 <i>Study samples</i>	46
8.2.1 Procedural samples (Papers I, II and III)	46
8.2.2 Procedural control samples (Paper II).....	47
8.2.3 <i>Salmonella</i> dilution series (Paper II).....	48
8.2.4 Mock community (Paper III)	48

8.3	<i>Bacterial DNA extraction (Papers I, II, III)</i>	50
8.4	<i>Quantification of bacterial load (Paper II)</i>	51
8.5	<i>Illumina MiSeq sequencing (Papers I, II, III)</i>	51
8.5.1	MiSeq sequencing setup 1 (Papers I, II, III)	52
8.5.2	MiSeq sequencing setup 2 (Paper III)	53
8.5.3	MiSeq sequencing setup 3 (Paper III)	54
8.6	<i>Bioinformatics sequence processing</i>	55
8.6.1	QIIME1 (Papers I and II)	55
8.6.2	QIIME2 (Paper III)	55
8.7	<i>Decontamination strategies (Papers I, II, III)</i>	56
8.7.1	The remove all approach	56
8.7.2	Decontam	56
8.8	<i>Statistical analyses</i>	57
8.9	<i>Ethics</i>	58
9.	Summary of papers	60
9.1	<i>Paper I</i>	60
9.2	<i>Paper II</i>	61
9.3	<i>Paper III</i>	62
10.	Discussion	65
10.1	<i>Discussion on methods</i>	65
10.1.1	Study populations	65
10.1.2	Procedural samples	67
10.1.3	Bacterial DNA extraction	72
10.1.4	Determination of bacterial load	73
10.1.5	Library preparation for sequencing	74
10.1.6	Bioinformatics processing	78
10.1.7	Analyses	84
10.2	<i>Discussion of main results</i>	86
10.2.1	Paper I	86
10.2.2	Paper II	91
10.2.3	Paper III	94
11.	Conclusions	102

12. Future perspectives.....	103
13. References.....	104
14. Papers and supplementary material	116

1. Scientific environment

During the PhD period, I was affiliated to the Bergen Respiratory Research Group at the Department of Thoracic Medicine, Haukeland University Hospital and Department of Clinical Science, University of Bergen. Main supervisor was Associate Professor Rune Nielsen. Co-supervisors were Professor Tomas M. L. Eagan and Professor Harald G. Wiker.

The PhD project was conducted as part of the Bergen COPD Microbiome study (short name «MicroCOPD»). The MicroCOPD study was funded by unrestricted grants and fellowships from Helse Vest, Bergen Medical Research Foundation, the Endowment of timber merchant A. Delphin and wife through the Norwegian Medical Association and GlaxoSmithKline through the Norwegian Respiratory Society. The funding bodies had no role in the design of the study, data collection and analysis, interpretation of data, or in writing the manuscripts for the current thesis.

2. Acknowledgements

I would first of all like to express my gratitude to supervisors Professor Tomas M. L. Eagan and Professor Harald G. Wiker for giving me the opportunity to work on this project, and for always having an open door.

To main supervisor Associate Professor Rune Nielsen, thank you for your patience, motivation and encouragement throughout the project. I have learned so much under your guidance that I know I will bring with me moving forward.

I also want to thank Øyvind Kommedal, for your enthusiasm and for sharing your knowledge in microbiology.

To my colleagues on the «fifth floor», thank you for making every work day, a day to look forward to – especially Sonja Ljostveit, Tove Skogmo, and Tharmini Kalanathan. A special thank you to Professor Audun Nerland, for sharing your expertise and your encouraging words. It has meant a lot.

To colleagues Ingvild Haaland, Einar Marius Martinsen, Elise Leiten, Bahareh Jouleh, Solveig Tangedal and Marta Erdal – I have enjoyed your company traveling to conferences around the world, coffee breaks and insightful discussions.

A very special thank you to Eli Nordeide, for your kindness and for always being there.

To my parents. I would not be where I am today without your continuous support and love. You have taught me that nothing in life comes for free and to aspire to do great things, one has to put in the work. And I can honestly say that without your countless hours of help babysitting, I would not have been able to put in the work to complete this thesis. Thank you!

And last but not least, to the two most important people in my life - Thea and Andreas. Thank you for always reminding me of what is really important. I am so proud of you both.

3. Abbreviations

BAL: bronchoalveolar lavage

BR: bronchoscope rinse

COPD: chronic obstructive pulmonary disease

CR: catheter rinse

CT: cryotube

DADA2: divisive amplicon denoising algorithm 2

ICS: inhaled corticosteroids

LLL: left lower lobe

LUL: left upper lobe

MC: mock community

NCS: negative control sample

OTU: operational taxonomic unit

OW: oral wash

PBAL: protected bronchoalveolar lavage

PBS: phosphate-buffered saline

PE: paired-end sequencing

PSB: protected specimen brush

QIIME: Quantitative Insights into Microbial Ecology

RLL: right lower lobe

RML: right middle lobe

SBS: sequencing by synthesis

SDS: *Salmonella* dilution series

SVL: small volume lavage

4. Abstract

Background

Studies on the lung microbiome face unique methodological challenges tied to the low bacterial load of acquired samples and the increased susceptibility to bacterial DNA contamination. Contamination may be introduced from i) the upper airways during sampling and ii) reagents, kits and the general laboratory environment during laboratory processing steps. Few publications exist on validity and reliability of applied methods of sampling, laboratory processing and bioinformatics analysis.

Objectives

The objective of the thesis was to address some of the methodological issues that remain unresolved in the field of lung microbiome research. In the first paper, we sought to determine whether protected (via a sterile catheter) bronchoscopic sampling techniques would reduce the influence of bronchoscopic carryover from the upper airways. In paper II, we examine the impact of laboratory contamination on airway samples and explore the expected inverse relationship between sample bacterial load and influence of contamination. We also compare different bioinformatic strategies to dealing with contamination. In paper III, we sought to determine whether processing samples through longer laboratory workflows would increase susceptibility to contamination, and to explore impact of choice of 16S rRNA gene variable region (V3 V4 or V4) on the presentation of the airways microbiome.

Methods

Study samples were collected from participants enrolled in the Bergen COPD Microbiome study (short name «MicroCOPD»). Samples included oral washes (OW), bronchoscopically acquired protected specimen brushes (PSB), protected bronchoalveolar lavages (PBAL), small-volume lavages (SVL) and negative control samples (NCS) consisting of PBS used for collection of all samples.

Bacterial DNA was extracted using a combination of enzymatic and mechanical lysis methods and processing through the FastDNA Spin Kit (MP Biomedicals). Bacterial community composition was determined by high-throughput sequencing of the bacterial 16S rRNA gene using the Illumina MiSeq sequencing platform. Three library preparation setups were included in the thesis, varying in number of PCR steps (1- or 2-steps) and target marker gene region (16S rRNA gene V3 V4 or V4): Setup 1 (2-step PCR; V3 V4 region); Setup 2 (2-step PCR; V4 region); Setup 3 (1-step PCR; V4 region). Papers I and II were based on setup 1. Paper III included all three setups. Bacterial load was determined by quantitative PCR targeting the 16S rRNA gene region V1 V2 (paper II).

Bioinformatics processing steps were performed using the Quantitative Insights Into Microbial Ecology (QIIME) bioinformatic package, versions 1 (papers I and II) and 2 (paper III). Strategies for decontamination varied across papers and included i) keeping samples intact (i.e. do nothing), ii) removing all sequences observed in NCS and iii) the removal of sequences identified as contaminants using the Decontam R package tools. In paper I, sequences observed in NCS were removed. In paper II, all three strategies were applied and compared. In paper III, Decontam was used.

Results

Analyses for paper I were based on the underlying assumption that the more similar the bronchoscopically acquired specimens (PSB, PBAL and SVL) were to the OW sample, the greater the influence of upper airway contamination. Between sample comparisons were made based on three parameters: i. taxonomy, ii. alpha diversity and iii. beta diversity. Across all three parameters, similarity to the OW sample decreased in order SVL>PBAL>PSB.

In paper II, an estimated 10-50% of the bacterial community profiles for the lower airway samples (PSB, PBAL) were derived from laboratory contamination. This was determined based on comparison to a dilution series of known bacterial composition and load. The DNA extraction kit was identified as the main contamination source. On comparison of the three decontamination strategies, we found that the Decontam

R package provided a balance between keeping and removing sequences found in both NCS and study samples.

In paper III, we found that the number of sequences and ASVs decreased in order setup1>setup2>setup3. This appeared to be associated with increased taxonomic resolution when targeting the V3 V4 region (setup 1) and an increased number of small ASVs in setups 1 and 2. For setups 1 and 2, we interpreted this as a result of contamination in the 2-step PCR protocol and sequencing across multiple runs (setup 1). Analyses of taxonomic composition revealed that genera *Streptococcus*, *Prevotella*, *Veillonella* and *Rothia* dominated all setups, but that relative abundances differed. Analyses of beta diversity revealed that while OW samples clustered together regardless of number of PCR steps, samples from the lower airways (PSB, PBAL) separated. Removal of contaminants identified in Decontam did not resolve differences across setups.

Conclusions

We show that protected bronchoscopic sampling techniques (PSB, PBAL) may provide protection from oropharyngeal carryover and should be the preferred sampling technique in future studies (paper I).

We demonstrate that bacterial load will vary across airway sample types and that bacterial contamination from the laboratory will have an increased impact on samples of lower bacterial load (paper II). We recommend that estimates of contamination are reported in all studies. We also recommend the use of contaminant identification tools based on statistical models that limit subjectivity (e.g. Decontam).

Finally, we demonstrate that differences in number of PCR steps (1- or 2-steps) will have an impact on final bacterial community descriptions, and more so for samples of low bacterial load (e.g. lower airway samples) (paper III). Our findings could not be explained by differences in contamination levels alone, and more research is needed to understand the underlying mechanisms contributing to the observed protocol bias.

5. List of publications

Paper I

Grønseth R, Drengenes C, Wiker HG, Tangedal S, Xue Y, Husebø GR, *et al.*
Protected sampling is preferable in bronchoscopic studies of the airway microbiome.
ERJ Open Res. 2017;3

Paper II

Drengenes C, Wiker HG, Kalanathan T, Nordeide E, Eagan TML, Nielsen R.
Laboratory contamination in airway microbiome studies. BMC Microbiology.
2019;19

Paper III

Drengenes C, Eagan TML, Haaland I, Wiker HG, Nielsen R. Exploring protocol bias
in airway microbiome studies: one versus two PCR steps and 16S rRNA gene regions
V3 V4 versus V4. Manuscript (Under revision, BMC Genomics)

6. Introduction

Chronic obstructive pulmonary disease (COPD) is a progressive disease of the lower airways, characterized by chronic inflammation and airflow obstruction [1]. Much research has been directed towards understanding the factors triggering the onset and progression of COPD. We have learned that there may be a genetic component (most important being alpha-1 antitrypsin deficiency) [2], and that smoking, air pollution, and occupational exposures [3] are important risk factors. However, enough understanding to enable the development of effective therapies, or the prevention of disease development has not been reached using current research methods.

Advancements in DNA sequencing technologies have however provided us with a new angle by which we can study the airways and airway diseases. Through massive parallel sequencing of the bacterial 16S ribosomal RNA gene (16S rRNA gene), we can establish the type of bacteria present in a sample and their relative abundances. The application of sequence-based techniques has already revolutionized our understanding of the role of bacteria in the lower airways – the biggest revelation so far being that the lungs are not sterile, even in healthy individuals [4]. Research efforts have since this newfound understanding been directed towards finding a potential link between the bacterial communities of the airways and the development and progression of disease, COPD being most studied. Progress in this relatively new field of lung “microbiome” research has in large been characterized by an urge to rapidly publish data comparing healthy and diseased states. Few publications however exist on validity and reliability of applied methods of sampling, laboratory processing and bioinformatics analyses.

The aim of the current PhD work is to address some of the methodological issues that remain unresolved in the field of lung microbiome research. In the following introductory sections (6.1 - 6.3), the general workflow for generating data on the bacterial component of the airway microbiome is described from steps of sampling, to library preparation for high-throughput sequencing, and finally bioinformatics analyses. The focus of discussion will be on inherent sources of bias and error, many

of which pertain to all fields of microbiome research regardless of site being studied. In the final introductory section (6.4), specific challenges related to the low bacterial load of the lower airways, and the resulting increased susceptibility to contaminating bacterial DNA is discussed.

In the literature, there is some inconsistency in the definition and usage of the terms “microbiome” and “microbiota”, and the terms are often used interchangeably [5]. Herein, the term “microbiota” is used to describe the microorganisms that make up a sampled community. “Microbiome” is used to describe the collection of genomes from these microorganisms. As the basis of the analyses for the current PhD work is the bacterial 16S rRNA gene, the usage of the terms “microbiome” and “microbiota” is limited to bacteria.

6.1 Sampling the lower airway microbiome

Obtaining valid (uncontaminated) lower airway microbiome samples is challenging. First of all, the lower airways are relatively inaccessible. Although percutaneous procedures exist, most sampling procedures must involve the passage of a sample (e.g. sputum) or sampling device (e.g. bronchoscope) through the upper airways. Regardless of route of sampling, which may be performed via the oral or nasal passage, contamination from the upper airways is more or less inevitable. Adding to the severity of the contamination issue is the difference in bacterial load between the upper and lower airways, which has been measured to be several logs greater in the upper airways [6]. Therefore, even minute amounts of carryover from the upper to lower airways during sampling may be enough to confound the analyses of a lower airway sample – this effect has however not been thoroughly studied in the literature. Furthermore, natural processes connecting the upper and lower airways (e.g. microaspiration, mucosal dispersion, inhalation), lead to an expected overlap (or similarity) between the sampled microbiota of the two sites [6, 7]. Recognizing when a sample is contaminated or not remains a challenge.

Despite the aforementioned issues, the field of lung microbiome research has pushed forward and studies have been published using a wide range of sampling techniques - often with little concern about the potential for upper airway contamination beyond its mention as a potential weakness in the discussion section of their reports [8]. The degree to which different sample types are of sufficient quality for microbiome analyses, particularly in terms of minimizing the influence of upper airway contamination, has not been thoroughly evaluated in the existing literature. A discussion on the most commonly used sampling techniques for studying the lung microbiome (sputum and bronchoscopy) is given below.

6.1.1 Sputum sampling

A priori, sputum samples are the most vulnerable to upper airway contamination. During the sampling procedure, sputum (i.e. mucus) is coughed up from the lower airways and expelled from mouth and into a container [9]. By passage through the mouth, the sample is in direct contact with the bacterial communities of the upper airways, rendering contamination more or less inevitable. Despite this, many researchers have still opted for sputum sampling. It is inexpensive, non-invasive, repeatable and used routinely in the clinical setting. In addition, samples can be readily acquired from most subjects, irrespective of age or health status. The debate on the validity of microbiome studies based on sputum sampling, however persists [10].

Besides upper airway contamination, several other factors must be considered when sampling sputum. First of all, sputum can be collected either spontaneously or when sputum production is low, or the procedure is difficult for the subject (e.g. children), it may be induced [9]. Sputum induction involves the inhalation of nebulized hypertonic saline solution that triggers mucus secretion and irritation, leading to coughing. Induced sputum may be collected at different times. Based on the analysis of cell (e.g. neutrophils, alveolar macrophages) and protein (e.g. mucin, SP-A) composition in studies on inflammatory markers, it has been proposed that earlier samples are representative of the proximal airways, whereas later samples are representative of the distal airways (i.e. alveoli) [9, 11]. Spontaneous sputum in turn,

is most likely representative of the proximal airways. Provided one accepts that different bacterial populations are found within the different regions of the lungs (e.g. as proposed by the adapted island model of lung biogeography [12]), one may expect that different microbiota populations will also be represented when sampling by one or the other technique (spontaneous or induced) or when samples are collected at different time points (induced). It is currently unclear whether spontaneous and induced samples can be used interchangeably in studies of the airway microbiome, and studies addressing the issue have been conflicting [10, 13].

6.1.2 Bronchoscopy sampling

Sampling by bronchoscopy is currently considered the gold standard in the lung microbiome field. Bronchoscopy, or flexible video bronchoscopy is an endoscopic technique for examination and sampling of the airways and lungs. The endoscope (i.e. bronchoscope) is inserted through the mouth or the nose, and under local or general anaesthesia passed through the vocal cords and into the lower airways. The flexible bronchoscope has a diameter of 2-7 mm and contains a working channel that is used for instillation of fluids (medication, sampling fluids), as well as insertion of various instruments for sampling or delivering treatment in the airways and lungs. A number of different sample types can be collected through the bronchoscope including endo- or transbronchial biopsies, bronchoalveolar lavages (BAL) and specimen brushings (SB). BAL and SB are most commonly used for sampling the microbial communities of the alveolar space and conducting airways, respectively.

Sampling BAL involves instilling a set volume of liquid into the lower airway region to be sampled, and then suctioning the liquid back through the bronchoscope working channel. When sampling BAL, an estimated 1/40 of the total lung surface area (i.e. 17500 cm²) is covered [14]. The amount of liquid instilled may vary according to both the subject being examined and what the examiner is looking for. The amount of volume returned is lower than that instilled and may differ between study subjects as a natural consequence of anatomical variations and diseases of the lung and airways. The manner by which BAL sampling is performed has not been standardized in studies of the lung microbiome. BAL may be fractionated, for example by instilling

2x50 mL liquid in turn from the same segment. The resulting «BAL return 1» and «BAL return 2» may be pooled together or kept separate. SB sampling involves passing a specimen brush through the working channel to the sampling point and brushing the targeted region. SB sampling typically covers 1 cm² of the airway mucosa - i.e. a significantly smaller area than that which is covered when sampling BAL.

Sampling by bronchoscopy (whether by BAL or SB) comes with the added advantage of enabling a more targeted sampling than sputum. This however means that also the «biogeography» of the lower airways must be considered when deciding on an appropriate sampling scheme. Dickson et al. [12, 14] introduced “the adapted island model of lung biogeography” to explain differences in microbiota that one may expect to find across lung sites in health. The model assumes the upper airways are the main source community for the lower airways, and that upper-lower airway similarity will decrease as one moves further down the lower airways. Further the model assumes that bacterial communities in the lung do not replicate and that the bacterial composition at any one site is determined by processes of immigration (e.g. inhalation, microaspiration and mucosal dispersion) and elimination (e.g. the «mucociliary escalator», cough, local host immune cells) [14]. Particularly illustrative was their observation that brushings from the right upper lobe were more similar to the upper airways, than that from the left upper lobe [12]. The authors explained these findings as a likely result of differences in the angle by which the left and right main bronchi leave the trachea; the sharper upward angle of the left main bronchus appears to direct microaspirated bacteria down the right main bronchus. Importantly, the authors also conclude that the observed variation found across intrapulmonary sites within one subject, is less than that observed across different subjects – the take home message being that in health multiple sampling of different sites within the lungs may not be important.

The diseased lung however is a different matter. Willner et al. [15] show in their study that variation exists between different sites in the CF lung. Erb-Downward et al. [8] found the same in the study of the COPD lung. These changes are likely a

result of regional differences in the lung that may occur in the diseased state, and that hinder processes of elimination and result in favorable conditions for growth of certain bacteria. Thus, particularly in studies including subjects with disease, it may be important to sample multiple sites within the lungs.

6.1.3 Bronchoscopy and upper airway contamination

Although a more «protected» approach than sampling sputum, bronchoscopy is not without risk of contamination from the upper airways. To reach the lower airway sampling point, the bronchoscope must pass through the upper airways, either via the oral or nasal route. Both routes are heavily populated with distinct upper airway microbiota [4, 6, 7, 16, 17], and both the outside of the scope and the inner working channel may carry with it contamination from the upper airways.

To minimize the risk of bronchoscopic carryover from the upper to lower airways, several different preventative measures have been observed in the literature. In some studies participants have been instructed to rinse their mouths with antiseptic mouthwash (e.g. Listerine) prior to sampling [7, 18, 19]. In other studies, investigators have attempted to avoid the high bacterial load of the oral cavity by sampling via the nasal rather than the oral route [8, 20, 21]. However, as described in more detail later, the common passage through the supraglottic region, leaves the effect of choice of sampling route questionable. Most studies report that they avoid suctioning prior to passage through the vocal cords [6, 7, 12, 18, 20]. However, this likely provides little protection against the influence of contamination from the outside of the bronchoscope. Most studies also report that when sampling multiple lower airway sites, care is taken not to retract the bronchoscope back up through the upper airways (i.e. above the vocal cords) [7]. Furthermore, several studies have performed SB sampling through a sterile wax-plugged catheter passed through the bronchoscope working channel. The resulting sample is commonly referred to as a “protected” specimen brush (PSB), reflecting findings from a culture-based study indicating that sampling via a sterile inner catheter (preferentially with a plug at the scope tip) provides protection from contaminants found within the scope channel [22]. These effects have however not been thoroughly studied in the context of the

more sensitive culture-independent methods used in microbiome studies. The MicroCOPD study [23] (for which the work for the thesis is a part of), is the only study for which protected bronchoalveolar lavage (PBAL) sampling has been performed.

Few studies on the lung microbiome have directly examined the relative contributions of contamination introduced from the outside and/or inside of the bronchoscope working channel – for the studies that have attempted to do so, conclusions have been contradicting. Charlson et al. [6] sought to examine the contamination picked up by the bronchoscope after passage through the upper airways via the oral route. They passed the bronchoscope to the supraglottic region (i.e. above the vocal cords) and back. Samples were collected from the outside bronchoscope tip and the inner bronchoscope working channel. They found that their samples were indistinguishable from oral (OW) and oropharyngeal (OP) samples and lower airway samples (BAL and PSB), but distinct from nasopharyngeal (NP) and negative control samples [6]. Their results therefore confirmed that both the outside and the inside of the bronchoscope working channel carried contaminants from the upper airways that could confound their analyses of the lower airways. Dickson et al. [14] examined the influence of bronchoscopic carryover when sampling PSB via the oral route. They passed the bronchoscope to just below the vocal cords, where they performed PSB sampling of the lumen space. In this study, the samples were indistinguishable from negative control samples and the authors concluded that bronchoscopic carryover had minimal influence on the analysis of their lower airway samples [14]. The two studies are however not directly comparable. While Charlson et al. actively sampled the outside and inside of the bronchoscope channel, Dickson et al. only sampled the airway lumen and an eventual coating of biofilm that the protected brush comes near when the wax-plug of the PSB is ejected. When actual sampling of the mucosal wall is performed, there is likely a greater risk of direct contact between the bronchoscope tip and the sampled microbiota. Thus, while protected sampling appears to provide an efficient barrier against contaminants found within the working channel, the degree of protection against contaminants found on the outside of the bronchoscope remains unclear.

To validate their studies, investigators have used indirect methods to show that their lower airway samples are authentic (i.e. uncontaminated) representations. A common argument is based on the expectation of a “dilution” effect in serially sampled BAL [6, 7, 14, 20, 24]. The assumption is that if the bacterial communities detected in the lower airways are a result of upper airway contaminants having been brought down by the bronchoscope, the similarity of the samples to the upper airways will decrease with each successive sampling event. Charlson et al. [6] used measures of bacterial load to conclude on the degree of upper airway carryover by the bronchoscope. They observed a dilution effect between the first and second BAL return collected from the same wedged position at sampled site A of the right middle lobe (RML). When comparing BAL return 2 (site A) with a third BAL collected from an adjacent site B (also RML), they found similar levels of bacteria. They concluded that upper airway carryover mainly influenced the first BAL return.

Not all studies have observed a dilution effect, and this has been used as evidence that upper airway contamination is negligible. Segal et al. [20, 25] and Bassis et al. [7], for instance did not observe a dilution effect in their respective studies when comparing BAL samples collected from the lingula and RML. In contrast to the previous study, where serial BAL was collected from the same wedged position, herein the samples compared were collected from different lungs (lingula and RML). An alternative explanation for the lack of an observed dilution effect may therefore be that the dilution effect was masked by the introduction of intrapulmonary contamination when repositioning the scope for sampling of the second site. If we accept the model of bacterial topography in the lungs as presented by Dickson et al. [14], the pulmonary site for which lung microbiota can be expected to be the most similar to the upper airways is the carina. When sampling across the left and right lungs, it can therefore become difficult to distinguish between carryover from the upper airways and intrapulmonary contamination – and consequentially so, difficult to conclude on the presence or lack of a dilution effect. In addition as described earlier, Dickson’s adapted island model of lung biogeography [12], predicts that microaspirated bacteria will be directed down the right main bronchus rather than left

bronchus. Thus we can expect a greater inherent similarity of microbiota, in terms of composition and load, between the upper and lower airways when sampling the RML than when sampling the lingula. Based on theoretical models, we are therefore reminded that observations of a dilution effect should perhaps be limited (at least) to samples from the same lung.

A second argument commonly used to show that samples are uncontaminated by the upper airways is the observation that similar community descriptions are obtained when sampling via the oral or nasal route [20, 21]. Since these two sites hold distinct microbiotas, the argument is based on the expectation that if contaminated, the community descriptions should reflect one or the other of these two source communities. However, little is mentioned about the fact that both the nasal and oral cavities funnel to a common passage located above the vocal cords at the entry to the lower airways (i.e. the supraglottic region). In theory, one may expect that this region will hold bacterial communities more similar to the oral cavity than the nasal cavity due to an increased flow of saliva relative to nasal fluid (in health) [7]. Thus, an alternative interpretation of the observed similarity of lower airway community descriptions when sampling by either nasal or oral route is that the contamination signal reflects the supraglottic region, which in turn likely reflects a composite signal from both nasal and oral sites – and is likely dominated by the communities that resemble that found in the oral cavity.

In summary, sampling the lower airways is difficult due to the potential confounding issue of contamination from the upper airways. As described, different studies have come to different conclusions regarding the impact and degree of upper airway carryover being brought down by the bronchoscope during sampling. It can perhaps be agreed upon that protected sampling procedures (e.g. PSB) appear to provide an efficient barrier to upper airway carryover from within the scope channel. Despite this, no study has examined the benefits of sampling protected BAL (PBAL) in studies of the lung microbiome.

6.2 Amplicon-based marker gene analyses

Amplicon-based marker gene sequencing workflows are most common in studies of the bacterial airway microbiome. Although much variation exists across protocols, most all laboratory workflows fit into the general framework outlined in Figure 1.

There are two key pieces of information that are obtained using amplicon-based marker gene analyses targeting the 16S rRNA gene. First of all, we are able to establish the type of bacteria that are found in the sampled community (i.e. membership), as each amplicon sequence will reflect a bacterial taxa which can be identified when matched up against a database of known sequences. Second, we are able to determine the relative abundances of each of these members, based on the relative proportions of the different amplicons in the amplicon pool.

The accuracy of amplicon-based marker gene analyses, will depend on the degree to which information regarding both membership and relative abundance is accurately transferred through each step of the laboratory workflow – including all steps of DNA extraction, library preparation for sequencing and sequencing itself. It is well established that both error and bias may be introduced at multiple steps within this framework. Herein, errors are defined as inaccurate representations of the marker gene sequence. Bias is used to describe inaccurate representations of the relative abundances of bacterial community members [26].

In the following section, each step outlined in Figure 1, will be described in turn, with emphasis on methodological pitfalls along the way that may introduce error and/or bias to the sequencing data. In section 6.3 current bioinformatic approaches to dealing with some of these issues will be reviewed.

6.2.1 DNA extraction

The DNA extraction step (Figure 1, step A), has been recognized as one of the main sources of bias in the general microbiome field. While protocols for DNA extraction may vary in many regards, most important is perhaps the chosen method for bacterial cell lysis, for which there are many different examples in the literature (e.g. mechanical [6, 12], enzymatic, chemical or a combination [27]). Differences in cell wall structure across bacteria will render different types of bacteria more or less resistant to the various methods of cell lysis. If these differences are not accounted for, we can obtain a biased picture of the sampled community, already at this first step of the sequencing workflow.

Peptidoglycan is an important structural component of the bacterial cell wall, and the main target in most cell lysis methods [28]. Peptidoglycan consists of chains of alternating N-acetylglucosamine (NAG) and N-acetylmuramic acid (NAM) sugar derivatives [28]. These chains are in turn linked together via short peptides. In the broadest of terms, we can distinguish between two groups of bacteria classified according to cell wall structure - the gram-positive and the gram-negative bacteria. The gram-positive bacteria have a thicker peptidoglycan layer than the gram-negative bacteria and therefore the former are considered more resistant to most cell lysis procedures. However, depending on choice of lysis method, also more subtle differences in peptidoglycan structure may be important. When using enzymatic lysis methods for instance, small differences in peptidoglycan structure can render some bacteria more or less vulnerable to the lytic activity of a particular enzyme [29, 30]. An example of this involves the commonly used enzyme lysozyme, which exerts its lytic activity by cleaving the glycosidic bond between NAG and NAM. For bacteria with O-acetylated NAM (e.g. *Neisseria gonorrhoeae*, *Staphylococcus aureus*), lysozyme is unable to bind sufficiently to the peptidoglycan substrate [30]. Bacteria with this modification are therefore resistant to treatment with lysozyme and other enzymes are needed (e.g. mutanolysin and lysostaphin [29]).

To ensure accurate representation in terms of both membership and abundance, protocols for DNA extraction should ideally be tailored to the bacterial communities

found in the samples under study. This however requires *a priori* knowledge of the bacterial communities in these samples – this is knowledge we usually do not have, and particularly not for the lower airways. It might therefore be tempting to use the whole arsenal of cell lysis tools available in order to secure accurate community representation. However, due consideration must also be made towards maintaining the integrity and yield of the extracted DNA, for which will be processed through a number of additional protocol steps post DNA extraction (Figure 1, steps B and C).

DNA integrity can be greatly impacted by choice of lysis method. When employing mechanical lysis methods there is an increased risk of DNA shearing and fragmentation. Because genomic DNA is released at an earlier stage of DNA extraction from the more easy-to-lyse gram-negative bacteria, these community members are likely more vulnerable to fragmentation than gram-positive bacteria. The main concern in amplicon-based microbiome studies, is that the fragmented DNA will increase the formation of recombinant amplification products (i.e. chimeras) in downstream steps of PCR (Figure 1, step B) [31]. As will be discussed further in a later section, chimeras represent a major source of error in microbiome analyses workflows because they may be interpreted as novel sequences (i.e. bacteria). Maintaining the integrity of the isolated DNA is therefore important for accurate representation of community composition, and choice of cell lysis method must carefully balance the goal of equal extraction efficiencies and DNA integrity against one another.

In addition, concentration of DNA obtained after DNA extraction (i.e. DNA yield) is variable across DNA extraction methods [29, 32]. Although it may seem reasonable that obtaining higher DNA yield, will result in better representation of community membership by also increasing signal from rare taxa, studies have shown that increased DNA yield does not necessarily equate to better community representation [29]. However, when processing low biomass samples the issue of contamination becomes relevant and increased DNA yield may be particularly important [32]. The discussion on low biomass samples will be elaborated on in section 6.4.

6.2.2 Library preparation for sequencing

After DNA extraction, the next step in the amplicon-based marker gene sequencing workflow is library preparation for sequencing (Figure 1, step B). As described earlier, this entails PCR amplification of the target marker gene to be sequenced and the addition of index sequences necessary for sample multiplexing.

Marker gene amplification

The bacterial 16S ribosomal RNA (16S rRNA) gene, is the most commonly targeted marker gene in amplicon-based microbiome studies. The approximately 1500 base pair long gene encodes a structural RNA component of the bacterial ribosome, and is a critical component of the cellular process of protein synthesis (i.e. translation of DNA to protein). The vital function of its gene product means that the 16S rRNA gene is found in all bacteria – and thus it serves as the perfect marker gene for capturing the full collection of bacteria in the sampled community [33, 34]. Although highly conserved, along its full length, the 16S rRNA gene consists of alternating variable and conserved sequences [33, 35, 36]. Conserved sequences are similar across all bacteria. The variable regions (for which there are nine (V1-V9)), on the other hand vary enough to allow bacterial identification to genus and sometimes even species level [37].

The popularity of the 16S rRNA gene as a target in microbiome studies comes in part from its gene structure enabling optimal use of current molecular tools. Current high-throughput sequencing technologies have a limitation on the maximum length of the DNA that can be sequenced; the power of high-throughput sequencing technologies lies in the large number of sequences that can be sequenced simultaneously, not the length of each of these sequences. At the time of writing this is approximately a third of the full-length 16S rRNA gene (approximately 600 bp). The structure of the 16S gene is convenient as “universal” PCR primers targeting the conserved regions within the gene allow for isolation and amplification of one or more of the shorter variable region(s) within the gene from (in theory) all bacteria in the sampled community. The length of these shorter amplicons is suitable for sequencing, and importantly have

been shown to be as informative as the full length sequences [38]. It is however important to note here that the optimal choice of target gene region has not been agreed upon, and some variation in results can occur based on choice of target variable region.

Indexing

Indexing is a method by which each amplicon is given a label or “address” sequence that links it back to the sample from which it was PCR amplified. The index sequence is attached to the amplicons during PCR, by their inclusion in PCR primer sequences. Indexing is necessary because amplicon libraries from all samples to be sequenced on the same sequencing run are mixed together in the final stages of library preparation prior to sequencing. A more detailed description is given below.

Figure 1 is a simplified representation of a typical library preparation workflow and shows steps as it occurs for one sample. In practice, this is usually performed for 96 samples at a time, using 96-well PCR plates. The final step of library preparation involves combining aliquots from all 96 samples to generate one sample that is further processed through the sequencing protocol (i.e. multiplexing). Post sequencing, sequences are assigned back to their samples via the unique sample specific “index” sequence(s) that were added to each amplicon during PCR steps of library preparation (i.e. demultiplexing) [39, 40].

As described in the Figure 1 text, index sequence(s) can be added during the same PCR step for which the target marker gene is amplified (i.e. 1-step PCR protocol) or during a separate PCR dedicated to the process of indexing (i.e. 2-step PCR protocol). Most common has been the use of a 1-step PCR protocol for which PCR primers include both the marker gene targeting sequence and the index sequence. Notably, the supplier of the most commonly used sequencing platform (Illumina), have chosen to base their commercial protocol for microbiome analyses on a 2-step PCR approach.

Index sequences can be added to only one end of the amplicons (i.e. the single indexing approach) or to both ends (i.e. the dual indexing approach). When the dual

index approach is used, the design can be either combinatorial or non-redundant. The combinatorial approach takes advantage of the fact that relatively few primers are needed for multiplexing many samples. For example in the 96 sample setup, a total of 8 forward and 12 reverse primers with unique indexes is sufficient for multiplexing 96 samples.

Bias and error during library preparation

Post-sequencing analysis of bacterial community composition is built on the assumption that the pool of 16S amplicons generated during library preparation is an accurate representation of the original sample, in terms of both bacterial membership and abundance. In Figure 1 step B, the relative proportion of amplicons from each of the three bacteria perfectly reflects the relative proportions of these bacteria in the sampled community. However, in practice the PCR is not perfect in this regard and several factors may contribute to the introduction of bias and error.

PCR primer bias is a problem that has attracted a lot of attention because of the inability to correct for this bioinformatically. Recall that “universal” PCR primers used in microbiome studies are designed to target conserved regions within the 16S rRNA gene – the designation “universal” implies that the primers are able to target all bacteria in the sampled community with equal efficiency. The issue of primer bias arises because the conserved regions are in fact not 100% similar (or conserved) across all bacteria [26, 36]. Preferential amplification, and hence a biased overrepresentation, of gene sequences for which the conserved region more closely matches the primer sequences is a general concern in microbiome studies.

In the design of “universal” primers for amplification of all bacterial members in the sampled community, sequence variability is somewhat accounted for by the use of degenerate primers. Degenerate primers consist of a mixture of primers that vary only at the specific base positions that are less conserved. However, the use of degenerate PCR primers does not completely alleviate the issue of primer bias. First of all, it is not certain that variability across all bacterial genomes is accounted for. Second, is the issue of “GC content”. In DNA, the hydrogen bonding between guanine (G) and

cytosine (C) (3 hydrogen bonds) is stronger than that between adenine (A) and thymine (T) (2 hydrogen bonds). Thus the higher affinity of PCR primers with G or C to its target template sequence, may result in the preferential amplification of genome targets with GC rich primer binding sites [41]. Thus, despite the use of degenerate universal primers, the issue of primer bias remains.

In addition to PCR primer bias introduced as a result of the marker gene targeting sequence, there is the question of the influence of additional overhang sequences such as Illumina adaptor sequences and index sequences. It is not understood whether the use of longer primer sequences associated with the 1-step PCR protocol may interfere with amplification of the target marker gene when compared to the 2-step PCR protocol, that separates marker gene amplification and indexing [42].

Besides PCR primer bias, inherent differences in 16S rRNA gene copy numbers across bacterial genomes may also lead to a biased representation of the sampled community. The ribosomal RNA operon (*rrn*), which holds the 16S rRNA gene, is often found in multiple and variable copy numbers across bacterial taxa – copy numbers typically range from 1 to 15 copies per genome [43]. Bacterial genomes with higher marker gene copy numbers may be overrepresented in the pool of amplicons generated after PCR amplification. In downstream analyses, this may result in a false impression that bacteria which are low in relative abundance in the sampled community, but contain high marker gene copy numbers, predominate. Further complicating matters is that sequence variation may be found between the 16S rRNA gene copies found even within the same bacterial genome – in the actinomycete *Thermobispora bispora* for example, 6.4% sequence variation has been found between two 16S rRNA copies [44]. When sequenced, this variation may be interpreted as originating from different bacteria, inflating measures of diversity within the community.

While, the issues of primer bias, GC content and copy number variation discussed above are examples of factors that may introduce bias (i.e. skew in relative abundances), other factors may introduce errors to the sequencing data (i.e.

misrepresentation of the sequence itself). Erroneous sequences are a concern because they are often difficult to identify and distinguish from true sequences. If not corrected for, erroneous sequences may even be interpreted as originating from other bacteria than those which are present in the sampled community – thus introducing false positives to the study. Two types of erroneous sequence are commonly described in the literature – that introduced by the polymerase during PCR and chimeras. The impact of polymerase incorporated errors vary according to DNA polymerase, but an estimated error rate of 1 substitution per 10^5 - 10^6 bases can be expected [26]. Bioinformatic approaches to dealing with such misincorporated bases are limited and pose a particular challenge when performing post-sequencing quality filtering steps, as will be discussed in section 6.3.2.

Chimeras, represent another type of erroneous sequence. Chimeras are mixed PCR products derived from two or more parent sequences found within the sampled community - resulting in so-called “bimeras” or “multimeras”, respectively. They may form between sequences originating from different bacterial genomes but also from copy variants found within the same genome [45]. The rate of chimera formation has been reported to be as high as over 30% in some studies [46, 47]. Even within well curated public repositories, it is assumed that approximately 5% of the sequences are chimeras [35]. Chimeras form as a result of mistakes during PCR and several mechanisms have been proposed. Recall that PCR amplification is performed in cycles, consisting of the following three steps: i) template denaturation, ii) primer annealing and iii) extension. If the extension step is terminated prematurely, incomplete PCR products may form that contain both the universal primer sequence and sequences specific to the 16S variant for which the primer annealed. In the next PCR cycle, these may behave as primers for amplification of other 16S variants, resulting in the formation of mixed PCR products (i.e. chimeras). Wang and Wang [46] were able to show that by using longer extension times, the frequency of chimera formation decreased, providing support for this mechanism. Another perhaps less frequent mechanism of chimera formation, is that which results from damaged DNA templates. As discussed earlier, harsh DNA extraction procedures (e.g. bead beating)

are required for efficient lysis of some bacteria, and this may result in DNA breakage and fragmentation. During the extension step of PCR, an encountered break in the DNA template may result in the “jumping” of the incompletely extended primer to another template, again resulting in a mixed PCR product or chimera [31]. Several bioinformatic tools for identification and removal of chimeras have been developed, and will be discussed further in section 6.3.4.

6.2.3 High-throughput sequencing

The final step in the amplicon-based marker gene sequencing workflow is high-throughput sequencing of the amplicon libraries (Figure 1, step C). While a number of different sequencing platforms exist (e.g. 454, PacBio, Illumina), each with own characteristic error patterns, herein the focus will be on the most commonly used platform – the Illumina MiSeq.

A description of the MiSeq sequencing process is first in order. All sequencing steps are performed on an Illumina flow cell, which may be described simply as a glass surface covered with two types of short DNA sequences (i.e. oligonucleotides). The oligonucleotides are attached to the flow cell surface and are complementary in sequence to the Illumina adapters found at the ends of the amplicon libraries. These adapters were added during library preparation steps (Figure 1, step B).

Before the actual sequencing can begin, a preparatory step referred to as “cluster generation” must be completed. During this process, the amplicon template DNA libraries (denatured to single strands in the final steps of library preparation), are first attached to the Illumina flow cell surface via complementary base pairing to the oligonucleotides that are attached to the flow cell. Each bound DNA fragment is then amplified via so-called bridge amplification PCR, to generate clusters consisting of approximately 1000 copies of the amplicon template DNA. After successful cluster generation, the flow cell consists of millions of distinct clusters, evenly spaced out across the flow cell surface, with each cluster representing an amplicon from the original pooled library.

Sequencing is then performed using an approach termed sequencing-by-synthesis (SBS). In short, sequences are “read” by building the strands complementary to the fragments that make up each cluster – one base at a time. The process is also carried out simultaneously for all clusters spread out across the flow cell - hence the term massive parallel sequencing used for the technology. Sequencing begins with the addition of sequencing primer, which marks the start position for the sequencing reads. During each cycle of sequencing, four fluorescently labeled nucleotides (A, T, C, G) with reversible terminator labels, are allowed to flow across the flow cell surface. The appropriate nucleotide binds to the strand being synthesized through complementary base pairing with the template DNA fragments that make up each cluster. The cumulative fluorescent signal generated from each of the fragments within a cluster is then recorded, and used to determine the base call for that cluster. The fluorescent reversible terminator labels are then removed, and the process repeats itself until the pre-programmed number of cycles have been completed.

The accuracy of the MiSeq sequencing process is determined in large by so-called phasing and pre-phasing events described hereafter [48]. Recall that for each cycle of sequencing, base calls are determined from the signal collected from all identical fragments that make up a particular cluster. The signal intensity is therefore dependent on the simultaneous incorporation and detection of the same nucleotide (i.e. base) across all fragments within a cluster. However, the chemistry is not perfect, and for each cycle it is expected that for a small fraction of the fragments, sequencing will either slow down (phasing) or progress ahead (pre-phasing) of the rest of the fragments. Phasing may for example result if the terminator label is not removed after a completed cycle. In turn, pre-phasing may result if a nucleotide lacks the terminator label, enabling the incorporation of more than one nucleotide in a cycle. For each sequencing cycle, the fraction of fragments in a cluster impacted by phasing and pre-phasing events increases - and the sequencing signal for that cluster becomes more distorted. This results in increasingly high error rates towards the ends of sequencing reads, and is currently the main reason for the limitation on maximum read lengths using this technology [48, 49]. The errors typically manifest themselves as

substitution type errors (i.e. an A is called instead of G), in contrast to insertion or deletion type errors frequently seen for other platforms.

Because of the issue of phasing and pre-phasing, the sequencing of the marker gene template and index(es) are typically performed in separate reads. The introduction of fresh primer for each new read will “restart” the sequencing process for all fragments within the cluster, mitigating the cumulated effects of phasing and pre-phasing [50]. And as described later, the paired-end (PE) sequencing approach, uses separation of reads to expand on the maximum read length that can be achieved using the currently available chemistry.

6.3 Bioinformatic sequence processing

In the previous section 6.2, the amplicon-based microbiome sequencing workflow has been outlined, with a focused discussion on potential sources of error and bias. In the following section, current bioinformatic approaches to dealing with some of these issues, as well as limitations that remain, will be discussed. Examples will be taken from one of the most popular bioinformatic pipelines – Quantitative Insight into Microbial Ecology (i.e. QIIME1/QIIME2). Note that QIIME is a wrapper for numerous other tools (e.g. DADA2) and for the current discussion, the default algorithms implemented in the pipeline will be referred to.

6.3.1 Demultiplexing

Recall that prior to sequencing, amplicon libraries from all samples to be sequenced on the same sequencing run are pooled together (i.e. multiplexed) [39, 40]. Once sequencing has been completed, the index sequence(s), which are the same for all amplicons from the same sample, are used to reassign sequences back to the sample from which they originated (i.e. demultiplexing). The sequencing output (Figure 1, step C), may be retrieved in the form of already demultiplexed fastq files. Other times, the fastq files have not been demultiplexed and bioinformatics processing begins with demultiplexing.

Bias may be introduced during demultiplexing if sequences are not assigned back to the correct sample - a phenomenon referred to as index misassignment. There are several ways by which index misassignment may occur [51]. Primers may for instance be contaminated during their manufacture. Cross-contamination may also occur during library preparation by way of internal well-to-well contamination between samples placed next to one another on the PCR plate, or during sequencing due to the presence of indexed amplicons from previous sequencing runs. Beside the issue of cross-contamination, there is the issue of PCR or sequencing error. Errors during PCR or sequencing may result in the conversion of an index sequence to that of another index used in the same sequencing setup. Most indexes are however designed to ensure that multiple substitution errors would have to occur before any one index would begin to resemble another index [49]. Index misassignments may also be associated with mixed or overlapping clusters on the flow cell, resulting in the assignment of an entire index read from one cluster to a sequence read from an adjacent cluster, or the assignment of a sequence read from one cluster to the index read(s) from another cluster [50, 51]. The use of unique dual indexed libraries, instead of single indexed libraries have been shown to reduce the impact of index misassignments [50, 51]. Quality filtering of index sequences has been proposed as a mechanism for correcting for index misassignments [51], although this has not been implemented in most pipelines.

6.3.2 Quality filtering

The aim of the quality filtering step is to filter out erroneous sequences resulting from PCR point errors and sequencing error (chimeras are dealt with in a subsequent step). While the occurrence of PCR and sequencing errors may appear as rare events to be overlooked, when dealing with the millions of sequences generated in amplicon sequencing data - these errors if not corrected for may result in inflated measures of diversity [26, 49, 52, 53]. In the following sections, differences between PCR and sequencing error and challenges associated with the removal of each will be discussed.

The Phred Q score

Most bioinformatic approaches to dealing with erroneous sequences have targeted errors generated during the sequencing process. These errors are relatively easy to identify compared to PCR incorporated errors, due to the sequencer generated Phred quality scores (Q score) that accompany sequence data (i.e. fastq files). By definition, the Q score is a measure of the probability by which a base is called incorrectly during the SBS process. It is calculated using the formula $Q = -10 \log_{10} P$, where Q is the Q score and P is the estimated error probability [54]. A Q30 score (Q score of 30) will for example translate to a base call accuracy of 99.9%. Quality scores can be used as an indicator to trim off low quality bases at read ends or for error correction when handling PE data – each strategy described subsequently. However, care must be taken when interpreting results, as there has been some disagreement on the reliability of the association between Q scores and error probabilities. While Kozich et al. [55] reported that sequencing errors were highly associated with low Q scores for example, Schirmer et al. [48] found the association to be unreliable.

Quality trimming

The aim of quality trimming is in essence to determine the point along the read for which the Q scores begin to fall below the desired threshold values (recall that reduction in quality scores is expected across the read length (section 6.2.3). The method described by Bokulich et al. [52], was implemented in the QIIME1 pipeline and performed in the same step as demultiplexing. The methods applied when using QIIME2 differ (implemented in denoising algorithms, e.g. DADA2), and will be discussed in a subsequent section. In simple words, when applying the “Bokulich method”, sequences are screened for low quality bases, starting at the beginning of the sequence and progressing until the end of the sequence is reached. Decisions regarding the number of consecutive bases (default=3) that must maintain a user defined threshold Q score (default=3), the length of the sequence after trimming (default = 75% of the untrimmed sequence), and finally the maximum number of ambiguous base calls (default = 0), determine whether a sequence is kept or discarded. Deciding on the appropriate threshold Q score is perhaps the most challenging. While demanding a high Q score (e.g. Q30) increases the overall quality

of the sequencing data, one also risks losing accurate reads that have been assigned lower but perhaps acceptable Q scores (e.g. Q20).

PE error correction

The PE sequencing approach adds another dimension to quality filtering [55, 56]. As described previously (section 6.2.3), sequencing error rates tend to increase towards the end of sequencing reads (and Q scores decrease). Using a PE sequencing approach (i.e. separation of reads), it is possible to extend the maximum read length that can be obtained using current sequencing chemistries. When performing PE sequencing, the targeted region is sequenced from both ends of the DNA sequence, in two separate reads. By design, the marker gene targeting primers should be chosen so that the amplicons are short enough to allow for an overlap between PE reads. This will enable the merging of the paired reads to form a contiguous read, and the opportunity to resolve discrepancies at the read ends. Different choices can be made with regards to resolving discrepancies between the overlapping reads. For the most stringent filtering procedures, one can demand that the overlapping reads match perfectly - if they do not, then the sequence may be removed. Alternatively, the overlap can be used to correct for an incorrect base, as more often than not, one of the reads will have a higher Q score than the other at the given base position. While PE sequencing has mainly been used to achieve longer sequencing reads, others have suggested that the approach should be used for generating completely overlapping reads for improved error filtering [49].

As mentioned above, because the Q score is a measure of the accuracy of the SBS process, it does not capture PCR errors incorporated during library preparation or cluster generation; an incorporated PCR error may indeed generate a perfect Q score. Current methods for dealing with PCR error (and also potentially missed sequencing errors) have largely been based on grouping sequences together into clusters called operational taxonomic units (OTUs).

6.3.3 Clustering and denoising

While much variation exists across different bioinformatic pipelines, a central step performed in all pipelines is clustering to OTUs or denoising to ASVs.

Clustering to operational taxonomic units (OTUs)

The most common approach has been to cluster sequences into operational taxonomic units (OTUs) based on a shared sequence similarity threshold of 97% (or dissimilarity of 3%). The method was built on the work by Stackebrandt and Goebel et al. [37] who were able to demonstrate that at 97% 16S sequence similarity one could distinguish between bacteria at species level. A representative sequence from each OTU is assigned a taxonomic label, and the number of sequences in each OTU is used to conclude on the relative abundance of the particular taxa in the sampled community. The list of OTUs, and the number of sequencing reads binned to each OTU form the working OTU table for which all subsequent bioinformatic analyses are performed.

Clustering of sequences into 97% OTUs serves two purposes. First of all, it reduces the number of sequences that are processed further down the bioinformatic pipeline. A typical MiSeq sequencing run will generate millions of sequencing reads. After clustering to OTUs this volume is often reduced to the thousands. Second, clustering serves the purpose of error correction. Erroneous sequences with less than 3% incorporated error, will be placed in the same OTU group as the correct sequences [57]. The method is however still vulnerable to spurious OTUs that may form from sequences with greater than 3% incorporated error. The issue of spurious OTUs was addressed by Bokulich et al. [52]. The authors recommended the removal of small OTUs, defined as those for which there were fewer sequences than 0.005% of the total sequence count.

Denoising to amplicon sequence variants (ASVs)

Recently, new methods of handling amplicon sequencing data have been developed with the promise of effectively removing both PCR and sequencing error [58, 59].

These methods have collectively been referred to as “denoising”, and the product sequences after denoising are referred to as an exact sequence variants (ESV) or amplicon sequence variants (ASVs). The most commonly used denoising method is implemented in DADA2 (divisive amplicon denoising algorithm 2) [59]. In simple words, the DADA2 denoising algorithm builds on the assumption that true biological sequences will be present at a higher frequency than erroneous sequences. Because ASVs are assumed to be free of PCR and sequencing error, there is no need to cluster sequences at 97% as was performed for OTUs. ASVs are thus referred to as the equivalent of 100% OTUs, making it possible to distinguish between sequences that vary by just one nucleotide. In addition to denoising to ASVs, the DADA2 method incorporates all general steps typical for amplicon sequencing data including, filtering, dereplication, chimera removal and joining of PE reads.

6.3.4 Chimera removal

Several bioinformatic tools have been developed for identification of chimeras. Early tools (e.g. Bellepheron [60], Pintail [35]) were designed for analysis of full-length 16S sequences, and were later shown to perform suboptimally on short amplicon data [47, 61]. This triggered the further development of new tools. Some examples include ChimeraSlayer [47], Perseus [61], and UCHIME [62] - to name a few. Chimera removal is also an integrated part of the DADA2 method as described above, developed specifically for targeting denoised ASVs.

The existing chimera detection tools listed above, are more or less based on a similar strategy. In short, the sequence in question is compared to a pre-defined set of non-chimeric reference sequences [61, 62]. If the query sequence matches to two or more sequences in the reference set, it is assumed that it formed from a recombination event between these matched sequences during PCR – the query sequence is then flagged as chimeric. Different detection tools may vary with regards to the origin of the non-chimeric reference set. Using *de novo* abundance based approaches (e.g. as seen in Perseus [61], UCHIME-*de novo* [62] and DADA2) sequences of high abundance in the data to be analysed are considered non-chimeric, and used to build the non-chimeric reference set. The logic being that parent sequences must have gone

through a greater number of PCR cycles than the chimeras that form between them (i.e. the parent sequence must already exist). Using an alternative approach, the reference set is provided by the user and based on a preexisting database of chimera-free sequences (e.g. as seen in ChimeraSlayer [47] and UCHIME – reference mode [62]).

The accuracy of the chimera filtering step will in large be dependent on quality of the non-chimeric reference set. Using the *de novo* approach, it is for example assumed that preferential amplification of a chimera sequence did not occur during PCR, as this would place the chimera in the reference set. The accuracy of chimera filtering methods will also depend on the presence of sequencing errors, as with the short read lengths characteristic of amplicon sequencing data, differences to be detected and used for inferring a chimera are small [62]. Using the database based approach, another consideration is whether the community under study is adequately represented in the reference set, important because chimeras formed from parent sequences not found in the reference set will not be detected. For most datasets, the communities under study (e.g. the airways) are poorly characterized and the chosen database likely to be incomplete; thus the *de novo* approach is often preferred.

6.4 Low biomass samples from the lungs

In the following section, the framework presented in Figure 1 is put in context to studies on the lower airway microbiome, which presents with additional challenges related to the low bacterial load of acquired samples.

6.4.1 Sample bacterial load and contamination

It was the reports by Biesbroek et al. [32] and Salter et al. [63] that initially stirred up discussion regarding the degree to which accurate microbiota analyses could be achieved for samples holding low bacterial loads (for which the lungs are the classic example in the literature). Both studies demonstrated the existence of an inverse relationship between sampled bacterial DNA levels and the influence of bacterial DNA contamination introduced during laboratory processing steps (e.g. from DNA

extraction kits, PCR reagents, the technician etc.). The significance of these studies warrants the review of key findings from each.

Biesbroek et al. [32] were concerned about the variable levels of DNA obtained from low biomass nasopharyngeal swabs. To explore the impact of bacterial load on the analyses of microbiota composition, they set up an experiment based on a serially diluted sample of saliva. Bacterial community composition began to differ from the original undiluted sample at a concentration of 10^5 bacteria/mL. The observed shift in bacterial composition coincided with an increase in the levels of bacteria mapping to the phylum *Proteobacteria*. Although the authors could not with certainty show that these bacteria originated from laboratory contamination, this was their interpretation of the data. Further, they defined 10^6 bacterial cells/mL as the threshold bacterial load for which accurate microbiota analyses could be performed. When extrapolating their findings to data obtained from the low biomass nares and nasopharynx, they found that choice of DNA extraction method (four different methods compared) determined whether samples fell above or below their set threshold level of bacterial load.

The paper by Salter et al. [63] was published two years later. The authors recognized the difficulty in distinguishing between contaminants and actual members of the sampled community, e.g. as seen for *Proteobacteria* in the Biesbroek study [32]. They designed a similar dilution experiment to that conducted in the former study, but with an important modification. Rather than using saliva (or other complex natural sample), they based their study on a monoculture consisting of one bacterial species not likely to be introduced from the laboratory environment or from reagents and kits used for sample processing (*Salmonella bongori*). On analysis of the sequencing output, sequences classified to taxa other than *S. bongori* were interpreted as derived from contamination. In accordance with the Biesbroek study [32], they observed a clear inverse relationship between sample bacterial load and proportion of non-*S. bongori* sequences (i.e. contaminants).

In summary, these groundbreaking studies have raised questions regarding the quality of sequencing data generated in studies on the bacterial lung microbiome – the most

important questions regarding the proportion of sequences that can be expected to be representative of the sampled lung microbiota and not contamination introduced from the laboratory. However, as discussed in the following section, the lack of standards with regards to protocols for sampling and laboratory processing make it difficult to generalize on the state of the field.

6.4.2 Sample bacterial load varies across protocols

In the design of their studies, investigators are faced with several decisions along the laboratory pipeline, that can determine the strength of the signal from the sampled bacterial community. A description of some critical steps follows.

Sampling.

As discussed earlier (section 6.1.2), BAL and SB are the most common sample types in airway microbiome studies. The manner by which these samples are collected has however not been standardized, and different approaches may lead to differences in obtained sample bacterial load. This is mainly due to differences in dilution effects. When sampling BAL for instance, decisions regarding the amount of fluid to instill and whether to keep fractions separate or pooled, will determine the levels of bacteria in the collected sample. However, the decision on whether to fractionate BAL or not, is not straightforward due to the potential for bronchoscope carryover from the upper airways. As discussed earlier, the fractioning of BAL may be used as a method for which one can “dilute away” upper airway bronchoscope carryover [6, 24] – however this is at the cost of lowering the obtained sample bacterial load. When sampling SB, the volume of sampling fluid for which SB are placed in after sampling will be equally important – be it saline sampling fluid or buffer for DNA extraction. Also the number of SB taken per sampling site has not been standardized – and the greater the number of SB per site, the higher we may expect the levels of bacteria.

Eukaryote cell removal

Once samples have been collected, a decision must then be made on whether to keep or remove eukaryotic host cells that have been collected together with the bacterial

cells [64, 65]. Eukaryotic cell removal is performed by centrifugation of the sample at a speed that pellets these larger cells out of solution, while leaving the smaller bacterial cells in suspension. Dickson et al. [64] compared results when processing whole (eukaryotic cells kept) and acellular (eukaryotic cells removed) BAL. They found that eukaryotic cell removal resulted in lower community richness - implying the concomitant removal of bacterial cells associated with eukaryote host cells (for instance via biofilms). Important to the current discussion - they also observed a lower bacterial load after the removal of eukaryotic cells. Both these findings warrant the use of whole samples over acellular samples. While most studies process whole samples, some key studies in the field have also used acellular samples [20, 25, 66, 67]. In the study by Lozupone et al. [67] both whole and acellular samples were used in the same study.

The publications by Biesbroek et al. and Salter et al. emphasize the importance of securing a high bacterial load already at the stage of sampling [32, 63]. This in order to overpower the contaminating bacterial signal introduced in subsequent steps - starting with bacterial DNA extraction.

Bacterial DNA extraction

A decision regarding the input sample volume used for DNA extraction is also one to be made. There is no standard in the field, and large differences are found across studies. For instance, the input volume BAL used in some studies is as low as only a few mL [6, 23], while in other studies volumes approximating 100 mL have been used [14, 64]. As described above however, several factors will determine the levels of bacteria in these samples (i.e. sample type, dilution effects and the use of whole or acellular BAL). Also, clinical factors such as disease state and the use of antibiotics may be important.

In addition to input sample volume, a decision has to be made regarding choice of DNA extraction method. Importantly, differences across methods can result in variable DNA yields [29, 32]. This has not been a major concern in studies on samples carrying high bacterial loads (e.g. the gut). DNA yield can however have a

major impact on studies on samples carrying low bacterial loads [32, 63]. Recall that Biesbroek et al. [32], show in their study how choice of DNA extraction kit determined whether samples fell above or below their defined threshold bacterial load for which accurate microbiota analyses could be achieved.

Polymerase chain reaction (PCR)

The PCR is a central tool in the amplicon-based marker gene sequencing workflow. In studies of low biomass samples, an increased number of PCR cycles is often used to obtain adequate levels of DNA for sequencing [32]. However, error and bias associated with the PCR, is expected to increase with increased number of PCR cycles. Furthermore, increasing the number of PCR cycles appears to result in weaker signals from the sampled microbiota and an increased signal from contamination. This was demonstrated in the study by Salter et al. [63] on comparison of sequencing output generated by processing the *S. bongori* dilution series through 20 and 40 PCR cycles.

In conclusion to the introduction sections 6.1 - 6.4, there is a need for studies that i) add to the discussion on how to sample the microbiota of the lower airways with minimal influence of oropharyngeal carryover, ii) increases knowledge on the impact of laboratory contamination in a low-biomass setting, and strategies to handle this, and finally, iii) shed light on the impact of methodological choices related to marker gene amplification and primer selection.

7. Objectives

The main objective of the thesis was to evaluate the impact that various methodological choices could have on the presentation of the airway microbiome. Specifically we investigated:

1. The bacterial composition observed when employing protected and unprotected bronchoscopic sampling methods of the lower airways (**paper I**).
2. The impact of laboratory contamination, and bioinformatic strategies for dealing with contamination in samples with low bacterial loads (**paper II**).
3. The impact of a one vs two step PCR protocol and choice of target amplicon region, 16S rRNA gene V3 V4 and V4 (**paper III**).

8. Materials and methods

8.1 Study design and participants

The PhD project was conducted as part of a single-center observational study: The Bergen COPD Microbiome study (short name «MicroCOPD»). The data collection for the MicroCOPD study was carried out at the outpatient clinic at the Department of Thoracic Medicine, Haukeland University Hospital (Bergen, Norway). The design of the MicroCOPD study has been published [23].

In brief, MicroCOPD is a bronchoscopy study designed to compare the airway microbiome of subjects with and without airway disease (i.e. COPD, asthma, healthy controls). The study was initiated in the spring of 2012, and two years of protocol development and a pilot phase followed. The first planned research bronchoscopy procedure for the main study was performed in April 2013. The last bronchoscopy was performed in June 2015.

All together, 323 bronchoscopies were performed on a total of 249 study participants (130 with COPD, 16 with asthma and 103 healthy controls). All participants were volunteers recruited to MicroCOPD from two previous studies conducted at the same department; the GeneCOPD follow-up study [68] and the Bergen COPD cohort study [69–71]. A small number of participants were also recruited by interest generated by the local press. In addition, asthma patients were recruited from a respiratory medicine private practice. There were several inclusion criteria. In the original protocol, there was an age limit of 40 years for COPD patients, and all COPD patients were above 40 years of age. As the study eventually also included asthma patients, this age limit was abandoned. This also applied to the requirement of a 10 pack year tobacco smoke exposure, enabling inclusion of also “never-smokers” with COPD for a total of three categories based on smoking habit (“current smokers”, “ex-smokers” and “never-smokers”). Airway obstruction was identified by post-bronchodilator spirometry, whereas all diagnoses were confirmed by experienced respiratory clinicians based on a comprehensive evaluation of patient history,

pulmonary function and radiologic imaging. Severity of airway obstruction was evaluated by the forced expiratory volume in 1 second (FEV1) in percent of predicted values based on Norwegian reference values [72]. The spirometry was performed on a Viasys Vmax ENCORE, and bronchodilation achieved by the administration of 0.4 mg salbutamol by a large-volume spacer at least 30 minutes before the procedure. Exclusion criteria included all factors indicating that the subject would not be able to tolerate research bronchoscopy. We postponed participation for subjects that had received antibiotics or corticosteroids the last 14 days.

Only a subset of the participants from the MicroCOPD study was included in the three papers for the current PhD project. Paper I included 125 participants (58 control subjects, 64 subjects with COPD and 3 subjects with asthma). Papers II and III, included the same 23 participants (9 control subjects, 10 subjects with COPD and 4 subjects with asthma).

8.2 Study samples

8.2.1 Procedural samples (Papers I, II and III)

Sample types included oral washes (OW), bronchoscopically acquired protected specimen brushes (PSB), protected bronchoalveolar lavages (PBAL), small-volume lavages (SVL), and negative control samples (NCS). A description of the sampling procedure used in the MicroCOPD study follows.

A sealed bottle (500 mL) of sterile phosphate-buffered saline (PBS) was opened prior to each procedure or used within 24 hours if multiple procedures were performed on the same day. The PBS had been sterilized by sterile filtration (0.22 μm) and autoclaving at 121 °C for 15 minutes by the manufacturer. 1 mL of the PBS was set aside for use as a negative control sample (NCS), never entering the study subject or being in contact with the bronchoscope.

Study participants were given 0.4 mg of salbutamol (a bronchodilator); this was required for pre-procedural lung function testing, but also a safety measure to prevent

the risk of bronchoscopy-induced bronchospasm. Before the bronchoscopy procedure, the participant was asked to gargle 10 mL of PBS for the collection of an oral wash sample (OW). Flexible video-bronchoscopy was then performed in supine position via the oral route. Each participant received local anaesthesia with lidocain; pre-procedurally as a spray to the tongue and oropharynx, and per-operatively through a catheter to the vocal cords, trachea, and bronchi. To minimize upper airway contamination to the scope working channel, no suctioning was performed before passage of the scope through the vocal cords. Having entered the lower airways, three protected specimen brushes (rPSB) were first collected from the right lower lobe (RLL) using double sheathed wax-plug protected specimen brushes (Conmed, USA) and placed in 1 mL PBS. Next, two fractions of protected bronchoalveolar lavage (PBAL1/PBAL2) were collected from the right middle lobe (RML); this by instilling 2 x 50 mL PBS through a wax-plugged catheter (Plastimed Combicath, France). The bronchoscope was then repositioned for sampling of the left lung. This was done without retracting the bronchoscope above the vocal cords. Three wax-plugged protected specimen brushes (IPSB) were collected from the left upper lobe (LUL) and placed in 1 mL PBS. Finally, a sample of small-volume lavage (SVL) was collected from the same segment as the IPSB by instilling 20 mL PBS directly through the working channel. For 100 participants, the left side was examined before the right side.

8.2.2 Procedural control samples (Paper II)

For the collection of procedural control samples, we returned to the bronchoscopy room and performed ten simulated (no patient) procedures over two days. For each procedure five samples were collected: a bronchoscope rinse (BR), a catheter rinse (CR), a protected specimen brush (PSB), a sample of phosphate buffered saline (PBS) transferred to a cryotube (CT) and a negative control PBS sample (NCS).

On each day, a sealed 500 mL bottle of PBS was opened for use in all five procedures. 1 mL PBS was transferred to both an eppendorf tube for collection of the NCS and to a cryotube for collection of the CT sample. Sampling PSB was performed by passing a wax-plug protected specimen brush (Conmed, USA) through

the bronchoscope working channel. The brush was exposed to the air and then returned back through the bronchoscope. Using sterile scissors, the brush end was cut off and transferred to an eppendorf tube with 1 mL PBS. This was repeated until three brushes had been collected per sampling. Sampling CR was performed by passing a wax-plugged catheter (Plastimed Combicath, France) through the bronchoscope working channel and instilling 50 mL PBS. This was collected in a 50 mL Falcon tube on the other end and then suctioned back up. Sampling BR was performed by instilling 20 mL PBS and collected in a serial connected lavage trap. Samples were aliquoted and stored at - 80 °C.

8.2.3 *Salmonella* dilution series (Paper II)

Salmonella enterica serovar Typhimurium (ATCC 14028, USA) was plated out on blood agar plates and incubated overnight at 37 °C. The following day, colonies were transferred to a tube containing sterile physiological water using sterile cotton swabs until a McFarland density of approximately four was reached. The suspension was used to prepare a ten-fold dilution series across a total of seven samples. Aliquots were stored at -20 °C.

8.2.4 Mock community (Paper III)

A mock community (MC) sample consisting of genomic DNA from 20 different bacterial species (17 genera) was included on all sequencing runs. The MC consisted of uneven levels of genomic DNA from the different species of bacteria, with the number of rRNA operons per species varying from 1000 to 1000000 counts (Table 1). The operon count (provided on the certificate of analysis) was used to calculate the relative abundance of the different bacteria in the sample.

The reagent was obtained through BEI Resources, NIAID, NIH, as part of the Human Microbiome Project: Genomic DNA from Microbial Mock Community B (Staggered, Low Concentration), v5.2L, for 16S rRNA Gene Sequencing, HM-783D.

Table 1. Mock community HM-783D

Species	Number of operons	Relative abundance
<i>Acinetobacter baumannii</i>	10000	0.22 %
<i>Actinomyces odontolyticus</i>	1000	0.02 %
<i>Bacillus cereus</i>	100000	2.19 %
<i>Bacteroides vulgatus</i>	1000	0.02 %
<i>Clostridium beijerinckii</i>	100000	2.19 %
<i>Deinococcus radiodurans</i>	1000	0.02 %
<i>Enterococcus faecalis</i>	1000	0.02 %
<i>Escherichia coli</i>	1000000	21.91 %
<i>Helicobacter pylori</i>	10000	0.22 %
<i>Lactobacillus gasseri</i>	10000	0.22 %
<i>Listeria monocytogenes</i>	10000	0.22 %
<i>Neisseria meningitidis</i>	10000	0.22 %
<i>Propionibacterium acnes</i>	10000	0.22 %
<i>Pseudomonas aeruginosa</i>	100000	2.19 %
<i>Rhodobacter sphaeroides</i>	1000000	21.91 %
<i>Staphylococcus aureus</i>	100000	2.19 %
<i>Staphylococcus epidermidis</i>	1000000	21.91 %
<i>Streptococcus agalactiae</i>	100000	2.19 %
<i>Streptococcus mutans</i>	1000000	21.91 %
<i>Streptococcus pneumoniae</i>	1000	0.02 %

8.3 Bacterial DNA extraction (Papers I, II, III)

The protocol for bacterial DNA extraction used in the MicroCOPD study was designed in-house by Tuyen Thi Van Hoang (UiB) and Professor Harald G. Wiker (UiB). The protocol includes both enzymatic and mechanical lysis methods, as recommended in the current literature [29]. A description follows.

The volume of sample used as input to the DNA extraction protocol varied with sample type. For procedural samples: 1800 μ l for OW and PBAL and 450 μ l for PSB and NCS. For procedural control samples: 1800 μ l for BR and CR, 550 μ l for PSB and 450 μ l for CT and PBS. For the SDS, an input volume of 500 μ l was used.

The samples were first treated with Sputasol (Oxoid Limited, England) (i.e. dithiothreitol) and incubated at 37 °C for 15 minutes on a thermomixer (1000 rpm); this to ensure a homogenous distribution of the bacterial cells in the sample. The volume of Sputasol added to each sample, was the same as the input sample volume for DNA extraction. Bacterial cells (and eukaryotic cells) were collected by performing a high speed centrifugation step, at 15700 g for 8 minutes. The resulting cell pellet was resuspended in 250 μ l PBS. Next, enzymatic lysis was performed by treatment with an enzyme cocktail solution consisting of 25 μ l lysozyme (10 mg/ml, Sigma-Aldrich, USA), 3 μ l mutanolysin (25 KU/mL, Sigma-Aldrich), 1.5 μ l lysostaphin (4000 U/mL, Sigma-Aldrich) and 20.5 μ l TE5 buffer (10 mM Tris-HCl, 5 mM EDTA, pH 8) and incubated at 37 °C for 1 hour on a thermomixer (350 rpm). Bacterial cells not sufficiently lysed by the enzymes, were collected by centrifugation at 15700 g for 15 minutes. The supernatant containing any extracted DNA was transferred to a new eppendorf tube and stored at 4 °C, while further processing of the hard to lyse pellet; this to prevent DNA shearing in the subsequent mechanical lysis step. Mechanical lysis was then performed on the pelleted cells by processing through the FastDNA Spin Kit (MP Biomedicals, USA). The pellet was first resuspended in 800 μ l CLS-TC lysis buffer and then transferred to a Lysing Matrix A tube (FastDNA Spin Kit). Mechanical lysis was performed using the FastPrep-24 instrument (MP Biomedicals) at a speed setting of 6.0 m/s for 40 seconds. The lysate was then

combined with the supernatant from the enzyme lysis step. Further processing was performed as described by the manufacturers for the FastDNA Spin Kit. DNA was eluted in a total volume of 100 μ l.

8.4 Quantification of bacterial load (Paper II)

Sample bacterial load was measured by probe-based quantitative PCR (qPCR) targeting the bacterial 16S rRNA gene region V1V2.

Primers with sequences [5'-AGAGTTTGATCCTGGCTCAG-3'] (forward) and [5'-CTGCTGCCTYCCGTA-3'] (reverse) were used together with probe [5'-6-FAM-TAACACATGCAAGTCTGA-BHQ-1-3'] (locked nucleic acid bases are underlined; 6-FAM: 6-carboxyfluorescein; BHQ-1: Black Hole Quencher-1) [6, 7, 14, 21]. Each reaction consisted of 10 μ l Takyon No ROX Probe MasterMix (2X) (Eurogentec, Belgium), 0.2 μ l of each forward and reverse primer (10 μ M), 0.15 μ l of the hydrolysis probe (10 μ M), 2 μ l sample and RT-PCR grade water (Thermo Fisher Scientific, USA) for a total volume of 20 μ l.

Cycling was performed on a Light Cycler 480 instrument (Roche) using the following conditions: an initial cycle at 95 $^{\circ}$ C for 5 minutes followed by 45 cycles of 95 $^{\circ}$ C for 5 seconds, 60 $^{\circ}$ C for 20 seconds and 72 $^{\circ}$ C for 10 seconds and a final extension cycle of 72 $^{\circ}$ C for 2 minutes. A standard curve was constructed from a 10-fold dilution series of genomic DNA from *E.coli* strain JM109 (Zymo Research, USA).

8.5 Illumina MiSeq sequencing (Papers I, II, III)

Analysis of bacterial community composition was performed by high-throughput amplicon-based sequencing of the bacterial 16S rRNA gene on the Illumina MiSeq sequencing platform. Three different setups for MiSeq sequencing were used in the project. The setups varied with regards to the number of PCR steps (one or two) and the target marker gene region sequenced (16S rRNA gene region V3 V4 or V4): Setup 1 (2-step PCR; region V3 V4); Setup 2 (2-step PCR; region V4); Setup 3 (1-step PCR; region V4).

8.5.1 MiSeq sequencing setup 1 (Papers I, II, III)

Sequencing setup 1 is based on the Illumina 16S Metagenomic Sequencing Library Preparation guide (Part # 15044223 Rev. B). The protocol consists of two PCR steps; the first for amplification of the target marker gene region to be sequenced, and the second for the addition of index sequences required for sample multiplexing.

In the first PCR, the 16S rRNA gene region V3 V4 was targeted using primers [5'-TCGCGGCAGCGTCAGATGTGTATAAAGAGACAGCCTACGGGNGGCWGCAG-3'] (forward) and [5'-GTCTCGTGGGCTCGGAGATGTGTATAAAGAGACAGGAC TACHVGGGTATCTAATCC-3'] (reverse). Illumina overhang adapter sequences are underlined. Gene specific sequence (not underlined) were taken from Klindworth *et al.* [73], and included four degenerate bases named according to IUPAC nucleotide nomenclature (N = any base; W = A or T; H = A, C or T; V = A, C, or G).

Each reaction consisted of 5 µl sample, 12.5 µl KAPA HiFi HotStart ReadyMix (2X) (Kapa Biosystems, South Africa), 0.5 µl of each primer (10 µM) and 6.5 µl RT-PCR grade water (Thermo Fisher Scientific, USA) for a total volume of 25 µl. PCR cycling was performed using the following program: an initial cycle at 95 °C for 3 minutes, followed by 45 cycles of 95 °C for 30 seconds, 55 °C for 30 seconds, 72 °C for 30 seconds, and a final extension cycle at 72 °C for 5 minutes. PCR cleanup was performed using Agencourt AMPure XP beads (Beckman Coulter, USA).

The second PCR was performed using primers from the Nextera XT Index kit (Illumina Inc., USA). The primers targeted the Illumina overhang adapter sequences added to each amplicon in the first PCR (underlined in the primer sequences given above). PCR cycling was performed using the following program: an initial cycle at 95 °C for 3 minutes, followed by 8 cycles of 95 °C for 30 seconds, 55 °C for 30 seconds, 72 °C for 30 seconds, and a final extension cycle at 72 °C for 5 minutes. PCR cleanup was performed using Agencourt AMPure XP beads (Beckman Coulter, USA).

Amplicon libraries were quantified using the Qubit dsDNA HS Assay Kit (Life Technologies, USA), normalized to 4 nM, and pooled. The pooled library was

denatured with NaOH, and diluted to 10 pM using a buffer provided in the MiSeq reagent kit v3 (Illumina). Finally, the diluted, denatured library pool was spiked (15%) with PhiX from the PhiX Control Kit v3 (Illumina). MiSeq sequencing was performed using 2x300 cycles of paired-end sequencing using reagents from the MiSeq reagent kit v3 (Illumina).

8.5.2 MiSeq sequencing setup 2 (Paper III)

Sequencing setup 2 was based on the two-step PCR protocol described in the Illumina 16S Metagenomic Sequencing Library Preparation guide (Part # 15044223 Rev. B).

In the first PCR, the 16S rRNA gene region V4 was targeted using primers [5'-TCGTC GGCAGCGTCAGATGTGTATAAAGAGACAGGTGCCAGCMGCCGCGGTAA-3'] (forward) and [5'-GTCTCGTGGGCTCGGAGATGTGTATAAAGAGACAGGGACT ACHVGGGTWTCTAAT-3'] (reverse). Illumina overhang adapter sequences are underlined. Gene specific sequence (not underlined) were taken from Caporaso *et al.* [74] and included four degenerate bases named according to IUPAC nucleotide nomenclature (M = A or C; H = A, C or T; V = A, C, or G; W = A or T).

Each reaction consisted of 5 µl sample, 12.5 µl KAPA HiFi HotStart ReadyMix (2X), 1.25 µl of each primer (10 µM), and 5 µl RT-PCR grade water (Thermo Fisher Scientific, USA) for a total volume of 25 µl. PCR cycling was performed using the following program: an initial cycle at 95 °C for 3 minutes, followed by 45 cycles of 95 °C for 30 seconds, 50 °C for 30 seconds, 72 °C for 30 seconds, and a final extension cycle at 72 °C for 5 minutes. PCR cleanup was performed using Agencourt AMPure XP beads (Beckman Coulter, USA).

The second Index PCR step, library quantification, normalization and preparation for sequencing was performed as described for sequencing setup 1. MiSeq sequencing was performed using 2x275 cycles of paired-end sequencing using reagents from the MiSeq reagent kit v3 (Illumina).

8.5.3 MiSeq sequencing setup 3 (Paper III)

Sequencing setup 3 was based on the one-step PCR protocol described in Kozich *et al.* [49]. Both steps of target gene amplification and indexing are incorporated into a single PCR step using primers that include the gene specific sequences, index sequences and the Illumina sequencing adapters.

The 16S rRNA gene region V4 was targeted using primers [5'-AATGATACGGCGA
CCACCGAGATCTACACNNNNNNNTATGGTAATTGTGTGCCAGCMGCCG
CGGTAA-3'] (forward) and [5'-CAAGCAGAAGACGGCCATACGAGATNNNNNN
NNAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT-3'] (reverse). Illumina sequencing adapters, indexes, pad and linker regions (detailed in Kozich *et al.* [49]) are underlined. The gene specific sequences (not underlined) are the same as for the primers used in the sequencing setup 2 (section 8.5.2), from Caporaso *et al.* [74].

Each reaction consisted of 5 µl sample, 18 µl Accuprime Pfx SuperMix (Thermo Fisher Scientific, USA), 1 µl of each primer (10 µM), for a total volume of 25 µl. PCR cycling was performed using the following program: an initial cycle at 95 °C for 2 minutes, followed by 45 cycles of 95 °C for 20 seconds, 55 °C for 15 seconds, 72 °C for 5 minutes, and a final extension cycle at 72 °C for 5 minutes. PCR cleanup was performed using Agencourt AMPure XP beads (Beckman Coulter, USA).

Sequencing was performed using sequencing primers [5'- TATGGTAATTGTGTGC CAGCMGCCGCGGTAA-3'] (read 1 primer), [5'-AGTCAGTCAGCCGGACTACH VGGGTWTCTAAT-3'] (read 2 primer) and [5'-TTAGAWACCCBDGTAGTCCG GCTGACTGACT-3'] (index read primer).

Library quantification, normalization and preparation for sequencing was performed as described for sequencing setup 1. MiSeq sequencing was performed using 2x250 cycles of paired-end sequencing using reagents from the MiSeq reagent kit v3 (Illumina).

*Degenerate bases are named according to the IUPAC nucleotide nomenclature (N = any base; W = A or T; H = A, C or T; V = A, C or G; M = A or C; B = C, G or T).

8.6 Bioinformatics sequence processing

Bioinformatics sequence processing was performed using the Quantitative Insights Into Microbial Ecology (QIIME) bioinformatic package (<http://qiime.org>) [75]. QIIME underwent an upgrade from QIIME1 to QIIME2 in January 1, 2018 and both versions were used in the project.

8.6.1 QIIME1 (Papers I and II)

For papers I and II, bioinformatic sequence processing steps were performed using the QIIME1 package, version 1.9.1. Sequences were first retrieved from the Illumina MiSeq in the form of two fastq files per sample – one for the forward read and one for the reverse read (i.e. demultiplexed, paired-end reads). Sequence processing began with the removal of PCR primer sequences. This was performed by instruction to trim off the first 17 and last 21 bases, which corresponds to the length of the forward and reverse primer sequences. The forward and reverse reads were then joined together to form contiguous sequences, using the default “fastq-join” method. We required a minimum of 100 bases of overlap between the forward and reverse reads. Quality filtering was performed by removal of sequences with a base quality (Phred) score of less than 20. In paper II, chimeras were removed after identification using the VSEARCH [76]. The working operational taxonomic unit (OTU) table was generated by clustering of sequences into 97% OTUs, using UCLUST [77] and the GreenGenes reference database (v.13.8) [78]. Small OTUs consisting of fewer than 0.005% of the total sequence count in the dataset were discarded. Taxonomy assignment was performed using the naïve bayesian RDP Classifier [79] together with the GreenGenes reference database (v13.8). Sequences were aligned using PyNAST [80] and a phylogenetic tree was constructed using FastTree [81].

8.6.2 QIIME2 (Paper III)

Demultiplexed paired-end sequencing reads (fastq files) were retrieved from the MiSeq sequencer and imported into the QIIME2 environment (release 2019.1). The DADA2 denoising method was applied using the dada2 denoise-paired plugin, resulting in i) the removal of primer sequences and low quality bases at read ends, ii)

the joining of paired end reads, ii) the removal of chimeras and iv) denoising to amplicon sequence variants (ASVs). An additional round of chimera filtration was performed using the `vsearch uchime-denovo` plugin. Abundance filtering was then performed by removal of ASVs with fewer sequences than 0.005% of the total number of sequences and removal of ASVs not found in at least two samples. Taxonomy was assigned using the feature-classifier `classify-sklearn` plugin together with a Naïve Bayes classifier (pre-trained on the full-length Greengenes 13_8 99% OTU reference database) available on `qiime2.org`. ASVs that classified above the phylum level were removed, in addition to ASVs that classified to mitochondria, chloroplasts or archaea.

8.7 Decontamination strategies (Papers I, II, III)

Two approaches to dealing with laboratory contamination were used in the thesis. In papers I and II, the “remove all” approach was used. In papers II and III, tools available within the the `Decontam` R package were used [82].

8.7.1 The remove all approach

When using the “remove all” approach, all OTUs found in NCS were discarded from the corresponding procedural samples collected under the same bronchoscopy procedure. In brief, the main working OTU table was first split by subject ID. OTUs found in NCS were removed along with samples with zero sequence counts. The resulting OTU tables (now free of NCS OTUs and NCS samples) were merged back together, generating the final decontaminated OTU table.

8.7.2 Decontam

Contaminant OTUs/ASVs were identified in procedural samples using the `isContaminant` function available in the `Decontam` [83] R package. The two algorithms that `Decontam` uses to identify potential contaminants are based on either the prevalence of ASVs/OTUs in NCS versus actual samples (prevalence based approach) or the co-variance of OTUs/ASVs with the total amount of DNA in samples, measured before standardizing the amount of DNA and loading samples into

the MiSeq. The possession of both negative controls and DNA quantitation data (qubit measurements) enabled us to take use of the approach “either”, within the `isContaminant` function. We set the user defined threshold value to 0.5 (default=0.1) for both algorithms. For the prevalence based approach, a setting of 0.5 implies that the OTU/ASV was as common in NCS samples as in procedural samples, whereas for the frequency-based approach, a threshold of 0.5 implies that the underlying statistical models for the sequence being a contaminant is as likely as the model for the sequence not being a contaminant. Our choice of the 0.5 setting was based on the intuitively available interpretation of this value, in addition to facilitate comparison with the “remove all” strategy.

The output from the `isContaminant` analysis (method=“either”, threshold=0.5) was a list of all OTUs/ASVs identified as contaminants by either the “prevalence” or “frequency” based contaminant classification algorithms.

8.8 Statistical analyses

The samples in a microbiome study all have different yield in terms of number of sequences. To take this into account some analyses are performed on proportions (e.g. most of the taxonomic analyses show the relative abundance at various taxonomic levels). For other analyses a number of sequences is required as input. Many of these analyses require an equal number of sequences in each sample. The analyst can choose to normalize to an equal number of sequences, or rarefy. The latter implies drawing (randomly) a set number of sequences from each sample, while discarding those not randomized [84].

We chose to rarefy our samples for alpha- and beta-diversity analyses, trying to balance between the exclusion of samples with few remaining sequences, and at the same time not discarding the signal from low-biomass samples of the airways.

The study population was divided into subjects with- and without airway diseases. It is well known that the clinical characteristics of these differ, and it was not an objective of the three papers to investigate these. Thus, no formal statistical testing

was performed on the characteristics of the participants, but was presented in the first table of all three papers.

Differences in taxonomic composition of samples were displayed in terms of plots of average composition of samples in all three papers. Since these were compositional data (i.e. relative abundances), we in paper I applied a beta distribution and used the *betafit* command in *Stata* to test for differences in the taxonomic composition between sample types.

In paper II we compared the estimated bacterial burden in the different sample types. Statistical significance testing was performed on logarithmically transformed outcomes by nonparametric trend tests in *Stata*.

For alpha-diversity (rarefied samples) we in paper I compared the Faith phylogenetic index by sample type using Wilcoxon matched pairs signed rank test as well as a nonparametric trend test (both performed in *Stata*). In the other two papers we did not perform formal tests of alpha diversity measures.

For all three papers we chose unweighted UniFrac distance as our main outcome in beta-diversity analyses. In paper I, we compared beta diversity pairwise between different sample types, visualized by principle coordinates analyses. The significance of these differences was tested by applying a permanova test in QIIME1. Whereas we did not perform extensive beta diversity analyses in paper II, in paper III we used principle coordinates analyses to investigate the overlap between the NCS and procedural samples. In addition, we also visualized the overlap between procedural samples in experimental setups 2 and 3 (before and after Decontam).

Beside using R, QIIME1 and QIIME2 we have used the statistical software Stata SE for Mac OS (versions 13 to 16, Stata Corp, College Station, TX, USA).

8.9 Ethics

MicroCOPD was conducted in accordance with the ethical principles outlined in the Declaration of Helsinki. The Regional Committee for Medical and Health Research

Ethics approved the study (REK Nord, project number 2011/1307). All study participants provided written informed consent.

9. Summary of papers

9.1 Paper I

Protected sampling is preferable in bronchoscopic studies of the airway microbiome

For the first paper, we set out to determine whether protected (via a sterile catheter) bronchoscopic sampling techniques would reduce the influence of bronchoscopic carryover from the upper airways. We compare three sampling techniques: i. protected specimen brushings (PSB), ii. protected bronchoalveolar lavage (PBAL) and iii. (unprotected) small volume lavage (SVL).

Samples were collected from a total of 125 participants, consisting of 67 subjects with obstructive lung disease (3 asthma and 64 COPD patients) and 58 healthy controls. Our bronchoscopy sampling scheme was designed to mitigate the impact of confounding factors, including microaspiration and intrapulmonary contamination. This required sampling at multiple intrapulmonary sites within each participant, and in a strictly specified order: three PSB from the right lower lobe (rPSB), two fractions of PBAL from the right middle lobe (PBAL1/PBAL2), three PSB from the left upper lobe (lPSB) and SVL from the left upper lobe. For a subset of participants (n=49), the left lung was sampled before the right lung. An oral wash (OW) sample was collected prior to each bronchoscopy procedure, for use as an upper airway reference sample, along with a negative control sample (NCS) consisting of PBS used for collection of all samples.

Analyses of bacterial community composition was performed by MiSeq sequencing of the bacterial 16S rRNA gene V3 V4 region. Bioinformatics processing was performed using QIIME1. Our approach to contamination was to remove all sequences observed in NCS.

Our analyses were based on the underlying assumption that the more similar the bronchoscopically acquired specimens (PSB, PBAL, SVL) were the OW sample, the greater the influence of upper airway carryover on these samples. Our assessment

was based on three parameters: i. taxonomy, ii. alpha diversity and iii. beta diversity. Across all three parameters, similarity to the OW sample decreased in the order: OW>SVL>PBAL>PSB. Our analysis of taxonomic composition revealed an increase in the proportion of *Proteobacteria* and a simultaneous decrease in the proportion of *Firmicutes*. Measures of alpha diversity (Faith's PD) decreased across sample types. By analysis of PCoA plots of unweighted UniFrac distances (beta diversity) we found that the overlap between OW and SVL samples was greater than that for OW and PBAL samples and OW and PSB samples.

9.2 Paper II

Laboratory contamination in airway microbiome studies

For paper II, we sought to determine the impact of laboratory contamination on airway microbiome analyses, and to explore the expected inverse relationship between sample bacterial load and influence of contamination when processing samples through the MicroCOPD laboratory pipeline. Furthermore, we set out to determine the optimal bioinformatic approach to dealing with contamination post sequencing. Analyses included quantitative PCR and targeted amplicon sequencing of the bacterial 16S rRNA gene.

Samples were collected from 23 participants from the MicroCOPD study, consisting of 14 subjects with obstructive lung disease (COPD, asthma) and 9 healthy controls. Sample types collected from each participant included oral washes (OW), two fractions of protected bronchoalveolar lavage (PBAL1/PBAL2), protected specimen brushes (PSB) and an aliquot of the phosphate buffered saline (PBS) used for collection of all samples (NCS). Additional procedural control samples were collected after ten simulated bronchoscopy procedures (no patient). Samples included a bronchoscope rinse (BR), a catheter rinse (CR), a protected specimen brush (PSB), a sample of PBS transferred to a cryotube (CT) and a sample of PBS used for collection of all samples. Molecular grade water samples were also processed through the DNA extraction protocol without the introduction of PBS. For assessment of the

relationship between sample bacterial load and the influence of contamination, we included a ten-fold dilution series of a sample of *Salmonella* (SDS).

Analysis of the SDS revealed an inverse relationship between sample bacterial load and the influence of contamination. When extrapolating these findings to quantitative data obtained for airway samples (PBAL, PSB), we found that an estimated 10-50% of the bacterial community profiles could be traced back to contaminating bacterial DNA introduced from the laboratory. The OW sample appeared unaffected by contamination. On examination of procedural control samples (BR, CR, PSB, CT, PBS), molecular grade water samples processed through DNA extraction and PCR water samples introduced post DNA extraction, the DNA extraction kit was identified as a main contamination source.

We compared three different bioinformatic approaches for removal of contamination: i) keep all samples intact (i.e. do nothing), ii) remove all OTUs seen in NCS, and iii) correction based on statistical models (i.e. the Decontam R package). Contaminant removal based on Decontam appeared to provide a balance between keeping and removing OTUs found in both NCS and study samples.

9.3 Paper III

Exploring protocol bias in airway microbiome studies: one versus two PCR steps and 16S rRNA gene regions V3 V4 versus V4

The lung microbiome has been studied using a wide range of protocols for high-throughput sequencing of the bacterial 16S rRNA gene. We set out to determine the impact of number of polymerase chain reaction (PCR) steps (1- or 2-steps) and choice of target marker gene region (V3 V4 or V4) on the presentation of the upper and lower airway microbiome.

The study included samples from 23 participants in the MicroCOPD study, consisting of subjects with (n=14) and without (n= 9) obstructive lung disease (COPD, asthma). Samples collected included oral washes (OW), protected specimen brushes (PSB)

from the right lower lobe and protected bronchoalveolar lavages (PBAL) from the right middle lobe. An aliquot of phosphate buffered saline used for the collection of all procedural samples was used as a negative control sample (NCS) for assessment of contamination. A PCR water sample, introduced post DNA extraction was also included for distinguishing between contamination introduced before and after DNA extraction. A mock community (MC) sample consisting of genomic DNA from 20 different bacterial species was included as a positive control.

Samples were processed through three different library preparation setups varying in number of PCR steps and targeted marker gene region: Setup 1 (2-step PCR; V3 V4 region), Setup 2 (2-step PCR; V4 region), and Setup 3 (1-step PCR; V4 region). Sequencing was performed on the Illumina MiSeq.

The number of sequences and amplicon sequence variants (ASVs) decreased in order setup 1 > setup 2 > setup 3. The observation appeared to be associated with an increased taxonomic resolution when sequencing the V3 V4 region (setup 1) and an increased number of small ASVs in setups 1 and 2. The latter was considered the result of increased susceptibility to contamination when following the 2-step PCR protocol (setup 1 and 2) and when sequencing across multiple sequencing runs (setup 1).

Comparison of contamination profiles (based on NCS) revealed the dominance of ASVs assigned to family *Enterobacteriaceae* across all three setups. For setup 3, an additional ASV attributed to genus *Escherichia coli* (family *Enterobacteriaceae*) dominated. The same ASV also dominated the PCR water sample in setup 3, and we interpreted this as a contaminant introduced with PCR reagents for this setup. The elevated levels of *Escherichia coli* in the MC sample processed in setup 3 further supported this interpretation.

Comparison of procedural samples revealed that the same taxa dominated across all setups, although at different relative abundances. Analyses of beta diversity revealed that OW samples clustered together regardless of number of PCR steps. PSB and PBAL samples separated. The removal of Decontam contaminants did not resolve the

differences between setups. This indicated that mechanisms related to sample bacterial load, other than contamination was driving the observed protocol bias.

10. Discussion

10.1 Discussion on methods

10.1.1 Study populations

Paper I

Bronchoscopic studies on the lower airway microbiome have in general been limited by small population sizes. The bronchoscopy procedure is invasive and costly, and the implementation of larger studies has therefore shown to be challenging. For some studies, the need for more power has been addressed by combining data from different centers [18, 19, 67]. However, an important weakness of large multicenter initiatives, is that differences in methods for sampling, processing and sequencing across centers may introduce experimental bias to the data that is difficult to adjust for.

With 249 participants enrolled in the MicroCOPD study, we were for the first paper in a position to conduct a high power single-center investigation that would meet the increasing demand for such studies in the field. As sampling and sequencing for MicroCOPD was still ongoing as we started our analyses, we could not include all participants from the main study. Our analyses were therefore limited to a subset consisting of 125 participants, encompassing a relatively even proportion of subjects with obstructive lung disease ($n=67$) and control subjects ($n=58$). Furthermore, the participants with obstructive lung disease were represented by users and non users of inhaled corticosteroids (ICS). The study population also included current-, ex- and never-smokers.

We found no widely accepted method to estimate sample size, but compared to 23 (out of 25) previous bronchoscopy studies (published by 2016), we had a larger number of participants [4, 6–8, 12, 18–21, 25, 27, 64, 66, 67, 85–95]; the exception being two multicenter studies conducted as part of the LHMP [19, 67], where one might question the validity of this approach, with significant protocol differences between centers. In relation to our research question, exploring differences in upper

airway contamination levels across bronchoscopic sampling techniques, we found that only five out of the 25 studies included protected sampling techniques [6, 12, 27, 67, 88]. And only three studies included both protected and unprotected sampling techniques [6, 12, 27]. For these latter studies, the number of study participants did not exceed 15.

In summary, this first paper gains power from a relatively large heterogenous study population. The full data-set collected in the MicroCOPD study, is however more than twice the size of the current paper. Although tempting, we decided not to look into details regarding disease state, use of ICS and smoking habits, as this would be the focus for later publications in the MicroCOPD study. In hindsight, we however acknowledge that smoking status may have confounded our analyses. Studies have indicated that smoking may alter the microbiota composition of both the upper and lower airways. Importantly, the upper airways appear to be more impacted than the lower airways [8, 18]. For participants in the “current-smokers” and “ex-smokers” categories, there is therefore the possibility that smoking may have expanded the distance between the lower airway samples and the upper airway samples. With only a few participants in the “never-smokers” category, stratification by smoking status was however not possible.

Papers II and III

For papers II and III, we included 23 participants from the MicroCOPD study - a low number compared to that which was available from the full data-set collected in the main study. However, the objectives of papers II and III were directed at resolving methodological issues associated with laboratory processing steps and the influence of variable sample bacterial load. Due to inherent bias expected to be found across sequencing runs, it was important that where possible, all study samples were included on the same sequencing run, for which there are 96 slots. We also wanted to take advantage of the multiple sample types collected per participant in the MicroCOPD study. The inclusion of multiple sample types per participant and

various technical control samples enabled us to analyze samples from 23 participants in these two studies.

Whilst likely not of critical importance to the research question for papers II and III, we sought to obtain a balance in the number of participants in the healthy and diseased categories (9 control subjects, 14 subjects with obstructive lung disease). This because our hypotheses were driven by the expectation that methodological issues would be directly tied to sample bacterial load, and we recognized the potential for obtaining higher sample bacterial load in participants with respiratory disease.

10.1.2 Procedural samples

Since the aim of the first paper was to gain knowledge regarding airway sampling in microbiota studies, this section of the discussion mainly revolves around the first paper. Our findings, however, greatly influenced the design of the two subsequent papers.

Paper I

We sought to explore the potential of reducing the influence of oropharyngeal contamination through the use of protected bronchoscopic sampling techniques. As described in the introduction, protected sampling involves sampling via a sterile wax-plugged catheter that is passed through the working channel of the bronchoscope. We hypothesized that the catheter would provide protection from contaminating bacteria present in the bronchoscope working channel. Second, we wanted to compare the performance of protected sampling techniques with the most common sample type utilized in studies of the lung microbiome – the unprotected bronchoalveolar lavage (BAL) sample.

In addition to a large heterogenous study population (providing external validity), we included multiple sample types from each study participant – the number of which to our knowledge has not been seen in any previous study. Also, the sampling scheme used in the MicroCOPD study enabled us to account for the difficult and potentially confounding issue of microaspiration and intrapulmonary contamination (providing

internal validity). This was particularly important for the current paper as assessment of bronchoscopic carryover, was based on the similarity between the upper airway OW sample and lower airway samples collected across different sampled sites.

To mitigate the impact of the aforementioned confounding factors, sampling in the MicroCOPD study was performed by collection of different samples across multiple sites and in a strictly specified order: three protected specimen brushes (rPSB) from the right lower lobe (RLL), two fractions of protected bronchoalveolar lavage (PBAL1/PBAL2) from the right middle lobe (RML), three protected specimen brushes (iPSB) from the left upper lobe (LUL), and (unprotected) small volume lavage (SVL) from the LUL. For a smaller subset of participants, the left lung was sampled before the right lung. An oral wash sample (OW) was collected before each bronchoscopy procedure for representation of the upper airway microbiota. For the current paper, we included all sample types collected in the MicroCOPD study. This enabled the evaluation of three sampling techniques: PSB, PBAL, and SVL.

While PSB sampling has commonly been used in studies of the lung microbiome, we were not aware of any study that had previously performed BAL through a protective catheter (i.e. PBAL). This despite the potential benefits of protected sampling via a catheter, particularly against contaminants found inside the working channel as described in the introduction. We found that only three studies included both protected PSB and unprotected BAL [6, 12, 27]. The SVL sample collected in the MicroCOPD study, was obtained by sampling directly through the bronchoscope working channel. As such, comparisons could be drawn to the commonly used unprotected BAL.

The comparison of all sampling techniques would ideally be based on sampling from the same pulmonary site. However, any one sampling event may alter the microbiota composition at the sampled site. Our comparison of the five different sample types, therefore included sampling at three different pulmonary sites – the right lower lobe (RLL), the right middle lobe (RML) and the left upper lobe (LUL). We searched the literature for studies that could help us predict the degree to which sampling across

multiple sites might confound our analyses. The few studies we found, described two different situations when studying healthy [12] and diseased subjects [8, 15].

For healthy subjects, we can look to the adapted island model of lung biogeography introduced by Dickson and colleagues [12], which was described in the introduction. Recall that the model predicts that in healthy individuals, we can expect the bacterial communities of the lungs to resemble that of the upper airways. The model further predicts that as we move down the lower airways, the similarity to the upper airways decreases. We found the model particularly relevant to our paper, as the anatomical distance to the upper airways varies considerably between our three sampled sites (RLL, RML, LUL). Recall again that our analyses were based on the underlying assumption that the more similar our samples were the representative upper airway sample (OW), the greater the influence of contamination by bronchoscopic carryover. The adapted island model of lung biogeography would assume that inherent differences in lung microbiota across sites would result in an expected decrease in similarity to the OW sample in the order of $LUL > RML > RLL$; that is even without the influence of bronchoscopic carryover or contamination that we were looking for. We recognized that the model introduced a potential confounding factor in our study design that we needed to account for in our analyses. In the diseased state, we also can expect greater difference in microbiota across sites that are independent of the influence of microaspiration [8, 15]. However, given the difference in bacterial load between the upper and lower airways, we might also expect that upper airway carryover using unprotected sampling techniques could blur out differences in microbiota across sites in health and even diseased states. This is however currently unclear, and thus the inclusion a large study population with representation from both healthy and diseased states was critical.

As the question of differences in microbiota across sites within the same individual is unclear, it was important that the MicroCOPD sampling scheme also accounted for the potential of contamination across sampling sites. First of all, PSB samples were always collected before lavage (PBAL, SVL) sampling. This to minimize the influence of residual sampling fluid on the surrounding sites. In our literature search,

our critique of one of the key papers in the field by Dickson et al. [12] comparing PSB and BAL sampling, was the sampling of BAL before PSB. Second, for a subset of the participants, we sampled the left lung before the right lung. The importance of this was twofold. First of all, one might predict that the first sampled site would be more influenced by oropharyngeal carryover, as for each subsequent sampling event this contamination would be more “diluted”. Second, repositioning the bronchoscope from one lung to the other involves passing the carina, a site for which microaspirated bacteria likely accumulate [14]. Thus, moving from one lung to the other, contaminating bacteria from the carina may give a false impression of oropharyngeal contamination in our analyses. By sampling also the left lung first, we could assess the influence of moving from left to right lung, right to left lung.

Besides reducing the potential impact of intrapulmonary contamination during sampling, the sampling scheme enabled us to address the difficult issue of distinguishing between microaspirated bacteria and oropharyngeal carryover by the bronchoscope. First of all, by sampling PSB at the two sites with which the anatomical distance to the upper airways varies the most (RLL and LUL), we could determine whether differences in lung biogeography (and thereby microaspiration) was affecting our interpretation of the impact of oropharyngeal carryover. Second, the LUL was sampled using both the most protected sampling technique (PSB) and the sampling technique most vulnerable to oropharyngeal carryover (SVL). This to enable a direct comparison of these two sampling modalities, and therefore the ability to distinguish between the true microbiota likely found in PSB and increased oropharyngeal contamination in SVL. BAL from the RML (same wedged segment) was fractionated to PBAL1 and PBAL2, enabling assessment of the dilution effect described in the introduction. A dilution effect from PBAL1 to PBAL2 could indicate influence of oropharyngeal contamination.

One weakness may be that, even our extensive sampling might not provide the detail needed to resolve the issue of how best to sample the airways. We did not collect samples from the last upper airway site for which the bronchoscope passes before reaching the lower airways (i.e. the supraglottic region). As described in the

introduction, this region likely consists of bacteria from the nasal and oral cavities, and particularly in diseased individuals, the bacterial communities derived from the nasal cavity may be relevant. Samples from this region may therefore be more representative of microbiota that are “microaspirated”. Samples from the lumen below the vocal cords might have told us more about what the bronchoscope brings down to the sampling sites, and samples from the carina and main bronchi might have shed light on the fraction of upper airways microbiota that reaches the lower regions. However, such comprehensive sampling would have greatly extended the procedure time with an associated increased need of sedation/anaesthesia and possibly also procedural complications.

There are also alternatives to our protected sampling methods. For instance, some investigators use a two-scope technique where they change bronchoscope after anesthetizing the vocal cords; this to minimize contamination of the bronchoscope used for sampling. However, this implies more movement up and down the airways, and opens other routes of contamination. Protection might also be accomplished using a balloon catheter when performing BAL. This might minimize a “washing effect” of the bronchoscope tip when sampling lavage, and also increase yield by sealing the sampling site and preventing leakage to other parts of the airways.

After sample collection, the next step in our microbiome analysis workflow was bacterial DNA extraction. As discussed in the introduction, controversy exists on whether eukaryote host cells should be removed from samples before proceeding with DNA extraction [64, 65]. While some studies have used acellular samples (eukaryote cells removed) [20, 25, 66], most studies have used whole samples (eukaryote cells kept). For the MicroCOPD study (and hence all papers of the thesis), we found the use of whole samples to be the most valid choice, as important members of the lower airway microbiota may be associated with eukaryote cells, for instance via biofilms [64]. However, there is currently no consensus in the field regarding the optimal sample type (acellular or whole). In fact some studies have even combined datasets obtained using both acellular and whole BAL, as in the multicenter LHMP study [67]. While the optimal sample type remains to be determined, we argue that

consistency in methods is most important and for such studies reliability/validity is questionable.

Papers II and III

Four different sample types were analyzed from each of the 23 participants included in the papers II and III: oral washes (OW), protected specimen brushes from the right lower lobe (rPSB), protected bronchoalveolar lavage (PBAL) from the right middle lobe, and negative controls samples (NCS).

Importantly, we included both high and low biomass airway samples. The lower airways were represented by two different sampling techniques (PSB and PBAL). The SVL sample was not included, as in our hands protected sampling procedures appeared to provide protection from upper airways (paper I). In addition, the included samples were taken from the same lung, thereby minimizing the impact of intrapulmonary contamination [12, 14]. By including the same participants and samples in papers II and III, we strengthened our analyses and were able to compare our different bioinformatic approaches. In paper II, the bacterial load was established for these samples, providing more validity to the conclusions made for paper III based on assumed differences in bacterial load.

10.1.3 Bacterial DNA extraction

Papers I, II and III

As discussed in the introduction, the bacterial DNA extraction step may introduce bias to a study if the genomic DNA is not extracted with equal efficiency from all bacterial members of the sampled microbiota. The protocol for DNA extraction used in the MicroCOPD study, was designed in-house and based on what we perceived as the best of knowledge currently available for securing optimal bacterial community representation.

Samples were first treated with a combination of the three lytic enzymes (lysozyme, mutanolysin and lysostaphin), as recommended in Yuan et al. [29]. By using a

combination of three enzymes we addressed the potential resistance of bacteria to the lytic activity of any one particular enzyme. Bacterial cells that were not sufficiently lysed on treatment with the enzyme cocktail, were subjected to mechanical lysis by bead beating. Importantly, genomic DNA that had been successfully isolated by treatment with enzymes, were removed before proceeding with the bead beating step. This to avoid the shearing of genomic DNA, for which had been successfully extracted on treatment with the enzymes and hence potentially minimizing the formation of chimeras in subsequent PCR steps, as described in the introduction.

A weakness to our study is perhaps that we did not validate our DNA extraction protocol against a mock community sample (MC). While this would have provided us with an indication of the differences in extraction efficiencies across bacteria with different cell wall structures, it can however be argued that a MC sample will anyhow not accurately reflect the true complexity of a natural sample.

We recognize the possibility that differences in extraction efficiency across bacterial taxa may have impacted our analyses. Recall for instance that for paper I, our assessment of upper airway carryover by the bronchoscope was based on the degree of similarity between the lower airway samples and OW samples. If a taxon found exclusively in OW samples is extracted with low efficiency, the OW samples may appear to be more similar to lower airway samples than is the actual case. Analyses for paper III were aimed at elucidating error and bias associated with laboratory processing steps occurring after DNA extraction. By using the same DNA extracts as input to each of the three library preparation setups compared, we minimized the potential for bias associated with DNA extraction.

10.1.4 Determination of bacterial load

For paper III, the levels of bacteria in our samples were determined by probe-based quantitative PCR (qPCR) targeting the bacterial 16S rRNA gene region V1 V2. We chose to use the same primer/probe set utilized by several others in the field [6, 7, 14, 21]. The standard curve used for determining absolute bacterial numbers was constructed from a ten-fold dilution series of *E. coli* genomic DNA.

The main weakness with our method, is that we did not account for the presence of human DNA in our samples. Human DNA may act as a competitive inhibitor in PCR, resulting in low reaction efficiencies. This is particularly a concern for samples where the levels of human DNA are high and the levels of bacterial DNA are low. Glassing and colleagues [96, 97] have shown that the presence of human DNA may have an impact not only on the 16S sequencing results for community profiling [97], but also on quantitative assessments of bacterial load [96] as discussed here (both are methods that build on the PCR). Glassing et al. [96] found that submucosal intestinal tissue samples (low biomass) containing high amounts of human DNA generated C_T values (i.e. threshold cycle values) greater than that obtained for no template controls – i.e. their results indicated that there were higher levels of bacteria in the no template controls than in their tissue samples. We cannot dismiss the possibility that the presence of human DNA may have influenced our results. This particularly so because the samples used for construction of the standard curve – a pure culture of genomic *E.coli* DNA – was devoid of human DNA. One may therefore question whether possible differences in reaction efficiency between the standard samples and the samples being tested due to differences in human DNA content, could have led to an underestimation of bacterial load.

10.1.5 Library preparation for sequencing

Papers I and II

For the first two papers of the thesis, library preparation for sequencing was performed using the commercial protocol by Illumina with title *16S Metagenomic Sequencing Library Preparation* (Part # 15044223 Rev. B). Methodological issues associated with choice of protocol and target marker gene region are central themes in paper III and is discussed in section 10.2.3. The number of PCR cycles used for amplification of the target marker gene was increased from 25 cycles as specified in the commercial protocol, to 45 cycles. This was necessary in order to obtain adequate levels of DNA for sequencing. The issue of PCR cycle number was addressed in paper II and is discussed in section 10.2.2.

Paper III

Three setups for library preparation and sequencing are compared in the third paper of the thesis: setup 1 (2-step PCR; V3 V4 region), setup 2 (2-step PCR; V4 region), setup 3 (1-step PCR; V4 region). The three setups were chosen to best answer two questions. First of all, will a 2-step PCR protocol render samples more vulnerable to laboratory contamination than a shorter 1-step PCR protocol? And second, will choice of target marker gene region (16S rRNA gene V3 V4 or V4) have an impact on final bacterial community descriptions?

To address the first question, two protocols that differed with regards to the number of PCR steps was required. To ensure external validity, we sought to find protocols widely used in the field. For the 2-step PCR protocol (setups 1 and 2), the choice fell on the commercial protocol by Illumina. The protocol was used for preparing samples for sequencing in the MicroCOPD study, and hence the findings from paper III could be extended to interpreting results generated in previous and future papers in MicroCOPD. The 1-step PCR protocol was based on the protocol described in Kozich *et al.* [49].

To address the second question, we first had to decide on which 16S rRNA gene variable regions to compare. Studies on the lung microbiome have been based on a wide range of 16S rRNA gene targets - including V1 V2 [6, 86], V1 V3 [8, 18, 19], V3 V5 [7, 19, 21, 64, 85, 98], V3 [87, 99], V4 [12, 14, 25, 67] and V3 V4 [23] – a decision regarding which target regions to compare was not a given. We however decided on the regions V3 V4 and V4. The V4 region stood out as the optimal choice as studies have collectively shown that the region generates the most accurate estimates of the three commonly used parameters - alpha diversity [100], beta diversity [38] and taxonomic assignment [79]. The short length of the V4 region also comes with the added advantage of enabling fully overlapping PE sequencing reads – which as discussed in the introduction, has shown to be a powerful means by which error correction may be performed (section 6.3.2). With the development of sequencers with increased read lengths, and novel denoising strategies (e.g. DADA2

[59]) longer regions are now becoming more favourable. The V3 V4 region was chosen as the second target as this was the region targeted in the MicroCOPD study, again enabling the expansion of current findings to the previous and future work in the MicroCOPD study.

Papers I, II and III

A third consideration, relevant to all three library preparation setups is the potential of contamination between study samples processed together on the same sequencing run (i.e internal contamination). Such internal contamination has been perceived as less of a concern than external contamination derived from bronchoscopic carryover or the general laboratory environment. We can distinguish between two types of such internal “between sample” contamination - that introduced during steps of library preparation for sequencing and that introduced during the sequencing process itself [101].

Internal contamination can arise during library preparation as a result of “well-to-well” contamination between samples that are placed next to each other on the PCR plate [101]. Library preparation for sequencing was performed for 96 samples at a time on 96-well PCR plates and involved multi-protocol workflows. Each individual protocol involved an extensive amount of manual pipetting, for which we used multichannel pipettes.

The issue of internal contamination across samples on the same PCR plate may be particularly relevant to the MicroCOPD study, due to the inclusion of multiple samples types and differences in bacterial load across sampled sites. Because samples were organized on the PCR plate according to bronchoscopy procedure and not sample type, samples of high bacterial load (i.e. OW) were placed directly adjacent to samples of low bacterial load (i.e. lower airway samples, NCS). The potential for well-to-well contamination in the direction of OW to lower airway sample (PSB, PBAL, SVL) is concerning because of the expected overlap in bacterial communities between the sampled sites. In much the same manner as described for microaspiration earlier, such internal contamination may for instance confound our analyses aimed at

comparing the similarity between OW samples and each of the different lower airway sample types (paper I). Well-to-well contamination in the direction of lower airway to NCS sample is also a concern, as this could lead to the false labelling of authentic airway community members as external contaminants (i.e. those derived from the laboratory environment or reagents). We did not conduct any experiments to determine the frequency of such contamination in the MicroCOPD study. We did however take the preventative measure of strictly using pipette tips with filters to avoid carryover by contaminated pipettes.

Internal contamination may also occur during the sequencing process itself due to the issue of index misassignment (section 6.3.1). All three setups used in the current thesis were based on a so called combinatorial dual indexing approach based on 8 forward primers with unique so called “i5” index sequences and 12 reverse primers with unique “i7” index sequences. Amplicons derived from different samples may therefore share the same “i5” or “i7” index sequence, but not both. As with the issue of well-to-well contamination, we did not experimentally determine the frequency of index misassignment in our library preparation setups. The use of unique indexes for all, could have reduced the potential for index misassignment.

While the issue of internal contamination is not addressed in the current thesis, one may question whether the greater number of steps associated with a 2-step PCR protocol (setups 1 and 2) relative to a 1-step PCR protocol (setup 3) could leave samples more vulnerable not only to external contamination, but also internal contamination such as that described here. First of all, as samples are processed through a greater number of protocol steps in the 2-step PCR approach, there might be more opportunity for well-to-well contamination. Also, the timing of the index PCR step may be important. Recall that indexing is performed so that sequences are labeled according to the sample from which they originate. One may therefore expect that the earlier addition of index sequences in a 1-step PCR approach could lower the impact of well-to-well contamination compared to a 2-step PCR approach. This has however not been evaluated in the literature.

Finally, also relevant to all three papers is the issue of repeatability of analyses. Ideally all laboratory processing, including sequencing, should have been replicated to avoid spurious findings dominating our conclusions. Even if we did not replicate analyses in paper II and III, hopefully the number of samples processed and consistency in our findings mitigate some of this weakness in our design.

10.1.6 Bioinformatics processing

Paper I

Bioinformatic sequence processing and analyses were conducted using tools available within the QIIME1 package. As with laboratory protocols, we found that there was no consensus in the field with regards to bioinformatic sequence processing steps or analyses. Most decisions were therefore made based on the default settings and recommendations provided by the QIIME development team [102]. Nonetheless, we had to make decisions regarding stringency of quality filtering and the approach to dealing with NCS.

Quality filtering.

In two key steps of the pipeline we had to make a decision regarding the stringency of quality filtering. First, when joining paired-end reads, a decision had to be made regarding the degree of overlap between the forward and reverse reads, and whether or not we would allow any discrepancies between the two overlapping regions. We demanded a minimum overlap of 100 bases and allowed for zero discrepancy. This was quite strict and in hindsight it is clear that we may have lost reads that could have been “saved” by error-correction using the read with a higher base quality score (section 6.3.2). Further down the pipeline, quality filtering was performed using default settings in QIIME1 [52] (6.3.2) but for which we increased the default quality score set to 3 to 19, thereby filtering out reads with lower Q scores than 20. With such a high threshold, we may have filtered out many accurate sequences together with erroneous sequences.

Approach to NCS.

A main strength of the MicroCOPD study, is the collection of procedural NCS for each separate bronchoscopy procedure. The NCSs were never in contact with the study participant or the bronchoscope, but were processed through all subsequent laboratory steps (DNA extraction, library preparation for sequencing and sequencing) alongside the procedural samples. This provided us with a unique opportunity to address the issue of contamination from the laboratory. However, the literature provided us with few guidelines for dealing with NCS and approaches varied across studies. The simplest approach found in our literature search, was the removal of all OTUs or taxa observed in NCS (i.e. the “remove all” approach) [21, 86]. Although simple, the approach does not account for taxa naturally overlapping both NCS and airway samples. *Pseudomonas* is for example commonly found in the lungs [21], while also a typical laboratory contaminant captured in sequenced NCS [20, 63]. The “remove all” approach also does not account for the potential impact of internal contamination that may occur in the direction airway sample to NCS, as described earlier (section 10.1.5). Another approach involved the select removal of probable contaminants based on reports from previous publications (i.e. the “black-list” approach) [63, 103]. The obvious limitation with the “black-list” approach, is that contamination may vary greatly from study to study, and even within the same study, variation can be expected across different time points and when using different lots of reagents and kits for sample processing [63]. Furthermore, we found that others had removed sequences observed in NCS based on arbitrarily chosen abundance thresholds [99]. However, we found that it is unclear where to draw the line with regards to a set abundance threshold level. The perhaps most sophisticated approach found in our literature search, was the application of the neutral model of community ecology for detection of likely contaminating OTUs [19]. The method was however used in few publications and we found it challenging to perform with our limited bioinformatics experience. While the above mentioned methods were based on identification of contaminating sequences to remove from the dataset, alternative approaches were based on the removal of entire samples that overlapped with NCS in ordination space [18, 20, 99]. Regardless of chosen method of handling NCS, a

common limitation to all methods is the possibility that not all contaminants are represented in the NCS [103].

With the backdrop described above, we found that there was no clear best approach to dealing with contamination and NCS. For paper I, we therefore chose the simplest approach – the removal of all OTUs found in NCS. As NCS were collected for each bronchoscopy procedure, OTUs were only removed from the corresponding procedural samples collected under the same procedure. Thus limitations associated with the “remove all” approach, with regards to potential impact of internal contamination (resulting in the removal of biologically relevant taxa) would only impact samples collected under the same bronchoscopy procedure. Despite the risk of removing OTUs naturally occurring in both airway and NCS samples due to an external contamination source potentially influencing all samples on the same sequencing run (e.g. DNA extraction kit or PCR reagents), we found the “remove all” approach satisfactory in light of our research question. This because the same OTUs were subtracted from the OW and lower airway samples. We therefore reasoned that our comparison of the upper airway OW to the lower airway samples would not be largely affected. In contrast, if samples were kept intact (i.e. no contaminant removal strategy), the influence of contamination would likely be greater on samples with low bacterial load (i.e. the lower airway samples) – possibly inflating the difference between upper and lower airway samples. The challenges associated with handling NCS in analyses for paper I prompted us to address the issue in the second paper for the thesis, as described in the subsequent section.

Paper II

The bioinformatic processing steps for paper II were conducted as for paper I, using tools available within the QIIME1 package. For the current paper we however made it an objective to explore bioinformatic strategies for handling NCS for removal of laboratory contamination. The focus of the current discussion will therefore be on our choice of strategies to compare.

As described for paper I, the inclusion of NCS was quickly becoming a requirement for publication in most journals. We however found that guidelines for handling NCS once they were collected were still lacking, even now years after the first paper was published. The bioinformatic field had however picked up on the issue and several tools were under development. Of particular interest to us was the Decontam package available in R [82], which was developed directly for dealing with contamination in amplicon-based studies. For the current paper, we sought to compare three strategies for dealing with contamination: i) keep all samples intact (i.e. do nothing), ii) remove all OTUs found in NCS and finally iii) remove OTUs identified as contaminants by Decontam.

While the first two approaches were rather straightforward, the Decontam R package presented us with several options. Contaminant identification is performed using the Decontam *isContaminant* function using one of several methods; including the “frequency” or “prevalence” based methods. A third method “either”, combines the former two methods. The choice of method (“prevalence”, “frequency” or “either”) in any study will first of all depend on the availability of auxiliary data. Negative control samples are required when performing the “prevalence” based method. DNA quantitation data are required when performing the “frequency” based method. The “either” method uses both the “prevalence” and “frequency” methods and as such requires both negative control samples and quantitation data. For the current thesis, both negative control samples and DNA quantitation data (qubit measurements) were available, enabling us to take use of the Decontam approach “either”.

It was decided that the strictest Decontam method would be chosen, for best comparison to the «removal all» approach. Therefore, we chose the «either» method, securing maximum contaminant identification where one of either the «frequency» or «prevalence» based method would fail. The methods were validated on a ten-fold dilutions series of *Salmonella* (SDS), for which we were able to confirm that the «either» method was most effective.

The frequency based approach and its dependency on total DNA measurements might be criticized. The total DNA measurement before loading the MiSeq is based on rather crude methods and will also measure non-bacterial DNA. However, the PCR has selectively amplified microbial DNA, and both our sequencing results and the sub-study investigated with qPCR has revealed a bacterial load that is worth examining. Furthermore, the non-bacterial DNA would most likely serve to weaken associations in a non-discriminant manner, and not affect the validity of identified contaminants.

Paper III

For the third paper, bioinformatic sequence processing steps were conducted using tools available within the QIIME2 package. A central tool in the pipeline is DADA2 [59], which has the primary function of denoising sequences to ASVs. The DADA2 workflow also includes all steps of filtering, dereplication, chimera removal and the merging of PE reads – all steps of which are executed in a single command using the DADA2 plugin in QIIME2. We chose to perform additional steps of chimera removal and abundance filtering post DADA2 processing – procedures that have not (yet) been recommended or expected to be necessary when working with ASVs. A discussion on our decision to include additional these steps follows.

Chimera removal.

We chose to perform two rounds of chimera removal because different algorithms will vary with regards to both sensitivity and specificity. The first round of chimera removal was performed as an integrated part of the DADA2 workflow, for which the *de novo* based approach is `BimeraDenovo()` is used [59]. The method is applied after denoising to ASVs and is highly specific for the detection of exact chimeras formed between two parent sequences (bimeras). For the detection of chimeras formed between more than two parent sequences (multimeras), we applied a second round of *de novo* based chimera removal using the `vsearch uchime-denovo` method [76], which is also available as a plugin in QIIME2. The method originally developed with OTUs in mind, offers less specificity but higher sensitivity to chimeras – possibly higher

risk of false positive identification. To my knowledge the use of two filtering procedures on ASV data has not been benchmarked although this has been discussed vaguely on the QIIME2 user forum. While the use of two rounds of chimera detection may seem excessive and with an added risk of false positive detection, we found it appropriate due to experimental conditions that may have left our data particularly vulnerable to recombination events during PCR; i.e. a high number of PCR cycles [46] and bead beating during DNA extraction [31]. In addition, studies have suggested that when using a one step protocol (setup 3), where longer primers are used, chimera formation may be increased.

The additional “Bokulich” filter step.

An additional abundance filtering step was also performed. In short, recall that ASVs for which there were fewer sequences than 0.005% of the total number of sequences were removed. This threshold abundance level was rather arbitrarily chosen, and based on the recommendations provided by Bokulich et al. [52] when performing quality filtering measures on OTU based data. As described in the introduction, the method was originally intended to reduce the number of spurious OTUs generated as a result of PCR and sequencing error, for which should not in theory be an issue when working with ASVs. ASVs are however a relatively new unit in the microbiome field and there are few recommendations for handling ASV data – and particularly so with regards to low biomass samples. We suspected that low abundant ASVs may reflect contaminants, undetected chimeras or organisms with little biological relevance, and chose to filter based on the “Bokulich” method as performed also for papers I and II when working with OTUs. To us it seemed unlikely that the 23 samples from the airways would hold over 1000 different taxa.

The bioinformatic pipeline used differed for different sample types depending on the question being asked and purpose for these samples in subsequent analyses steps. While the procedural samples were processed through all steps of the bioinformatics pipeline - including denoising by DADA2, chimera removal by VSEARCH, removal of small ASVs (i.e. the “Bokulich method”), removal of ASVs not classified at

minimum the phylum level and finally the removal of ASVs identified as contaminants using Decontam - the mock community (MC) and negative controls were handled differently. For MC samples, the main question we sought to answer was whether the three setups were equally efficient at recovering the different MC members. We also wished to determine the impact of contamination (i.e. all ASVs assigned to taxa not expected in the MC). Bioinformatic processing steps were therefore limited to DADA2, the additional chimera removal step by VSEARCH and the removal of ASVs that did not classify at minimum to phylum level. Removal of small ASVs (i.e. the “Bokulich method” described above) and removal of contaminants identified using Decontam, was not performed as we also wanted our analyses to capture the impact of contamination. The negative control samples – including both procedural NCS and PCR water samples, were processed through all steps of the pipeline except for the removal of contaminants identified using Decontam. As such the MC sequencing output is more reflective of the “raw” sequencer generated data, as small ASVs were not removed.

10.1.7 Analyses

Analyses were based on the OTU (QIIME1) or ASV (QIIME2) tables generated after bioinformatics processing of samples as discussed above. Common parameters used for microbiome analyses include i. taxonomy, ii. alpha- and iii. beta-diversity.

Taxonomy

Analyses of taxonomic composition were performed in all three papers using average relative abundance of taxa in the sampled communities. While this is a common approach in the literature, it is important to note that taking averages may distort conclusions if major or minor taxa are driven by extreme samples lacking taxa or overrepresented by taxa, for example as a result of internal contamination. While for paper I, the study population was large and this may not be an issue, it must be acknowledged that for papers II and III this may have affected our interpretation of the data.

Alpha diversity

Within sample comparisons (alpha diversity) were made using Faiths PD (paper I). For papers II and III alpha diversity was not assessed beyond the mention of number of sequences and OTUs/ASV generated per sample. However, without rarefaction, accurate comparisons across samples could not be made. But, nevertheless, this still provided us with useful information regarding differences in sequence depth.

Beta diversity

Between sample comparisons (beta diversity) were made using PCoA of unweighted UniFrac distances. By choosing unweighted rather than weighted UniFrac, every detected OTU in the samples were given equal significance regardless of relative abundance. This is likely important in order to recover small differences between samples. The upper and lower airway communities, for instance, appear from the literature and own analyses to be dominated by many of the same taxa. We may therefore expect that the use of the weighted UniFrac metric could result in the masking of small differences between samples. On the other hand, the limitation with the use of unweighted UniFrac is that equal significance is also given to OTUs/ASVs derived from contamination. Our bioinformatic pipeline, however included the filtering of small OTUs/ASVs, which may have reduced the impact of low abundant contaminants on analyses of beta diversity.

Statistical analyses

Analyses of microbiota based on next generation sequencing data is a relatively new field. Both the nature and magnitude of data available for scientists has changed substantially over the last few years, and we now analyze millions of sequences from several hundred samples. The sequences are again organized in several hundreds (at least) units (ASVs, OTUs, taxonomic levels). Associated features of the resulting data sets make statistical analyses particularly challenging.

First the number of variables is very high, resulting in a multiple comparison problem. This is a problem well known from genetic studies, and although statistical corrections are available (Bonferroni, FDR), these tend to over-compensate [104].

Second, there is a large number of zero values which makes choosing a statistical distribution to base tests on, difficult [105]. Third, many of the parameters have a compositional distribution, which excludes many conventional statistical methods. And finally, the data presented in the current thesis are mostly paired in some way or another, and compared parameters are not independent of each other.

Currently, there are a plethora of suggested workarounds for most of these problems. However, there is no agreed-upon solution, and when choosing one approach you often face limitations that necessitates analyses by yet another method. The results might conflict, and the researcher might end up in a conflict between full disclosure and the need to present a clear message.

Nevertheless, most of the objectives in the current thesis were to shed light on methodological issues. We therefore chose to focus on descriptive analyses, and visualizations of these, to provide investigators with as much information as possible without categorical conclusions based on uncertain statistical tests.

10.2 Discussion of main results

10.2.1 Paper I

Our evaluation of the impact of upper airway contamination when using different bronchoscopic sampling techniques was based on the underlying assumption that the more similar the lower airway specimens (PSB, PBAL and SVL) were the OW sample, the greater the influence of upper airway contamination on these samples. Between sample comparisons were made based on three parameters: i. taxonomy, ii. alpha diversity and iii. beta diversity.

Taxonomy

For comparison of taxonomic composition, we looked at the average relative abundance of the most prominent phyla by sample type. A clear trend with decreasing similarity to the OW sample in the order OW>SVL>PBAL1>PBAL2>rPSB>IPSB

was observed. Driving this effect was an increase in *Proteobacteria* and a concomitant decrease in *Firmicutes* across sample types.

The order by which similarity to the oral wash sample decreased across sampling types was in accordance with our prediction of which sampling techniques would offer the most protection from upper airway contamination. SVL being the only sample type that was unprotected, showed greatest resemblance to the OW sample, as expected. The IPSB sample for which sampling was performed from the same site as SVL, showed the least resemblance to the OW sample of all sample types. This indicated that differences in susceptibility to upper airway contamination and not lung biogeography, was responsible for the observed differences between samples. Also as expected, PBAL samples showed less similarity to the OW sample in the PBAL return 2, compared to PBAL return 1. As described in the section 6.1.3, this may indicate a dilution effect as upper airway contamination may be “diluted off” after the first PBAL sampling. Despite the use of protected sampling when using PBAL, the dilution effect may still be prominent if the outside of the bronchoscope is a major source of bronchoscopic upper airway carryover. Contamination from the outside of the bronchoscope channel may also be captured better by washings (i.e. PBAL) than brushings (i.e. PSB).

The increase in *Proteobacteria* and simultaneous decrease in *Firmicutes* across sample types was however more challenging to interpret. *Proteobacteria* have previously been associated with contamination from the laboratory that is more pronounced in samples of low bacterial load. This was for instance found in the study by Biesbroek et al. [32] across serially diluted samples of saliva (section 6.4.1). Thus, the increase in *Proteobacteria* as samples become less similar to OW, may reflect a decrease in total bacterial load that have left these samples more influenced by laboratory contamination. The higher levels of *Proteobacteria* in PSB related to PBAL may also be expected as the input volume to DNA extraction was lower for PSB compared to PBAL (450 µl for PSB and 1800 µl for PBAL), thereby possibly securing a lower bacterial load for these samples. The observed simultaneous decrease in *Firmicutes* across sample types may also reflect a reduced signal from the

sampled microbiota. Indeed, *Proteobacteria* includes important bacteria of the airways such as *Haemophilus*, *Legionella*, *Pseudomonas* and *Burkholderia*. Without having quantified differences in bacterial load across sample types, we could however not further conclude on these possibilities. In addition, our approach to laboratory contamination was to remove all taxa found in NCS. The accuracy by which this approach is able to address all laboratory contaminants is not known, but is a central topic in paper II.

Alpha diversity

Alpha diversity measurements shed light on the level of biodiversity found *within* a single sample. In its simplest form this can be a count of the number of different operational taxonomic units (OTUs), amplicon sequence variants (ASVs) or bacterial species. The Faith's phylogenetic diversity (Faith's PD) metric used in the current study, takes into account not only the number of OTUs, but also how phylogenetically similar these are to one another. It does this by adding together the total branch length between all OTU placements in the phylogenetic tree. Thus, a sample will be more diverse the greater the number of OTUs and the greater the phylogenetic distance between these OTUs.

We found that diversity decreased across sample types as the level of protection from upper airway contamination increased: OW>SVL>PBAL1>PBAL2>rPSB>lPSB. Importantly, this trend was the same as that observed when comparisons of taxonomic composition were made. Sampling the left or right lung first did not have a great impact on measurements of alpha diversity.

Beta diversity

Beta diversity measures the degree of (dis)similarity *between* samples. As with alpha diversity, we decided to incorporate phylogenetic information into our analyses. Using the UniFrac metric [106], the distance between samples was measured based on the branch length in the phylogenetic tree that is shared between their bacterial communities. The calculated distances between all samples were stored in a distance

matrix, and the (dis)similarity between samples visualized by principal coordinates analysis (PCoA) plot.

Using PCoA of unweighted UniFrac distances, we compared the OW samples to each of the five different sample types. We found that the similarity to the OW sample decreased in the order of SVL>PBAL>PSB. It was difficult to visually assess whether the overlap between PBAL1 and OW or PBAL2 and OW was greater and likewise, whether the overlap between IPSB and OW or rPSB and OW was greater. However, the general trend with regards to the three sampling techniques (PSB, PBAL, PSB) was in agreement with our results for both taxonomy and alpha diversity.

By performing a permutational multivariate analysis of variance (PERMANOVA) test on the calculated UniFrac distance matrix, we could ascertain whether the (dis)similarities visualized in PCoA space were significant. This was particularly important for evaluation of the sample types mentioned above, that visually appeared to overlap equally with the OW samples in PCoA space. We found that all comparisons of OW and sample type were significant. Importantly, we were also able to confirm that the differences between OW and sample type increased in the order SVL< PBAL1< PBAL2< rPSB< IPSB as indicated by an increasing pseudo F-statistic.

Our results led us to conclude that protected sampling methods (PBAL, PSB) diminish the influence of oropharyngeal carryover by the bronchoscope. Our conclusion was however not in complete agreement with the findings presented in the three papers found in our literature search, for which both protected and unprotected sampling was performed [6, 12, 27].

Charlson et al. [6] compared PSB samples (from the LLL) and BAL fluid (from the RML) obtained from six healthy subjects. Using PCoA analyses on weighted UniFrac distances, they found that all samples from lungs (irrespective of sample type) clustered together with samples collected from the upper airways (oral washes (OW),

oropharyngeal swabs (OP)). Thus, protected sampling did not appear to influence the degree of similarity to upper airway samples. The study however lacked power in terms of the number of study participants and only one PSB sample was collected per subject. The sequencing depth was also lower than in our study. The authors also did not report results from PCoA analyses on *unweighted* UniFrac distances, which may have resulted in a different interpretation of the data. The combination of weighted (as opposed to unweighted) UniFrac analyses, and low sequencing depth may explain why differences were not detected between lower and upper airway samples, as seen in our study.

Dickson *et al.* [12] compared PSB samples (from RUL and LUL) and BAL fluid (from RML and lingula) obtained from 15 healthy subjects. Based on principal component analyses of beta-diversity, they found that samples from the lungs (on average) clustered separately from that of the upper airway sample (PSB from the supraglottic region). They found no clustering by sample type, but noted that PSB samples from the RUL, showed the greatest resemblance to the upper airway sample. Collectively their results indicated that protected sampling techniques did not influence the degree of similarity to the upper airway sample. However, the sampling of BAL before PSB might have resulted in residual BAL fluid influencing the sampled PSB site. In the design of the sampling scheme used in the MicroCOPD study, we were careful to always sample PSB before BAL.

Hogan *et al.* [27] compared PSB samples of mucus plugs and BAL fluid obtained from nine patients with CF. PSB and BAL were sampled from multiple lobes on the right lung (RUL, RML, RLL). On comparison of samples taken from the same lobe, they found that measures of alpha diversity (based on the Simpson Diversity Index metric) were consistently higher in PSB than in BAL. This was in direct contrast to our results. However, the CF lung likely reflects a completely different scenario than the healthy and even diseased COPD/asthma lungs sampled in our study. First of all, the CF lung is not considered particularly low biomass and therefore the impact of oropharyngeal contamination during sampling may be negligible. Second, mucus

plugs may form specific niches for microbial colonization that may impact the comparison of sampling techniques. Their study did not include healthy control subjects. Our study included a greater number of participants, for which the observed decrease in diversity (SVL>PSB), was observed for both healthy and diseased states.

10.2.2 Paper II

Our analyses for the second paper were aimed at i) estimating levels of contamination across different airway sample types, ii) determining the main contamination source when processing samples through the MicroCOPD laboratory workflow, and iii) exploring bioinformatic approaches to dealing with the issue of contamination.

Bacterial load varies with sample type

The bacterial load was determined for the following procedural samples: OW, rPSB, PBAL1, and PBAL2. We found that bacterial load varied with sample type and decreased in order OW>PBAL1>PSB>PBAL2 ($p < 0.001$, non parametric trend test). We did not find differences in bacterial load across diseased states, which may reflect the fact that our diseased subjects had a fairly high lung function. Due to a low number of study participants, we therefore did not stratify our analyses on disease category. A discussion on our interpretation of these results follows.

The average bacterial load in the samples from the lungs was highest for PBAL1. The bacterial load in PBAL2 and PSB samples were approximately an order of magnitude lower. PBAL1 and PBAL2 were obtained from the same wedged position of the RML, and the same volume sampling fluid was instilled. Thus, these samples can be directly compared. The observed decrease in bacterial load across these sample types (PBAL1>PBAL2) could be interpreted in several ways. First of all, it could mean that the first lavage fraction (PBAL1) collects a larger portion of the resident microbiota, by primarily sampling the more proximal airways, and also “cleaning up” the secretions that the bronchoscope might bring with it from its passage down the airways. However, the decrease in bacterial load could also be a result of a dilution effect, as lavage yield tends to increase in second fraction (PBAL2). Charlson et al.

[6] also compared the bacterial load obtained when sampling two fractions of BAL from the same site (site A). As us, they observed a decrease in bacterial load from the first to second BAL fraction. In addition, they sampled BAL from an adjacent site (site B), for which they found similar levels of bacteria as in the second return of the sampled site A. Their interpretation was that the first BAL fraction was contaminated with bacteria from the upper airways. Although we did not include sampling from an adjacent site “B”, the Charlson study provides us with another interpretation of our data – that PBAL1 may be more susceptible to contamination from the upper airways than PBAL2. This is also in agreement with results from paper I, where we found that PBAL1 was more similar to the upper airway OW sample than PBAL2 in terms of taxonomy, and measures of both alpha and beta diversity. PSB samples represent a different sampling modality than PBAL, and therefore we could not make comparisons across these sample types.

The comparison of bacterial levels across studies is difficult because of the lack of standards in the field with regards to protocols for sampling and DNA extraction - and quantification of bacterial load is performed on samples after processing through these steps. To illustrate this, we can again look to the study conducted by Charlson et al. [6], for which we can find some similarity in protocols to those used in the current study. As us, they sampled BAL from the RML by instilling 50 mL saline and used 1.8 mL of the returned BAL as input to their DNA extraction protocol. Our protocols for DNA extraction however differed. Without conducting a head-to-head comparison of the two DNA extraction protocols, it is unclear whether observed differences in sample bacteria load are due to actual differences in sampled bacterial levels or a result of differences in extraction efficiency between protocols. Equally important is likely differences in contamination levels introduced when processing samples through different DNA extraction kits. Commonly used DNA extraction kits are not free of bacteria, and differences in contamination levels across kits can be expected [63]. Importantly, the measured bacterial load will reflect both the sampled microbiota and contamination introduced from sampling and DNA extraction [63].

Bacterial load and impact of laboratory contamination

Estimates of contaminant levels for samples with varying bacterial load were made using the “Salter approach” described in the introduction. In brief, our analyses included a ten-fold dilution series of *Salmonella* (SDS). For evaluating the impact of varying PCR cycle number, the SDS was processed through two PCR protocols differing only in the number of PCR cycles (30 and 45 cycles).

On analysis of the sequencing output for the SDS samples, we observed the expected inverse relationship between sample bacterial load and the proportion of sequences mapping to taxa other than *Salmonella* (i.e. contamination) [32, 63]. At an input of between 10^3 and 10^4 *Salmonella*/mL, we observed that contaminants constituted more than 50% of the bacterial community profile for a sample. Despite differences in protocols, our results were in accordance with the findings by Salter et al. [63]. This may be explained by the use of similar DNA extraction kits (both from MP Biomedicals, FastDNA Spin Kit). Further supporting our interpretation is the study by Biesbroek et al. [32], who found that choice of DNA extraction kit determined whether low biomass samples fell above or below their set threshold bacterial load for which contamination begins to dominate. When processing the SDS through an increased number of PCR cycles, we observed only a small increase in the proportion of non-*Salmonella* taxa (i.e. contamination). Thus the impact was low, validating the protocol used in the MicroCOPD study (45 cycles used).

The SDS experiment was used to estimate levels of contamination in the different airway sample types (OW, PSB, PBAL). The OW sample appeared to be unaffected. For the lower airway samples (PSB, PBAL) however, an estimated 10-50% of the sequencing output was expected to be derived from contamination. A limitation to the Salter approach is that it does not capture PCR incorporated error (e.g. chimeras) that may be associated with a more complex natural sample (i.e. airway samples). As described in section 6.2.2, such erroneous sequences may result in sequences mapping to taxa not found in the sampled community. Although different than external contamination derived from the laboratory environment and reagents, such PCR incorporated errors would have the same impact on the bacterial community readout – i.e. the lowering of the relative proportion of true sequences from the

sampled community. We may therefore expect that the procedural samples would be more impacted by PCR cycle number than what was estimated using the SDS.

The objective of the SDS experiment was to demonstrate a method by which levels of contamination could be reported on in airway microbiome studies, and not to define a threshold bacterial load applicable to all studies. Important because contamination levels may vary greatly across studies due to differences in protocols. For a more detailed discussion on protocol effects on bacterial load, the reader is directed to section 6.4.2.

The SDS sample was also used as a tool to determine optimal methods and settings when performing contaminant removal using the Decontam R package tools. We recognize also that because the SDS is a less complex sample than the procedural samples, findings may not be directly transferable. On comparison of decontamination strategies, we could have also included other approaches, such as contaminant identification using the neutral community model [19].

10.2.3 Paper III

Paper III was aimed at (dis)proving our hypothesis that contaminating bacterial DNA introduced during laboratory processing steps would render samples processed through the longer 2-step PCR protocol (setups 1 and 2) more vulnerable to laboratory contamination than when processed through the 1-step PCR protocol (setup 3). We also wished to explore differences that may result from targeting the 16S rRNA gene V3 V4 region versus the V4 region.

By processing the same DNA extracts through each setup it was possible to mitigate potential bias from differences in contamination introduced by the DNA extraction kit. This was important as we have previously shown that the DNA extraction kit is a main source of contamination in our experiments (paper II).

The differences in sequencing output generated by processing samples through each library preparation setup was based on four separate analyses: i) comparison of sequences and ASV retained at each bioinformatic processing step, ii) comparison of

a mock community sample processed through each setup, iii) comparison of contamination profiles for each setup by examination of ASVs recovered in NCS, and finally iv) comparison of community descriptions obtained from procedural samples before and after the removal of contaminating ASVs using the Decontam strategy.

Bioinformatics processing steps

The comparison of the three different setups began with examination of the number of sequences and ASVs retained at each step of the bioinformatic pipeline. We expected that the removal of error and bias at each step would result in increasingly more similar datasets. Indeed, we did observe that the number of sequences and ASVs became more similar; however at the end of the pipeline differences remained with the number of sequences/ASVs still decreasing in the order: setup 1 > setup 2 > setup 3. A closer examination of the number of sequence/ASVs retained at each step provided insight into the differences in the raw sequencing output generated when processing samples through each setup. A discussion on the most telling observations and our interpretation follows - namely that resulting from the additional filtering of small ASVs and that from the additional chimera removal step.

The perhaps most interesting observation, was that the additional filtering step for removal of low abundant ASVs (i.e. the “Bokulich method”), led to the greatest reduction in the number of ASVs across all three setups. The impact was greatest for setups 1 and 2, both of which were based on the 2-step PCR protocol. These observations were in accordance with our prediction that small ASVs likely represent low abundant contaminating sequences, and that samples processed through the longer 2-step PCR protocol would be more susceptible to contamination than samples processed through the shorter 1-step PCR protocol. The greatest proportion of these “contaminating” small ASVs were observed in setup 1, and this was substantially greater than for setup 2, for which was based on the same number of PCR steps. We recognized that our observations could be a result of setup 1 samples being spread across four separate sequencing runs; this because contamination may vary across sequencing runs and ultimately may have led to the observed higher diversity of

sequences representative of contaminants. Analyses of MC samples included on each of four sequencing runs for setup 1, confirmed that contamination profiles differed across the four sequencing runs for setup 1. When reanalyzing the data on the subset of samples that were sequenced on the same sequencing run, this was confirmed, since the differences across setups were now lower.

Also interesting was that the additional round of chimera filtering, led to an additional loss of sequences and ASVs. The additional chimera removal step had greatest impact on sequence data derived from setup 1. Interpretation of this observation is not straightforward as there again are multiple possible explanations. The observation may indicate that the proportion of chimeras is truly greater when processing samples through setup 1. This would be in line with studies that have indicated that longer target amplicons will be more inclined to form chimeras, and multimeras. On the other hand, it may reflect the fact that algorithms for chimera detection have more difficulty in identifying chimeras, the shorter the sequence. Thus there is the possibility that chimera removal may not be as effective for setups 2 and 3 targeting the shorter V4 region.

Protocol effects on mock community

A mock community (MC) sample of known bacterial composition was included on each sequencing run for the current paper III. In general, MC samples are used to answer one central question – does my protocol generate data that accurately represents the sampled bacterial community? Although not a perfect representation of a natural sample (i.e. lower complexity), the MC is a valuable tool for which bias introduced during library preparation and sequencing can be estimated. The MC can take form as a mixture of bacterial cells or a mixture of their genomic DNAs. In the form of a bacterial cell mixture, the MC is processed through all steps of the amplicon-based marker gene sequencing workflow from DNA extraction to sequencing. In the form of a mixture of genomic DNA, the DNA extraction step can be omitted, allowing assessment of downstream steps (PCR and sequencing) without the influence of DNA extraction as a confounding factor. The MC can be “staggered”

or “even” in composition, meaning that the different types of bacteria are represented at different or equal concentrations, respectively.

For the current paper choice of MC fell on mock community HM-783D (Bei Resources) consisting of genomic DNA from 20 different bacterial species at varying concentrations in the range of 1000 to 1000000 rRNA operon counts per species. A MC of genomic DNA was chosen, rather than a MC of bacterial cells, because we wished to validate laboratory steps post-DNA extraction. In addition, the use of genomic DNA was favorable because the number of rRNA copies for each species was known. Analyses were therefore not influenced by variation in copy numbers. A staggered community was chosen in order to assess the degree to which both low and high abundant MC members were recovered. For setup 1, the MC was included on all four sequencing runs enabling also the assessment of reproducibility/reliability for this setup. For setups 2 and 3, for which only one sequencing run was performed for each, it was not possible to determine reproducibility/reliability. This may represent a weakness in our interpretation of results from MC sequencing across the three setups.

We found that the three setups were equally efficient at recovering the high abundant members. For low abundant members, recovery varied across setups. Recall that the MC was processed on each of the four sequencing runs for setup 1. While the first run recovered all MC members, *Bacteroides* was missing on run II, and *Actinomyces* was missing on runs III and IV. For setup 2, all MC members were recovered. For setup 3, three genera were not recovered including *Propionibacterium*, *Actinomyces*, and *Enterococcus*. It thus appeared that the recovery of low abundant genera was an unreliable event across both 1- and 2-step PCR protocols (as demonstrated from output in both setups 1 and 3). However, setup 3 was most impacted as multiple genera were missing from data generated from the same sequencing run. One possible explanation for these observations was that the degenerate V4 primers in setup 3, were suboptimal matches to the sequences from these bacteria (i.e. primer bias). However, this was quickly dismissed as the setup 2 primers contained the same sequences for targeting the V4 region, and these same bacteria were recovered in data obtained from setup 2. Berry *et al.* [42], who also compared 1 and 2-step PCR

protocols, found in accordance with our findings, lower diversity for samples processed through a one step PCR protocol. They suggested another more likely explanation for observed differences – that the additional sequences required on primers used in 1-step PCR protocols (marker gene targeting sequence, index and adapter sequences) may interfere with primer-template interactions during PCR.

Based on these results, we concluded that the 1-step PCR protocol may be less apt for detection of rare taxa. This was in accordance with the analyses of bioinformatics processing steps, for which we observed a lower total ASV count when following setup 3 compared to setups 1 and 2. Our analyses of the MC however, forced us to rethink our previous conclusion. Recall that based on the analyses of bioinformatics processing steps, we interpreted the lower total number of ASVs in setup 3, and the more excessive removal of small ASVs for setups 1 and 2, following the “Bokulich model”, as an indication that samples were less prone to contamination when processing through a shorter 1-step PCR protocol. By analyses of the MC, we however learn that differences in contamination levels may not be directly linked to the length of the laboratory protocol. Rather, this may be a result of differences in PCR primer structure, for which primers required for the 1-step PCR protocol are less able to pick up on low abundant taxa – be it derived from contaminating bacteria or true members of the sampled community.

Protocol effects on contamination profiles

Our working hypothesis linked any observed differences in sequencing output to differences in susceptibility to laboratory contamination. We therefore proceeded with an examination of negative control samples – the procedural NCS and PCR water samples.

The procedural NCS were across all three setups dominated by ASVs belonging to the family *Enterobacteriaceae*. At ASV level we found that setup 1 was dominated by 3 different ASVs that all mapped to the genus *Gluconacetobacter* (Family *Enterobacteriaceae*). For setup 2, a single ASV was mapped to *Enterobacteriaceae* at no deeper depth than family level; however the relative abundance of this ASV was

the same as the cumulative abundance of the three ASVs mapping to *Gluconacetobacter* in setup 1- suggesting a common bacterial origin. The same ASV was also observed in setup 3. Overall this told us that *Enterobacteriaceae* dominated all samples and was likely introduced during DNA extraction steps. This was also in accordance with findings from the previous paper II, for which analyses of top 20 OTUs were conducted on the same samples using QIIME1 and the OTU based sequence clustering approach. Also noteworthy, is that the similarity of the results for the current paper (processed using QIIME2 and denoising to ASVs) to that generated for the previous paper II, provides internal validity with regards to bioinformatics processing steps and findings reported for both papers.

Perhaps the most interesting finding in our analyses of the top 20 NCS across setups, was the observation of a second ASV mapping to *Enterobacteriaceae* in setup 3. The ASV was in turn nearly undetectable in setup 2, having been observed in just 2/23 NCS samples (relative abundances of just 0.16% and 0.26%). Using NCBI blastn, this second *Enterobacteriaceae* ASV was identified as *Escherichia coli*. To further grasp the origin of this ASV, we looked to the PCR water samples. Indeed, we found that the ASV dominated the PCR water sample from setup 3 and was absent in the PCR water sample from setup 2. This told us that the ASV was a contaminant introduced during library preparation steps (i.e. post DNA extraction) in setup 3. To further back up our findings, we also searched for the ASV in the MC samples. We found the same ASV present in both MC samples regardless of sequencing setup 2 or 3 – this was as expected as *Escherichia coli* is a high abundant member of the MC with an expected relative abundance of 21.91%. In accordance with our finding that the *Escherichia* ASV also behaves as a contaminant in setup 3, we observed increased levels of the ASV for setup 3 compared to the expected and that observed in setup 2– thus in other words for setup 3, the ASV represented both a true member of the sampled community, but also a contaminant. In addition, as will be discussed in the subsequent section, the same *Escherichia* ASV was also observed in the low biomass airway samples (PSB, PBAL) processed through setup 3 (and not setup 2). Thus, several lines of evidence show that the *Escherichia* ASV is a result of a contaminant introduced from a library preparation reagent in setup 3. In summary,

this included the observation of the ASV in both procedural NCS and PCR water and also the observation of elevated levels of the ASV in the mock community. While internal contamination - such as well to well carryover in the direction MC to PCR water sample may be theoretically be possible, the sum of our observations suggest this to be an unlikely alternative explanation. For such an internal contamination event to occur, we would also expect to observe other MC members in the PCR water sample (for instance the *Rhodobacter*); this was not observed when not taking into account potential members that overlap between airway and MC samples (e.g. *Streptococcus*, *Staphylococcus*) as this may indicate contamination from airway samples.

Protocol effects on procedural samples

Having compared NCS across setups, we were next interested in finding out whether differences in contamination would influence our interpretation of the lower airway microbiome. We were particularly interested in tracing the *Enterobacteriaceae* ASVs that were found to dominate the contamination profiles (NCS) for all three setups as described above.

In accordance with expected patterns of contamination (section 6.4.1), we found that the proportion of *Enterobacteriaceae* was highest in the lower airway samples (PSB>PBAL) and nearly undetectable in OW samples. Also, consistent across setups was the observation that higher levels of *Enterobacteriaceae* were observed in PSB samples compared to PBAL samples. This is likely explained by differences in the bacterial load between the two sample types, rendering PSB samples more vulnerable to contamination. Recall that in paper II, it was established that the mean sample bacterial load decreased in order OW > PBAL1 > PSB > PBAL2. However, less sample volume was used as input to DNA extraction for PSB samples than PBAL samples (450 μ l PSB vs 1800 μ l PBAL), likely explaining the apparent increased impact of contamination on PSB samples. As expected from analyses of NCS, we also found that *Enterobacteriaceae* was found in greatest relative abundance in procedural samples in setup 3.

While the aforementioned observations were in line with expected patterns of contamination, the levels of *Enterobacteriaceae* were lower than the expected levels of contamination as estimated in paper II for setup 2 (section 10.2.2). We recognized that a significant proportion of contaminants were likely also represented by other taxa. For a more accurate assessment of contamination, comparison of samples was made before and after the removal of contaminants identified using the Decontam package tools. On analysis of unweighted UniFrac distances in PCoA space for setups 2 and 3, we found that high biomass OW samples clustered together regardless of setup, both before and after Decontam had been applied. The lower airway samples (PSB, PBAL) however separated according to setup 2 or 3. After removal of Decontam contaminants the overlap appeared to increase, but the observed separation according to setup was still apparent. We concluded that factors related to bacterial load, other than contamination was contributing to the observed protocol bias.

No other study on the airway microbiome has addressed the issue of protocol effects (1- vs 2-PCR steps) on the presentation of the airways. We shed light on an issue that needs to be investigated further in future studies. Particularly concerning is that our findings indicate that the similarity between upper and lower airway samples may be protocol dependent. Furthermore, we show that similar community descriptions obtained for upper airway samples should not be interpreted as evidence that datasets are comparable also for lower airway samples.

11. Conclusions

1. The bacterial composition of samples obtained by protected specimen brushes and protected bronchoalveolar lavage was less similar to oral wash samples than more unprotected sampling methods. Future investigators should take these findings into consideration, and take measures to prevent potential contamination from supraglottic regions.
2. Laboratory contamination was considerable in airway microbiome studies, and in particular the DNA extraction kits appeared to represent a major contamination source. However, bioinformatic strategies were able to correct for this, given availability of proper negative control samples.
3. A one-step PCR protocol yields results that differ taxonomically from a two-step PCR protocol. These differences are likely related to mechanisms in the PCR itself and not to contamination. Differences between V4 and a combined V3 V4 target amplicon are smaller and more likely related to taxonomic resolution.

12. Future perspectives

The field of lung microbiome research has up until now been characterized by an urge to rapidly publish data comparing healthy and diseased states. Few studies have however addressed validity and reliability of applied methods of sampling, laboratory processing and bioinformatics analyses. The findings in the current thesis underline the importance of addressing these issues in future studies.

While our findings have suggested that protected bronchoscopic sampling techniques may minimize the influence of oropharyngeal contamination, there is still a need for further investigation. Studies for which the upper airway representative sample is obtained from the supraglottic region (the last upper airway site that the bronchoscope passes on its entry to the lower airways) may for instance provide deeper insight into impact of bronchoscopic carryover. Furthermore, we need a deeper understanding of the impact of dilution effects when sampling and when deciding on the volume of sample that is further passed down the laboratory pipeline.

We found that differences in setups for library preparation for sequencing, related to the number of PCR steps (1- or 2-steps) led to different community descriptions for airway samples. Because differences in contamination levels alone could not explain these findings, we concluded that more research is needed to understand underlying mechanisms driving the observed protocol bias - these are likely related to the PCR. Besides impact of number of PCR steps, there is a need to investigate the impact of variation in PCR cycling conditions across studies. One study has for instance used touch-down PCR as a means to optimize their protocol [21]. The degree to which variations in PCR cycling may introduce bias is not known.

13. References

1. Barnes PJ. Immunology of asthma and chronic obstructive pulmonary disease. *Nat Rev Immunol.* 2008;8:183–92.
2. Stoller JK, Aboussouan LS. α 1-antitrypsin deficiency. *The Lancet.* 2005;365:2225–36.
3. Svanes \emptyset , Skorge TD, Johannessen A, Bertelsen RJ, Bråtveit M, Forsberg B, et al. Respiratory Health in Cleaners in Northern Europe: Is Susceptibility Established in Early Life? *PLOS ONE.* 2015;10:e0131959.
4. Hilty M, Burke C, Pedro H, Cardenas P, Bush A, Bossley C, et al. Disordered microbial communities in asthmatic airways. *PLoS ONE.* 2010;5:e8578.
5. Marchesi JR, Ravel J. The vocabulary of microbiome research: a proposal. *Microbiome.* 2015;3. doi:10.1186/s40168-015-0094-5.
6. Charlson ES, Bittinger K, Haas AR, Fitzgerald AS, Frank I, Yadav A, et al. Topographical Continuity of Bacterial Populations in the Healthy Human Respiratory Tract. *Am J Respir Crit Care Med.* 2011;184:957–63.
7. Bassis CM, Erb-Downward JR, Dickson RP, Freeman CM, Schmidt TM, Young VB, et al. Analysis of the Upper Respiratory Tract Microbiotas as the Source of the Lung and Gastric Microbiotas in Healthy Individuals. *mBio.* 2015;6:e00037-15.
8. Erb-Downward JR, Thompson DL, Han MK, Freeman CM, McCloskey L, Schmidt LA, et al. Analysis of the Lung Microbiome in the “Healthy” Smoker and in COPD. *PLoS One.* 2011;6. doi:10.1371/journal.pone.0016384.
9. Paggiaro L of the WGPL, Group M of the W, Chanez P, Holz O, Ind PW, Djukanović R, et al. Sputum induction. *European Respiratory Journal.* 2002;20 37 suppl:3s–8s.
10. Tangedal S, Aanerud M, Grønseth R, Drengenes C, Wiker HG, Bakke PS, et al.

Comparing microbiota profiles in induced and spontaneous sputum samples in COPD patients. *Respir Res.* 2017;18:164.

11. Gershman NH, Liu H, Wong HH, Liu JT, Fahy JV. Fractional analysis of sequential induced sputum samples during sputum induction: evidence that different lung compartments are sampled at different time points. *J Allergy Clin Immunol.* 1999;104 2 Pt 1:322–8.

12. Dickson RP, Erb-Downward JR, Freeman CM, McCloskey L, Beck JM, Huffnagle GB, et al. Spatial Variation in the Healthy Human Lung Microbiome and the Adapted Island Model of Lung Biogeography. *Ann Am Thorac Soc.* 2015;12:821–30.

13. Zemanick ET, Wagner BD, Robertson CE, Stevens MJ, Szeffler SJ, Accurso FJ, et al. Assessment of airway microbiota and inflammation in cystic fibrosis using multiple sampling methods. *Ann Am Thorac Soc.* 2015;12:221–9.

14. Dickson RP, Erb-Downward JR, Freeman CM, McCloskey L, Falkowski NR, Huffnagle GB, et al. Bacterial Topography of the Healthy Human Lower Respiratory Tract. *mBio.* 2017;8:e02287-16.

15. Willner D, Haynes MR, Furlan M, Schmieder R, Lim YW, Rainey PB, et al. Spatial distribution of microbial communities in the cystic fibrosis lung. *ISME J.* 2012;6:471–4.

16. Charlson ES, Chen J, Custers-Allen R, Bittinger K, Li H, Sinha R, et al. Disordered Microbial Communities in the Upper Respiratory Tract of Cigarette Smokers. *PLOS ONE.* 2010;5:e15216.

17. Consortium THMP, Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, et al. Structure, function and diversity of the healthy human microbiome. *Nature.* 2012;486:207.

18. Morris A, Beck JM, Schloss PD, Campbell TB, Crothers K, Curtis JL, et al. Comparison of the Respiratory Microbiome in Healthy Nonsmokers and Smokers.

Am J Respir Crit Care Med. 2013;187:1067–75.

19. Beck JM, Schloss PD, Venkataraman A, Twigg H, Jablonski KA, Bushman FD, et al. Multicenter Comparison of Lung and Oral Microbiomes of HIV-infected and HIV-uninfected Individuals. *Am J Respir Crit Care Med*. 2015;192:1335–44.

20. Segal LN, Alekseyenko AV, Clemente JC, Kulkarni R, Wu B, Chen H, et al. Enrichment of lung microbiome with supraglottic taxa is associated with increased pulmonary inflammation. *Microbiome*. 2013;1:19.

21. Dickson RP, Erb-Downward JR, Freeman CM, Walker N, Scales BS, Beck JM, et al. Changes in the Lung Microbiome following Lung Transplantation Include the Emergence of Two Distinct *Pseudomonas* Species with Distinct Clinical Associations. *PLoS One*. 2014;9. doi:10.1371/journal.pone.0097214.

22. Wimberley N, Faling LJ, Bartlett JG. A Fiberoptic Bronchoscopy Technique to Obtain Uncontaminated Lower Airway Secretions for Bacterial Culture. *Am Rev Respir Dis*. 1979;119:337–43.

23. Grønseth R, Haaland I, Wiker HG, Martinsen EMH, Leiten EO, Husebø G, et al. The Bergen COPD microbiome study (MicroCOPD): rationale, design, and initial experiences. *Eur Clin Respir J*. 2014;1.

24. Qvarfordt I, Riise GC, Andersson BA, Larsson S. Lower airway bacterial colonization in asymptomatic smokers and smokers with chronic bronchitis and recurrent exacerbations. *Respir Med*. 2000;94:881–7.

25. Segal LN, Clemente JC, Tsay J-CJ, Koralov SB, Keller BC, Wu BG, et al. Enrichment of the lung microbiome with oral taxa is associated with lung inflammation of a Th17 phenotype. *Nat Microbiol*. 2016;1:16031.

26. Schloss PD, Gevers D, Westcott SL. Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies. *PLOS ONE*. 2011;6:e27310.

27. Hogan DA, Willger SD, Dolben EL, Hampton TH, Stanton BA, Morrison HG, et

-
- al. Analysis of Lung Microbiota in Bronchoalveolar Lavage, Protected Brush and Sputum Samples from Subjects with Mild-To-Moderate Cystic Fibrosis Lung Disease. *PLOS ONE*. 2016;11:e0149998.
28. Shehadul Islam M, Aryasomayajula A, Selvaganapathy PR. A Review on Macroscale and Microscale Cell Lysis Methods. *Micromachines (Basel)*. 2017;8. doi:10.3390/mi8030083.
29. Yuan S, Cohen DB, Ravel J, Abdo Z, Forney LJ. Evaluation of methods for the extraction and purification of DNA from the human microbiome. *PLoS ONE*. 2012;7:e33865.
30. Brott AS, Clarke AJ. Peptidoglycan O-Acetylation as a Virulence Factor: Its Effect on Lysozyme in the Innate Immune System. *Antibiotics*. 2019;8:94.
31. Pääbo S, Irwin DM, Wilson AC. DNA damage promotes jumping between templates during enzymatic amplification. *J Biol Chem*. 1990;265:4718–21.
32. Biesbroek G, Sanders EAM, Roeselers G, Wang X, Caspers MPM, Trzeciński K, et al. Deep Sequencing Analyses of Low Density Microbial Communities: Working at the Boundary of Accurate Microbiota Detection. *PLOS ONE*. 2012;7:e32942.
33. Woese CR, Magrum LJ, Gupta R, Siegel RB, Stahl DA, Kop J, et al. Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence. *Nucleic Acids Res*. 1980;8:2275–93.
34. Woese CR. Bacterial evolution. *Microbiol Rev*. 1987;51:221–71.
35. Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ. At Least 1 in 20 16S rRNA Sequence Records Currently Held in Public Repositories Is Estimated To Contain Substantial Anomalies. *Appl Environ Microbiol*. 2005;71:7724–36.
36. Baker GC, Smith JJ, Cowan DA. Review and re-analysis of domain-specific 16S primers. *J Microbiol Methods*. 2003;55:541–55.
37. Stackebrandt E, Goebel BM. Taxonomic Note: A Place for DNA-DNA

Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *International Journal of Systematic and Evolutionary Microbiology*. 1994;44:846–9.

38. Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R. Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res*. 2007;35:e120.

39. Binladen J, Gilbert MTP, Bollback JP, Panitz F, Bendixen C, Nielsen R, et al. The Use of Coded PCR Primers Enables High-Throughput Sequencing of Multiple Homolog Amplification Products by 454 Parallel Sequencing. *PLoS One*. 2007;2. doi:10.1371/journal.pone.0000197.

40. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R. Error-correcting barcoded primers allow hundreds of samples to be pyrosequenced in multiplex. *Nat Methods*. 2008;5:235–7.

41. Polz MF, Cavanaugh CM. Bias in Template-to-Product Ratios in Multitemplate PCR. *Appl Environ Microbiol*. 1998;64:3724–30.

42. Berry D, Mahfoudh KB, Wagner M, Loy A. Barcoded Primers Used in Multiplex Amplicon Pyrosequencing Bias Amplification. *Appl Environ Microbiol*. 2011;77:7846–9.

43. Klappenbach JA, Saxman PR, Cole JR, Schmidt TM. rrndb: the Ribosomal RNA Operon Copy Number Database. *Nucleic Acids Res*. 2001;29:181–4.

44. Wang Y, Zhang Z, Ramanan N. The actinomycete *Thermobispora bispora* contains two distinct types of transcriptionally active 16S rRNA genes. *J Bacteriol*. 1997;179:3270–6.

45. Wang GC, Wang Y. Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. *Appl Environ Microbiol*. 1997;63:4645–50.

-
46. Wang GCY, Wang Y. The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species. *Microbiology*. 1996;142:1107–14.
47. Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, et al. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res*. 2011;21:494–504.
48. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res*. 2015;43:e37.
49. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. *Appl Environ Microbiol*. 2013;79:5112–20.
50. Kircher M, Sawyer S, Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res*. 2012;40:e3.
51. Wright ES, Vetsigian KH. Quality filtering of Illumina index reads mitigates sample cross-talk. *BMC Genomics*. 2016;17. doi:10.1186/s12864-016-3217-x.
52. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, et al. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat Methods*. 2013;10:57–9.
53. Huse SM, Welch DM, Morrison HG, Sogin ML. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental Microbiology*. 2010;12:1889–98.
54. Quality Scores for Next-Generation Sequencing. :2.
55. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon

Sequence Data on the MiSeq Illumina Sequencing Platform. *Appl Environ Microbiol.* 2013;79:5112–20.

56. Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics.* 2012;13:31.

57. Edgar RC, Flyvbjerg H. Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics.* 2015;31:3476–82.

58. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal.* 2017;11:2639–43.

59. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods.* 2016;13:581.

60. Huber T, Faulkner G, Hugenholtz P. Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics.* 2004;20:2317–9.

61. Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ. Removing Noise From Pyrosequenced Amplicons. *BMC Bioinformatics.* 2011;12:38.

62. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics.* 2011;27:2194–200.

63. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology.* 2014;12:87.

64. Dickson RP, Erb-Downward JR, Prescott HC, Martinez FJ, Curtis JL, Lama VN, et al. Cell-associated bacteria in the human lung microbiome. *Microbiome.* 2014;2:28.

65. Twigg *Homer L., Nelson DE, Day RB, Gregory RL, Dong Q, Rong R, et al. Comparison of Whole and Acellular Bronchoalveolar Lavage to Oral Wash

Microbiomes. Should Acellular Bronchoalveolar Lavage Be the Standard? *Ann Am Thorac Soc.* 2014;11 Suppl 1:S72–3.

66. Twigg HL, Knox KS, Zhou J, Crothers KA, Nelson DE, Toh E, et al. Effect of Advanced HIV Infection on the Respiratory Microbiome. *Am J Respir Crit Care Med.* 2016;194:226–35.

67. Lozupone C, Cota-Gomez A, Palmer BE, Linderman DJ, Charlson ES, Sodergren E, et al. Widespread colonization of the lung by *Tropheryma whippelii* in HIV infection. *Am J Respir Crit Care Med.* 2013;187:1110–7.

68. Sørheim I-C, Johannessen A, Grydeland TB, Omenaas ER, Gulsvik A, Bakke PS. Case-control studies on risk factors for chronic obstructive pulmonary disease: how does the sampling of the cases and controls affect the results? *Clin Respir J.* 2010;4:89–96.

69. Eagan TM, Aukrust P, Bakke PS, Damås JK, Skorge TD, Hardie JA, et al. Systemic mannose-binding lectin is not associated with Chronic Obstructive Pulmonary Disease. *Respir Med.* 2010;104:283–90.

70. Eagan TML, Aukrust P, Ueland T, Hardie JA, Johannessen A, Mollnes TE, et al. Body composition and plasma levels of inflammatory biomarkers in COPD. *Eur Respir J.* 2010;36:1027–33.

71. Eagan TM, Ueland T, Wagner PD, Hardie JA, Mollnes TE, Damås JK, et al. Systemic inflammatory markers in COPD: results from the Bergen COPD Cohort Study. *The European respiratory journal.* 2010.

72. Gulsvik A, Tosteson T, Bakke P, Humerfelt S, Weiss ST, Speizer FE. Expiratory and inspiratory forced vital capacity and one-second forced volume in asymptomatic never-smokers in Norway. *Clin Physiol.* 2001;21:648–60.

73. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* 2013;41:e1.

74. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci USA*. 2011;108 Suppl 1:4516–22.
75. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*. 2010;7:335–6.
76. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*. 2016;4. doi:10.7717/peerj.2584.
77. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 2010;26:2460–1.
78. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J*. 2012;6:610–8.
79. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Appl Environ Microbiol*. 2007;73:5261–7.
80. Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics*. 2010;26:266–7.
81. Price MN, Dehal PS, Arkin AP. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Mol Biol Evol*. 2009;26:1641–50.
82. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*. 2018;6:226.
83. Davis NM, Proctor D, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and

metagenomics data. *bioRxiv*. 2017;:221499.

84. McMurdie PJ, Holmes S. Waste Not, Want Not: Why Rarefying Microbiome Data is Inadmissible. *PLoS Computational Biology*. 2014;10:e1003531.

85. Venkataraman A, Bassis CM, Beck JM, Young VB, Curtis JL, Huffnagle GB, et al. Application of a Neutral Community Model To Assess Structuring of the Human Lung Microbiome. *mBio*. 2015;6:e02284-14.

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

87. Pragman AA, Kim HB, Reilly CS, Wendt C, Isaacson RE. The lung microbiome in moderate and severe chronic obstructive pulmonary disease. *PLoS ONE*. 2012;7:e47305.

88. Huang YJ, Nelson CE, Brodie EL, DeSantis TZ, Baek MS, Liu J, et al. Airway Microbiota and Bronchial Hyperresponsiveness in Patients with Sub-optimally Controlled Asthma. *J Allergy Clin Immunol*. 2011;127:372-381.e3.

89. Cabrera-Rubio R, Garcia-Núñez M, Setó L, Antó JM, Moya A, Monsó E, et al. Microbiome Diversity in the Bronchial Tracts of Patients with Chronic Obstructive Pulmonary Disease. *J Clin Microbiol*. 2012;50:3562–8.

90. Willner DL, Hugenholtz P, Yerkovich ST, Tan ME, Daly JN, Lachner N, et al. Reestablishment of Recipient-associated Microbiota in the Lung Allograft Is Linked to Reduced Risk of Bronchiolitis Obliterans Syndrome. *Am J Respir Crit Care Med*. 2013;187:640–7.

91. Borewicz K, Pragman AA, Kim HB, Hertz M, Wendt C, Isaacson RE. Longitudinal Analysis of the Lung Microbiome in Lung Transplantation. *FEMS Microbiol Lett*. 2013;339:57–65.

92. Cribbs SK, Uppal K, Li S, Jones DP, Huang L, Tipton L, et al. Correlation of the

lung microbiota with metabolic profiles in bronchoalveolar lavage fluid in HIV infection. *Microbiome*. 2016;4. <http://d-scholarship.pitt.edu/28907/>. Accessed 9 Aug 2020.

93. Molyneaux PL, Cox MJ, Willis-Owen SAG, Mallia P, Russell KE, Russell A-M, et al. The Role of Bacteria in the Pathogenesis and Progression of Idiopathic Pulmonary Fibrosis. *Am J Respir Crit Care Med*. 2014;190:906–13.

94. Garzoni C, Brugger SD, Qi W, Wasmer S, Cusini A, Dumont P, et al. Microbial communities in the respiratory tract of patients with interstitial lung disease. *Thorax*. 2013;68:1150–6.

95. Zakharkina T, Heinzl E, Koczulla RA, Greulich T, Rentz K, Pauling JK, et al. Analysis of the Airway Microbiota of Healthy Individuals and Patients with Chronic Obstructive Pulmonary Disease by T-RFLP and Clone Sequencing. *PLoS One*. 2013;8. doi:10.1371/journal.pone.0068302.

96. Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathog*. 2016;8. doi:10.1186/s13099-016-0103-7.

97. Glassing A, Dowd SE, Galandiuk S, Davis B, Jordan JR, Chiodini RJ. Changes in 16s RNA Gene Microbial Community Profiling by Concentration of Prokaryotic DNA. *Journal of Microbiological Methods*. 2015;119:239–42.

98. Dickson RP, Erb-Downward JR, Prescott HC, Martinez FJ, Curtis JL, Lama VN, et al. Analysis of Culture-Dependent versus Culture-Independent Techniques for Identification of Bacteria in Clinically Obtained Bronchoalveolar Lavage Fluid. *J Clin Microbiol*. 2014;52:3605–13.

99. Pragman AA, Lyu T, Baller JA, Gould TJ, Kelly RF, Reilly CS, et al. The lung tissue microbiota of mild and moderate chronic obstructive pulmonary disease. *Microbiome*. 2018;6:7.

-
100. Youssef N, Sheik CS, Krumholz LR, Najjar FZ, Roe BA, Elshahed MS. Comparison of Species Richness Estimates Obtained Using Nearly Complete Fragments and Simulated Pyrosequencing-Generated Fragments in 16S rRNA Gene-Based Environmental Surveys. *Appl Environ Microbiol.* 2009;75:5227–36.
101. Minich JJ, Sanders JG, Amir A, Humphrey G, Gilbert JA, Knight R. Quantifying and Understanding Well-to-Well Contamination in Microbiome Research. *mSystems.* 2019;4. doi:10.1128/mSystems.00186-19.
102. Navas-Molina JA, Peralta-Sánchez JM, González A, McMurdie PJ, Vázquez-Baeza Y, Xu Z, et al. Advancing our understanding of the human microbiome using QIIME. *Methods Enzymol.* 2013;531:371–444.
103. Marsh RL, Kaestli M, Chang AB, Binks MJ, Pope CE, Hoffman LR, et al. The microbiota in bronchoalveolar lavage from young children with chronic lung disease includes taxa present in both the oropharynx and nasopharynx. *Microbiome.* 2016;4. doi:10.1186/s40168-016-0182-1.
104. Jiang L, Amir A, Morton JT, Heller R, Arias-Castro E, Knight R. Discrete False-Discovery Rate Improves Identification of Differentially Abundant Microbes. *mSystems.* 2017;2. doi:10.1128/mSystems.00092-17.
105. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome.* 2017;5:27.
106. Lozupone C, Knight R. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Appl Environ Microbiol.* 2005;71:8228–35.

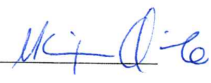
14. Papers and supplementary material

**Errata for
Methodological issues in airway microbiome studies**

Christine Drengenes



Thesis for the degree philosophiae doctor (PhD)
at the University of Bergen

17.11.20 Christine Drengenes 18.11.20 

(date and sign. of candidate) (date and sign. of faculty)

Errata

All pages	Change in page numbers
Pages 4	Missing sections in Table of contents: added “8.8 Statistical analyses”, “8.9 Ethics”, “10.2.3 Paper III”, “11. Conclusions”
Page 24	Missing word: “the” added
Page 34	Change from upper to lower case: “filtering”
Page 45	Change from upper to lower case “design”
Page 57	Error in subtitle number: “9.8 Statistical Analyses” – corrected to “8.8 Statistical analyses”.
Page 65	Error in number of participants. “With over 300 participants...” corrected to “With 249 participants”.
Page 67	Remove word “study”
Page 67	Change from upper to lower case: “samples”
Page 84	Misspelling: “over represented” corrected to “overrepresented”
Page 85	“beta- diversity” changed to “beta diversity”
Page 94	Missing subtitle: “Paper III” – corrected to “ 10.2.3 Paper III”
Page 102	Wrong line spacing: line spacing of 2.0 – corrected to 1.5.
Page 102	Misspelling: “Broncho alveolar lavage” – corrected to “bronchoalveolar lavage”
Page 102	Wrong wording: “The bacterial composition of samples obtained by protected specimen brushes and protected broncho alveolar lavage was <u>more</u> similar to oral wash samples than more unprotected sampling methods.” – corrected to “The bacterial composition of samples obtained by protected specimen brushes and protected bronchoalveolar lavage was <u>less</u> similar to oral wash samples than more unprotected sampling methods.”



Protected sampling is preferable in bronchoscopic studies of the airway microbiome

Rune Grønseth¹, Christine Drengenes^{1,2}, Harald G. Wiker^{2,3}, Solveig Tangedal^{1,2}, Yaxin Xue⁴, Gunnar Reksten Husebø^{1,2}, Øistein Svanes^{1,2}, Sverre Lehmann^{1,2}, Marit Aardal¹, Tuyen Hoang², Tharmini Kalanathan¹, Einar Marius Hjellevad Martinsen², Elise Orvedal Leiten², Marianne Aanerud¹, Eli Nordeide^{1,2}, Ingvild Haaland^{1,2}, Inge Jonassen⁴, Per Bakke² and Tomas Eagan^{1,2}

Affiliations: ¹Dept of Thoracic Medicine, Haukeland University Hospital, Bergen, Norway. ²Dept of Clinical Science, Faculty of Medicine and Dentistry, University of Bergen, Bergen, Norway. ³Dept of Microbiology, Haukeland University Hospital, Bergen, Norway. ⁴Computational Biology Unit, Dept of Informatics, University of Bergen, Bergen, Norway.

Correspondence: Rune Grønseth, Dept of Thoracic Medicine, Haukeland University Hospital, Jonas Lies vei, Bergen 5021, Norway. E-mail: nielsenrune@me.com

ABSTRACT The aim was to evaluate susceptibility of oropharyngeal contamination with various bronchoscopic sampling techniques.

67 patients with obstructive lung disease and 58 control subjects underwent bronchoscopy with small-volume lavage (SVL) through the working channel, protected bronchoalveolar lavage (PBAL) and bilateral protected specimen brush (PSB) sampling. Subjects also provided an oral wash (OW) sample, and negative control samples were gathered for each bronchoscopy procedure. DNA encoding bacterial 16S ribosomal RNA was sequenced and bioinformatically processed to cluster into operational taxonomic units (OTU), assign taxonomy and obtain measures of diversity.

The proportion of Proteobacteria increased, whereas Firmicutes diminished in the order OW, SVL, PBAL, PSB ($p < 0.01$). The alpha-diversity decreased in the same order ($p < 0.01$). Also, beta-diversity varied by sampling method ($p < 0.01$), and visualisation of principal coordinates analyses indicated that differences in diversity were smaller between OW and SVL and OW and PBAL samples than for OW and the PSB samples. The order of sampling (left *versus* right first) did not influence alpha- or beta-diversity for PSB samples.

Studies of the airway microbiota need to address the potential for oropharyngeal contamination, and protected sampling might represent an acceptable measure to minimise this problem.



@ERSpublications

Protected bronchoscopic sampling is most suitable for identification of a distinct airway microbiome <http://ow.ly/q1ly30eqB9M>

Cite this article as: Grønseth R, Drengenes C, Wiker HG, *et al.* Protected sampling is preferable in bronchoscopic studies of the airway microbiome. *ERJ Open Res* 2017; 3: 00019-2017 [<https://doi.org/10.1183/23120541.00019-2017>].

This article has supplementary material available from openres.ersjournals.com

Received: Feb 16 2017 | Accepted after revision: June 21 2017

Support statement: The current work has been funded through unrestricted grants and fellowships from Helse Vest, GlaxoSmithKline, the Endowment of Timber Merchant A. Delphin and Wife through the Norwegian Medical Association and Bergen Medical Research Foundation. Funding information for this article has been deposited with the Crossref Funder Registry.

Conflict of interest: Disclosures can be found alongside this article at openres.ersjournals.com

Copyright ©ERS 2017. This article is open access and distributed under the terms of the Creative Commons Attribution Non-Commercial Licence 4.0.



Introduction

High-throughput sequencing has opened up a new window in microbial ecology, enabling the characterisation of microbial communities in biological compartments thought to be completely sterile only a few years ago. The implications for health and disease are widely unexplored, but are likely to be significant [1]. Recent studies have found compelling evidence for the lungs to have a distinct microbiome [2], providing a bacterial presence with which our immune system interacts [3, 4]. As almost all pulmonary diseases have a local inflammatory component, there is a possibility of a disrupted microbiome being integral to disease pathogenesis.

Thus, there is a current push to characterise the pulmonary microbiome, and its relation to different pulmonary diseases. However, sampling the pulmonary microbiome is difficult. Sputum is fraught with significant contamination from the oral cavity, and percutaneous sampling is unpractical with a high risk of complications like pneumothorax or bleeding. The emerging gold standard for sampling is bronchoscopy. But bronchoscopy also has its technical challenges, besides issues of discomfort, cost and sedation. The bronchoscope must pass through either the oral or nasal cavity in addition to the pharyngeal cavity, and might carry contaminants from the upper airways to the lower biomass compartment of the lower airways. Samples are collected through the same bronchoscope working channel through which fluid is suctioned up and out. The different modes of sampling (bronchoalveolar lavage (BAL) brushings, biopsies) might be carried through catheters, which may or may not have a wax-sealed tip to ensure sterility. Added to this is the conundrum caused by the constant influx of microbiota by microaspiration and inhalation that probably is responsible for maintenance and creation of a large fraction of the lung microbiome [5].

In 25 studies of the human lung microbiome sampling the airway microbiome by bronchoscopy of healthy subjects [2–4, 6–9] and patients with chronic obstructive pulmonary disease (COPD) [10–14], asthma [15, 16], interstitial lung disease [17, 18], cystic fibrosis (CF) [19], HIV [20–23] and lung-transplanted subjects [24–27]; only five used protected sterile brushes (PSB) to avoid contamination from the working channel [7, 8, 16, 19, 22]. Some authors reported that suction was not used prior to entering the trachea [2–4, 6–10, 20, 22], and three studies used separate bronchoscopes for anaesthesia and sampling of some or all participants [3, 4, 7]. No study performed bronchoalveolar lavage (BAL) through a protected catheter (protected BAL), and no study with more than 20 sampled subjects has compared protected with unprotected sampling methods.

In preparation for the analyses of a large, ongoing COPD microbiome study [28], we sought to reduce contamination as well as assess the performance of different sampling techniques. In the current paper we present analyses to examine the degree of oropharyngeal influence on the airway microbiome applying protected bronchoscopic sampling techniques. In addition we present an analysis on the effect of sampling the left or right lung first.

Material and methods

The design of the entire MicroCOPD study has been published previously [28]. The current analysis includes 58 control subjects, 64 subjects with COPD and three subjects with asthma. All participants were at least 35 years old and were recruited from previous longitudinal case–control studies in addition to a few volunteers [29]. Subjects had neither acute respiratory symptoms nor any reported use of antibiotics or oral corticosteroids within the last 14 days prior to bronchoscopy. Other inclusion/exclusion criteria are listed in the supplementary material.

The Regional Committee for Medical and Health Research Ethics approved the study (REK Nord, project number 2011/1307). All participants provided written informed consent.

All participants received at least 0.4 mg of salbutamol through a spacer before the bronchoscopy procedure. Flexible video-bronchoscopy was performed *via* the oral route in supine position. No suction was used prior to having entered the trachea. All subjects received local anaesthesia with lidocaine both before and during the procedure. All but 18 subjects received mild sedation (alfentanil) parenterally. Participants were monitored according to current guidelines, and were observed for at least 2 h after the procedure [30]. Six procedural samples, of which five were obtained during bronchoscopy, were analysed for each participant: oral wash (OW); three protected specimen brushes (PSBs) from the right lower lobe (right PSB) and three from the left upper lobe (left PSB); two 50-mL fractions of protected bronchoalveolar lavage of the right middle lobe (PBAL1 and PBAL2); and small-volume lavage (SVL) in the left upper lobe. In addition, we included negative control samples (NCSS) from the same bottle of phosphate-buffered saline that was used for the procedure of the corresponding individual. For 49 subjects, we examined the left lung before the right lung. BAL and SVL were always collected after obtaining PSB samples. Protected specimen brushes and protected bronchoalveolar lavage are illustrated in supplementary figures S1 and S2.

Bacterial DNA was extracted using a combination of enzymatic lysis with lysozyme, mutanolysin and lysostaphin, and mechanical lysis methods using the FastPrep-24 as described by the manufacturers of the FastDNA Spin Kit (MP Biomedicals, LLC, Solon, OH, USA).

Library preparation and sequencing of the V3-V4 region of the 16S rRNA gene was carried out according to the Illumina 16S Metagenomic Sequencing Library Preparation guide (Part no. 15044223 Rev. B). The V3-V4 region was PCR amplified (45 cycles) and prepared for a subsequent index PCR step using primers adapted from KLINDWORTH *et al.* [31] as follows. 16S amplicon PCR forward primer (overhang adaptor sequences are underlined): 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG. 16S amplicon PCR reverse primer (overhang adaptor sequences are underlined): 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC. The samples were pooled and prepared for 2×300 cycles of paired-end sequencing on the Illumina Miseq sequencing platform using reagents from the Miseq reagent kit v3 (Illumina Inc., San Diego, CA, USA).

The chosen bioinformatic pipeline was Quantitative Insights Into Microbial Ecology (QIIME, <http://qiime.org>) v1.9.1. After creating a library of joined reads, operational taxonomic units (OTUs) were picked at a 97% similarity threshold, small OTUs and OTUs seen in negative control samples were removed, taxonomy was assigned to the OTUs and a phylogenetic tree was constructed after alignment. We used the GreenGenes version 13.8 as reference database [32]. Further details on the bioinformatic procedures can be found in the supplementary material.

Differences in relative abundance of taxa were evaluated by applying a beta distribution and non-parametric trend tests. Alpha-diversity was evaluated using Faith's phylogenetic diversity (PD), or "PD wholetree". Beta-diversity was estimated with unweighted UniFrac and visualised by principal coordinates analyses (PCoA) [33]. Diversity analyses require a similar number of sequences in each sample, which was ensured by rarefaction. Statistical significance for alpha-diversity and beta-diversity between sampling methods was evaluated by Bonferroni-corrected Wilcoxon matched-pairs test in Stata version 13.2 (Statacorp, Texas, USA) and permutational ANOVA (permanova) tests in QIIME, respectively.

Results

Only three subjects had asthma: two men and one woman. The 64 COPD subjects were slightly older, included more men and had a larger tobacco-smoking burden than the 58 control subjects (table 1).

For each of the 125 participants, seven samples were sequenced (negative control sample, OW, right PSB, PBAL1, PBAL2, left PSB, SVL). A total of 12.5 million sequences were obtained from the six procedural samples after bioinformatics clean-up, as described in the methods section. For alpha- and beta-diversity, we rarefied our data at 1000 sequences.

Taxonomy

Figure 1 shows the taxonomic classification by sampling method at the phylum level. As the degree of protection from influence of oral environment increased, the proportion of Proteobacteria increased, whereas Firmicutes diminished ($p<0.01$). At the genus level all sample types were dominated by streptococci, but the mean proportion of the largest *Streptococcus* OTU showed the same declining pattern by sample type (OW 14.5%, SVL 13.6%, PBAL1 11.8%, PBAL2 11.3%, right PSB 8.6% and left PSB 5.4%; non-parametric trend test $p<0.001$).

TABLE 1 Characteristics of 125 subjects of the MicroCOPD study

	COPD	Asthma	Control
Subjects	64	3	58
Males	34 (53.1%)	2 (67.7%)	34 (58.6%)
Current smokers	15 (23.4%)	0	16 (27.6%)
Ex-smokers	48 (75.0%)	2 (67.7%)	35 (60.3%)
Never-smokers	1 (1.6%)	1 (33.3%)	7 (12.1%)
Smoking exposure pack-years	28.49±16.08	20.88±24.22	22.83±18.55
FEV₁ % predicted	56.83±16.30	88.31±11.37	100.71±11.00
Age years	68.73±7.23	64.41±9.1	64.89±8.43
Use of inhaled corticosteroids	44 (68.8%)	1 (33.3%)	1 (1.7%)

Data are presented as mean±SD unless otherwise stated. COPD: chronic obstructive pulmonary disease; FEV₁: forced expiratory volume in 1 s.

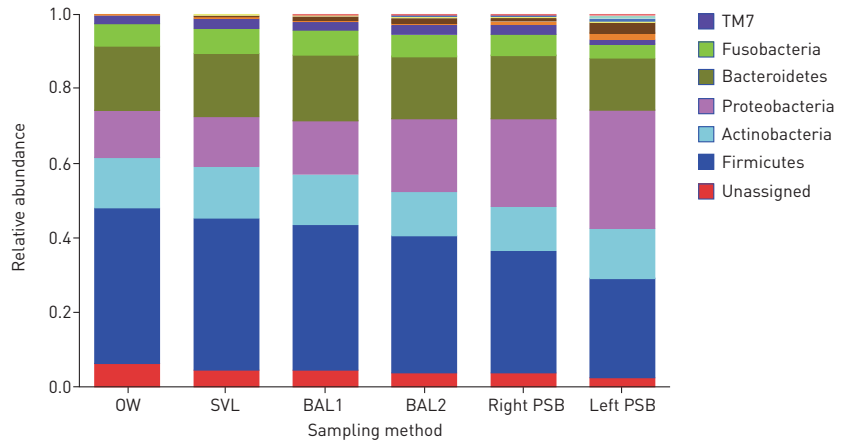


FIGURE 1 Mean taxonomic distribution at the phylum level, by sampling method, for all 125 individuals (unrarefied). OW: oral wash; SVL: small-volume lavage in the left upper lobe; BAL1: first fraction of protected bronchoalveolar lavage (BAL) from right middle lobe; BAL2: second fraction of protected BAL from right middle lobe; PSB: protected specimen brush from right lower lobe and left upper lobe. No legend for smallest phylae.

Alpha-diversity

Figure 2 shows a boxplot of the alpha-diversity metric, Faith’s phylogenetic diversity, by sampling method and by disease category, excluding the three asthma subjects. The phylogenetic diversity within a sample is an indication of richness as the diversity increases both when a higher number of different OTUs are present, and when the phylogenetic distance is larger within the phylogenetic tree (less genetically similar). Bonferroni-corrected Wilcoxon matched-pairs signed-ranks tests showed that the oral wash samples were more alpha-diverse than all other sampling methods ($p < 0.001$). The diversity was lower in COPD patients than controls, for most all sample types (figure 2). Importantly, the diversity decreased as the samples

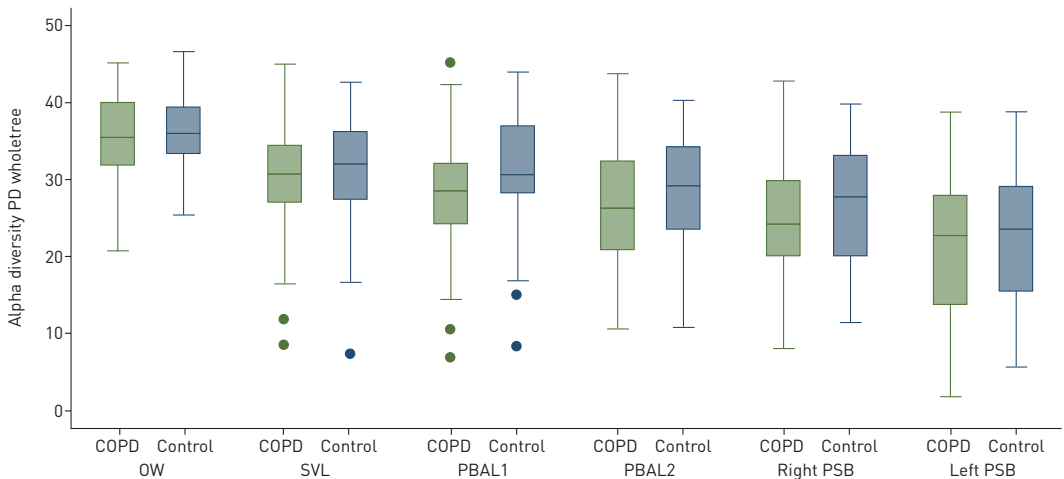


FIGURE 2 Box-plot of alpha-diversity measured by wholetree phylogenetic differences grouped according to sampling method and chronic obstructive pulmonary disease (COPD) status. Rarefied at 1000 sequences. OW: oral wash sample; SVL: small-volume lavage from left upper lobe; PBAL1: first fraction of protected bronchoalveolar lavage (BAL); PBAL2: second fraction of protected BAL; right PSB: protected specimen brush from right lower lobe; left PSB: protected specimen brush from left upper lobe.

were less exposed to potential oral and bronchoscope contamination (OW>SVL>PBAL1>PBAL2>rightPSB>leftPSB, non-parametric trend test $p<0.01$).

Beta-diversity

To compare between sample compositions (beta-diversity), we constructed principal coordinates analysis (PCoA) plots of unweighted UniFrac distances including all procedural samples. Figure 3 shows the PCoA plots for the oral wash *versus* each of the other sampling methods. Each dot represents a diversity measurement for one sample, and the OW sample is always shown in green. As can be seen, most respiratory tract samples clustered differently from the OW samples, but the visual impression is that the differences in diversity were smaller between OW and SVL and OW and PBAL samples than for OW and the PSB samples. Another way of comparing the beta-diversity was employed using a permanova test; estimating the beta-diversity between OW samples and each of the other sampling methods. This method tests to which degree the variation in a matrix of UniFrac distances can be explained by an imposed categorisation (*i.e.* sampling method). Overall permanova test confirmed that the beta-diversity differed by sampling method (pseudo F 8.73, $p=0.001$, 999 permutations). When the distance matrix was split according to the comparisons in figure 3, all were significant ($p<0.01$, permanova, corrected for multiple comparison), with the permanova pseudo F-statistic gradually increasing for the comparison of OW with SVL, PBAL1, PBAL2, right PSB and left PSB respectively, again indicating that PSB samples were more clearly separated from OW samples than SVL and PBAL.

Finally we investigated whether the order of sampling (left *versus* right lung first) influenced alpha- and beta-diversity in PSB samples. We found no significant difference in alpha- or beta-diversity for the right or the left PSBs as judged by phylogenetic diversity and unweighted UniFrac (supplementary figures S3 and S4).

Discussion

We have shown that protected BAL and protected brush samples differed more from oral wash samples than unprotected lavage through the bronchoscope working channel. Thus, unprotected sampling of the airway microbiome might convey an image of a microbiome that is more similar to the oral microbiome, than it would have been with protected sampling.

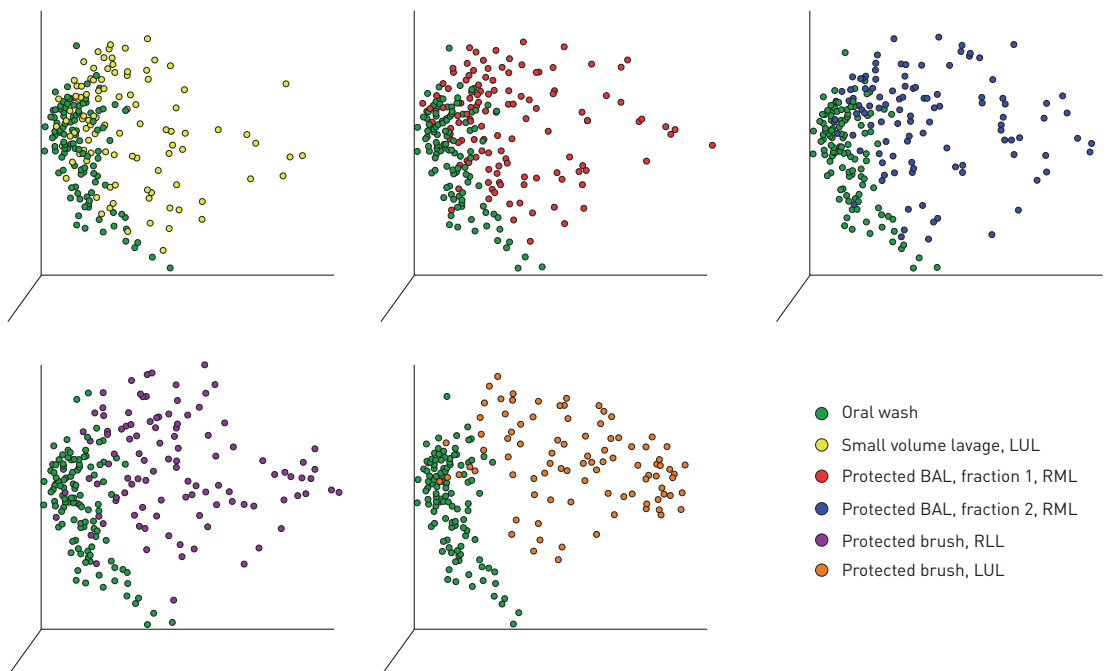


FIGURE 3 Principal coordinates analyses on unweighted UniFrac distance matrix comparing sampling methods in the MicroCOPD to oral wash samples. Rarefied at 1000 sequences. LUL: left upper lobe; BAL: bronchoalveolar lavage; RML: right middle lobe; RLL: right lower lobe.

To our best knowledge this is the first study that presents both protected brush *and* protected lavage sampling as compared with both the oral microbiome and unprotected sampling. With more than 120 examined subjects it is by today the largest single site bronchoscopy study of the lung microbiome.

As other authors we find evidence of a lung microbiome separated from the oral microbiome by a larger fraction of Proteobacteria and a proportionately lower fraction of Firmicutes [2, 8, 15, 20]. However, SEGAL and associates [3, 4] mainly found that the airway microbiome was characterised by enrichment from supraglottic areas of the respiratory tract, and in particular by *Prevotella* and *Veillonella* OTUs, which are Bacteroidetes and Firmicutes, respectively. They examined 49 subjects, with supraglottic brushes and BAL through the working channel, and observed that two clusters dominated airway samples: one dominated by OTUs present in negative control samples, and one dominated by OTUs present in supraglottic brushes. One interpretation might be that these two clusters represent two different modalities of contamination, the first one from laboratory procedures and the second from bronchoscopy carryover. SEGAL *et al.* argue that if it was bronchoscopic carry-over, they would have observed a dilutional effect when they compared a first BAL of the lingula, with the second BAL of the right middle lobe. However, this comparison was done for only 15 individuals, and anatomically one might expect lower biomass in the lingula than the right middle lobe.

Other authors have also investigated the possibility of bronchoscopic carryover. BASIS *et al.* examined oral wash samples of 12 subjects and compared them with a first BAL of the lingula and a second BAL of the right middle lobe [6]. They did not find any difference in quantitative PCR between the first and second BAL, and no difference in beta-diversity when comparing the OW with the two BALs. Their interpretation was that if there was significant carryover, there should have been observed some sort of dilutional effect. Nevertheless, the two sampled sites are separated by the carina, and the bronchoscope must be repositioned between sampling, and these two sites are indeed in different communication with the outside world, possibly leading to an *a priori* larger biomass in the right lung. Also, DICKSON *et al.* compared supraglottic brushes with PSB and BAL through the working channel [8]. In principal component analyses of beta-diversity they found no clustering by sample type, except that the supraglottic samples differed from the intrapulmonary sample communities. However, by performing unprotected BAL before PSB, residual BAL fluid might have affected the brush areas making them more similar to the BAL sample sites. Finally, 15 sampled subjects might not be sufficient to detect the differences we observed in the current study with more than 100 participants.

It is quite plausible that microbes migrate from the oropharyngeal cavity to the airways, generating a normal overlap between the oropharyngeal and airway microbiomes [5]. But as we have shown, co-existing sample contamination likely also is an issue. The oropharyngeal microbiome has a known large biomass, with a high diversity. By passing through this cavity, contamination to the outside of the bronchoscope including its tip is inevitable. Use of suction will contaminate the working channel [7]. Since the oral biomass is much greater than the airway biomass, even a small contamination will have a disproportionate effect on the supposed airway microbiome if the unprotected measurements are performed through the working channel. Using the working channel for unprotected lavage repeatedly at different lobes will lead to contamination from one lobe to another. Using larger volume lavage may negate this effect to some degree, but not eliminate the problem.

Results from the current study suggest that protected sheet sampling is the superior sampling methodology. Comparing unprotected SVL and PSB both taken from the upper left lobe in our study, SVL was most similar to the oral sample by visual assessment of the 10 most abundant taxa, and likewise both by alpha- and beta-diversity. A direct comparison of protected and unprotected lavage from the same lobe is impossible, as any washing will impact the contents of later washings. However, the diversity of PBAL from the right middle lobe was intermediate between that found in OW and that found in the PSB.

Besides the above-mentioned study by DICKSON and colleagues [8], only two other studies have compared PSBs to other sampling methods [7, 19]. CHARLSON *et al.* [7] sampled laboratory reagents, the bronchoscope itself during various parts of the procedure, and the oropharyngeal microbiome in addition to BAL through the working channel and PSBs. They concluded that the microbiome from the lower respiratory tract was indiscriminate from the oropharyngeal microbiome irrespective of sampling method. However, the study included only one PSB per sampling, had lower sequencing depth than the current study, included only six healthy individuals and there were no adjustments made for OTUs seen in the negative control samples [7]. HOGAN *et al.* compared PSB, and SVL samples of nine CF patients [19]. For eight CF patients who had PSB and SVL taken from the same lobe, diversity was consistently higher in the PSB samples [19], the opposite of our findings. HOGAN *et al.* employed the PSB only at visible mucus plugs, and the airways of adult CF patients are perhaps no longer representing a low biomass environment. In addition the number of study subjects was limited.

The main strength of our study was comprehensive sampling of a large, heterogeneous sample of subjects with and without COPD, while taking precautions to avoid excessive influence from laboratory and bronchoscopic contamination. However, some potential weaknesses should be acknowledged. First, we have not performed quantitative PCR, and thus cannot conclude regarding the amount of 16S rRNA gene copies in the samples before amplification. Second, our analyses do not include a mock community, and we are therefore not able to provide sequencing error rates for the current study. We could also have spiked our samples with bacteria that would have indicated the efficiency of our DNA extraction. Third, pre-bronchoscopy all participants received 0.4 mg salbutamol. This was done for obtaining pre-bronchoscopy post-bronchodilator lung function values, but had the added benefit of protecting against procedural bronchospasm. Salbutamol was given as an aerosol through large volume spacers that are cleaned daily, and we are not aware of reports on contamination through metered dose inhalers. Furthermore, since both patients and controls received salbutamol, our conclusions should not be affected. Fourth, some results are difficult to compare with those of other authors because of differences in DNA extraction, PCR amplification, sequencing and bioinformatic approach. This is the result of a field where standards for 16S rRNA gene amplicon studies of microbial communities currently do not exist. To facilitate reproducibility we have used well-documented analytic approaches and mostly default settings for our bioinformatic pipeline (QIIME), in addition to using primers and PCR recommendations from a major next-generation sequencing provider (Illumina). Regardless of this, we cannot rule out that some of our findings only pertain to the current set of methodological choices such as the choice of sequencing hypervariable region V3V4 [34]. To minimise the influence of small/spurious OTUs we have excluded singletons by using default settings in our OTU picking, and removed OTUs that constituted less than 0.005% of the total number of sequences.

Insights concerning the airway microbiome in disease and health might provide vital understanding of disease mechanisms and provide new targets for treating lung diseases such as COPD, asthma, cystic fibrosis and interstitial lung diseases. However, to date only a minority of studies have performed protected sampling, and might have been affected by exposure to exposure to microbiota encountered before reaching the sampled sites. We have shown that unprotected sampling is likely to be affected by this phenomenon, and we encourage the use of protected specimen brushes when sampling the airway microbiota.

Acknowledgements

The MicroCOPD study is a large undertaking with many contributing partners. The authors wish to thank Lise Monsen (Dept of Thoracic Medicine, Haukeland University Hospital, Bergen, Norway), Hildegunn Fleten (Dept of Thoracic Medicine, Haukeland University Hospital, Bergen, Norway), Randi Sandvik (Dept of Clinical Science, Faculty of Medicine and Dentistry, University of Bergen, Bergen, Norway), Tove Folkestad (Dept of Clinical Science, Faculty of Medicine and Dentistry, University of Bergen, Bergen, Norway), Ane Aamli (Dept of Clinical Science, Faculty of Medicine and Dentistry, University of Bergen, Bergen, Norway) and Kristina Apalseth (Dept of Thoracic Medicine, Haukeland University Hospital, Bergen, Norway) for help with data collection and aspects of laboratory handling.

R. Grønseth was the guarantor of the study and all authors had full access to all of the data in the study and take full responsibility for the integrity of the data and the accuracy of the data analysis. R. Grønseth, H.G. Wiker, M. Aanerud, P. Bakke and T. Eagan designed the study. R. Grønseth, H.G. Wiker, G.R. Husebø, Ø. Svanes, S. Lehmann, M. Aardal, T. Kalanathan, E.M. Hjeltestad Martinsen, E. Orvedal Leiten, E. Nordeide, I. Haaland, I. Jonassen, P. Bakke and T. Eagan took part in the data collection. T. Hoang, C. Drengenes, H.G. Wiker and T. Kalanathan performed DNA extraction and high-throughput sequencing analyses. R. Grønseth, Y. Xue, S. Tangedal, T. Eagan and I. Jonassen performed statistical and bioinformatic analyses. Data were interpreted by R. Grønseth, C. Drengenes, H.G. Wiker, S. Tangedal, Y. Xue, I. Haaland, I. Jonassen and T. Eagan. R. Grønseth, C. Drengenes, S. Tangedal and T. Eagan drafted the paper. All authors revised the draft and approved the version to be published.

References

- 1 Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. *Nat Rev Genet* 2012; 13: 260–270.
- 2 Morris A, Beck JM, Schloss PD, *et al.* Comparison of the respiratory microbiome in healthy nonsmokers and smokers. *Am J Respir Crit Care Med* 2013; 187: 1067–1075.
- 3 Segal LN, Alekseyenko AV, Clemente JC, *et al.* Enrichment of lung microbiome with supraglottic taxa is associated with increased pulmonary inflammation. *Microbiome* 2013; 1: 19.
- 4 Segal LN, Clemente JC, Tsay JC, *et al.* Enrichment of the lung microbiome with oral taxa is associated with lung inflammation of a Th17 phenotype. *Nat Microbiol* 2016; 1: 16031.
- 5 Dickson RP, Erb-Downward JR, Martinez FJ, *et al.* The microbiome and the respiratory tract. *Annu Rev Physiol* 2016; 78: 481–504.
- 6 Bassis CM, Erb-Downward JR, Dickson RP, *et al.* Analysis of the upper respiratory tract microbiotas as the source of the lung and gastric microbiotas in healthy individuals. *MBio* 2015; 6: e00037–15.
- 7 Charlson ES, Bittinger K, Haas AR, *et al.* Topographical continuity of bacterial populations in the healthy human respiratory tract. *Am J Respir Crit Care Med* 2011; 184: 957–963.
- 8 Dickson RP, Erb-Downward JR, Freeman CM, *et al.* Spatial variation in the healthy human lung microbiome and the adapted island model of lung biogeography. *Ann Am Thorac Soc* 2015; 12: 821–830.


- 9 Venkataraman A, Bassis CM, Beck JM, *et al.* Application of a neutral community model to assess structuring of the human lung microbiome. *MBio* 2015; 6: e02284-14.
- 10 Cabrera-Rubio R, Garcia-Nunez M, Seto L, *et al.* Microbiome diversity in the bronchial tracts of patients with chronic obstructive pulmonary disease. *J Clin Microbiol* 2012; 50: 3562–3568.
- 11 Einarsson GG, Comer DM, McIlreavey L, *et al.* Community dynamics and the lower airway microbiota in stable chronic obstructive pulmonary disease, smokers and healthy non-smokers. *Thorax* 2016; 71: 795–803.
- 12 Erb-Downward JR, Thompson DL, Han MK, *et al.* Analysis of the lung microbiome in the “healthy” smoker and in COPD. *PLoS One* 2011; 6: e16384.
- 13 Pragman AA, Kim HB, Reilly CS, *et al.* The lung microbiome in moderate and severe chronic obstructive pulmonary disease. *PLoS One* 2012; 7: e47305.
- 14 Zakharkina T, Heinzl E, Koczulla RA, *et al.* Analysis of the airway microbiota of healthy individuals and patients with chronic obstructive pulmonary disease by T-RFLP and clone sequencing. *PLoS One* 2013; 8: e68302.
- 15 Hilty M, Burke C, Pedro H, *et al.* Disordered microbial communities in asthmatic airways. *PLoS One* 2010; 5: e8578.
- 16 Huang YJ, Nelson CE, Brodie EL, *et al.* Airway microbiota and bronchial hyperresponsiveness in patients with suboptimally controlled asthma. *J Allergy Clin Immunol* 2011; 127: 372–381.
- 17 Garzoni C, Brugger SD, Qi W, *et al.* Microbial communities in the respiratory tract of patients with interstitial lung disease. *Thorax* 2013; 68: 1150–1156.
- 18 Molyneux PL, Cox MJ, Willis-Owen SA, *et al.* The role of bacteria in the pathogenesis and progression of idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2014; 190: 906–913.
- 19 Hogan DA, Willger SD, Dolben EL, *et al.* Analysis of lung microbiota in bronchoalveolar lavage, protected brush and sputum samples from subjects with mild-to-moderate cystic fibrosis lung disease. *PLoS One* 2016; 11: e0149998.
- 20 Beck JM, Schloss PD, Venkataraman A, *et al.* Multicenter comparison of lung and oral microbiomes of HIV-infected and HIV-uninfected individuals. *Am J Respir Crit Care Med* 2015; 192: 1335–1344.
- 21 Cribbs SK, Uppal K, Li S, *et al.* Correlation of the lung microbiota with metabolic profiles in bronchoalveolar lavage fluid in HIV infection. *Microbiome* 2016; 4: 3.
- 22 Lozupone C, Cota-Gomez A, Palmer BE, *et al.* Widespread colonization of the lung by *Tropheryma whipplei* in HIV infection. *Am J Respir Crit Care Med* 2013; 187: 1110–1117.
- 23 Twigg HL, Knox KS, Zhou J, *et al.* Effect of advanced HIV Infection on the respiratory microbiome. *Am J Respir Crit Care Med* 2016; 194: 226–235.
- 24 Borewicz K, Pragman AA, Kim HB, *et al.* Longitudinal analysis of the lung microbiome in lung transplantation. *FEMS Microbiol Lett* 2013; 339: 57–65.
- 25 Dickson RP, Erb-Downward JR, Freeman CM, *et al.* Changes in the lung microbiome following lung transplantation include the emergence of two distinct *Pseudomonas* species with distinct clinical associations. *PLoS One* 2014; 9: e97214.
- 26 Dickson RP, Erb-Downward JR, Prescott HC, *et al.* Cell-associated bacteria in the human lung microbiome. *Microbiome* 2014; 2: 28.
- 27 Willner DL, Hugenholtz P, Yerkovich ST, *et al.* Reestablishment of recipient-associated microbiota in the lung allograft is linked to reduced risk of bronchiolitis obliterans syndrome. *Am J Respir Crit Care Med* 2013; 187: 640–647.
- 28 Grønseth R, Haaland I, Wiker HG, *et al.* The Bergen COPD microbiome study (MicroCOPD): rationale, design, and initial experiences. *Eur Clin Respir J* 2014; 1: 26196.
- 29 Eagan TM, Ueland T, Wagner PD, *et al.* Systemic inflammatory markers in COPD: results from the Bergen COPD Cohort Study. *Eur Respir J* 2010; 35: 540–548.
- 30 Du Rand IA, Blaikley J, Booton R, *et al.* British Thoracic Society guideline for diagnostic flexible bronchoscopy in adults: accredited by NICE. *Thorax* 2013; 68: Suppl. 1, i1–i44.
- 31 Klindworth A, Pruesse E, Schweer T, *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* 2013; 41: e1.
- 32 McDonald D, Price MN, Goodrich J, *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* 2012; 6: 610–618.
- 33 Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 2005; 71: 8228–8235.
- 34 Hiergeist A, Reischl U, Priority PIMCQAP, *et al.* Multicenter quality assessment of 16S ribosomal DNA-sequencing for microbiome analyses reveals high inter-center variability. *Int J Med Microbiol* 2016; 306: 334–342.

RESEARCH ARTICLE

Open Access



Laboratory contamination in airway microbiome studies

Christine Drengenes^{1,2*} , Harald G. Wiker^{2,3}, Tharmini Kalanathan³, Eli Nordeide¹, Tomas M. L. Eagan^{1,2} and Rune Nielsen^{1,2}

Abstract

Background: The low bacterial load in samples acquired from the lungs, have made studies on the airway microbiome vulnerable to contamination from bacterial DNA introduced during sampling and laboratory processing. We have examined the impact of laboratory contamination on samples collected from the lower airways by protected (through a sterile catheter) bronchoscopy and explored various *in silico* approaches to dealing with the contamination post-sequencing. Our analyses included quantitative PCR and targeted amplicon sequencing of the bacterial 16S rRNA gene.

Results: The mean bacterial load varied by sample type for the 23 study subjects (oral wash>1st fraction of protected bronchoalveolar lavage>protected specimen brush>2nd fraction of protected bronchoalveolar lavage; $p < 0.001$). By comparison to a dilution series of known bacterial composition and load, an estimated 10–50% of the bacterial community profiles for lower airway samples could be traced back to contaminating bacterial DNA introduced from the laboratory. We determined the main source of laboratory contaminants to be the DNA extraction kit (FastDNA Spin Kit). The removal of contaminants identified using tools within the Decontam R package appeared to provide a balance between keeping and removing taxa found in both negative controls and study samples.

Conclusions: The influence of laboratory contamination will vary across airway microbiome studies. By reporting estimates of contaminant levels and taking use of contaminant identification tools (e.g. the Decontam R package) based on statistical models that limit the subjectivity of the researcher, the accuracy of inter-study comparisons can be improved.

Keywords: Microbiome, Contamination, Low biomass, Respiratory, 16S rRNA gene

Background

The most common method used for studying the bacterial communities of the lower respiratory tract is high throughput amplicon sequencing of the bacterial 16S ribosomal RNA (16S rRNA) marker gene [1]. Some studies use sputum samples [2, 3], with inevitable questions regarding the degree to which the samples are representative of the lower respiratory tract as opposed to contamination from the upper respiratory tract. The emerging gold standard for lower respiratory tract samples

is protected bronchoscopy (sampling via a sterile catheter) [4]. However, even with protected bronchoscopy the samples are processed through extensive laboratory workflows that include at minimum steps of bacterial DNA extraction, PCR amplification of the marker gene, and preparation for sequencing. Each step opens up the possibility for the introduction of contaminating bacterial DNA from the laboratory environment, with greatest impact on samples with the lowest bacterial load [5].

Accurate analysis of the lower respiratory tract microbiome will require separate consideration of both of the aforementioned contamination sources - that from the upper respiratory tract introduced during sampling and that introduced during laboratory processing steps. We have previously shown that protected bronchoscopy offers some protection from upper airway contamination

* Correspondence: Christine.Drengenes@gmail.com

¹Department of Thoracic Medicine, Haukeland University Hospital, Bergen, Norway

²Department of Clinical Science, Faculty of Medicine, University of Bergen, Bergen, Norway

Full list of author information is available at the end of the article



[4]. In the current study, we address the issue of contamination from the laboratory.

The impact of laboratory contamination is typically evaluated through the inclusion of negative control samples (NCS) that are processed through all steps of DNA extraction and library preparation for sequencing alongside the study samples. The approach is not perfect as one may expect to find taxa in the NCS that also belong to the bacterial communities of the sampled site. Researchers are thus faced with a difficult decision with regards to what to do with the information acquired from the NCS. Some groups have removed all taxa identified in NCS from their study samples [4, 6, 7]. Others single out taxa they believe likely represent contaminants [8]. Currently bioinformatic tools are being developed that aim to wriggle out the authentic microbiota signal using statistical models [9–11], but these have yet to be tested on lower respiratory tract sequencing data (e.g. Decontam [9]).

In the current paper we illustrate an effective workflow for evaluating the quality of lower respiratory tract samples for accurate assessment of bacterial composition. Objectives of the study were i) to determine the influence of contamination on lower respiratory tract samples as a function of bacterial load, ii) to determine the main source of contamination in our laboratory setting and iii) to explore common in silico approaches to dealing with contamination.

Results

In order to establish the bacterial load in protected airway samples collected using different sampling techniques, we included oral washes (OW), two fractions of protected bronchoalveolar lavage (PBAL1 and PBAL2) and protected specimen brushes (PSB) from 23 participants of the MicroCOPD study [12]. The subject characteristics are provided in Table 1.

Bacterial load varies with sample type

The bacterial load in the four sample types collected per subject was measured by probe based quantitative PCR (qPCR) targeting the bacterial 16S rRNA gene V1 V2 region. The bacterial load decreased in order OW > PBAL1 > PSB > PBAL2 ($p < 0.001$, non-parametric trend test) (Fig. 1). The mean number of bacteria ($\times 10^6$ /mL sample) was 34.2 (range 1.4 to 155.8) for OW ($n = 23$); 1.1 (range 1.7×10^{-3} to 6.6) for PBAL1 ($n = 23$); 0.7 (range 4.3×10^{-3} to 2.8) for PSB ($n = 20$) and 0.5 (range 19.9×10^{-3} to 5.1) for PBAL2 ($n = 23$).

Bacterial load and impact of laboratory contamination

Salter and colleagues [5] have previously illustrated the inverse relationship between the bacterial load in a sample and the influence of contamination on the bacterial

Table 1 Subject characteristics

	Controls	COPD	Asthma
Subjects	9	10	4
Age	63.0 ± 6.7	68.2 ± 5.2	63.6 ± 3.1
Men	6 (66.7%)	8 (80.0%)	2 (50.0%)
Current-smokers	2 (22.2%)	1 (10.0%)	0
Former-smokers	5 (55.6%)	9 (90.0%)	3 (75.0%)
Non-smokers	2 (22.2%)	0	1 (25.0%)
Smoker pack years	11.8 ± 6.1	25.2 ± 8.1	12.1 ± 6.2
FEV ₁ (% predicted)	97.0 ± 13.7	72.6 ± 23.2	101.6 ± 9.3
Inhaled corticosteroids	0	2 (20.0%)	3 (75.0%)
LABA	0	3 (30.0%)	1 (25.0%)
LAMA	0	4 (40.0%)	0

COPD chronic obstructive pulmonary disease, FEV₁ forced expiratory volume in 1 s, LABA long-acting beta-agonist, LAMA long-acting muscarinic antagonist. 1 smoker pack year = 20 cigarettes (one pack) smoked daily for 1 year. Age, smoker pack years and FEV₁ (% predicted) are presented as the mean ± standard deviation

community readout. Once we had established that the bacterial load varied with sampling technique (Fig. 1), we questioned whether the differences in bacterial load for each of the patient samples would also reflect differences in susceptibility to laboratory contamination. Using the Salter approach [5], we estimated the degree of contamination as a function of bacterial load (Fig. 2), and translated this to an estimate of contamination in the procedural samples (OW, PBAL, PSB). Using quantitative PCR we determined that the initial *Salmonella* sample had a concentration of 10^7 bacterial cells/mL. As expected the oral wash samples having a high bacterial load (mean of approximately 10^7 bacterial cells/mL), will not be greatly impacted by contamination. Samples from the lungs (PBAL, PSB) fell between dilution 2 and 3 (Fig. 2), with contamination representing 10–50% of the bacterial community readout. The impact of varying number of PCR cycles was low (Fig. 2).

Monitoring procedural contamination

Having learned that contaminating bacterial DNA likely represents a substantial proportion (10–50%) of the sequencing output for the lower airway samples in our study, we attempted to identify the main contamination source. We performed ten simulated bronchoscopy procedures (no patient) over two days to capture the environmental contaminants that may have been introduced during sampling.

All procedural control samples were sequenced together on the same sequencing run (Run A). Additional control samples were sequenced on a second run (Run B) and included samples of molecular grade water that were processed through the DNA extraction protocol without the introduction of PBS. Although sequenced

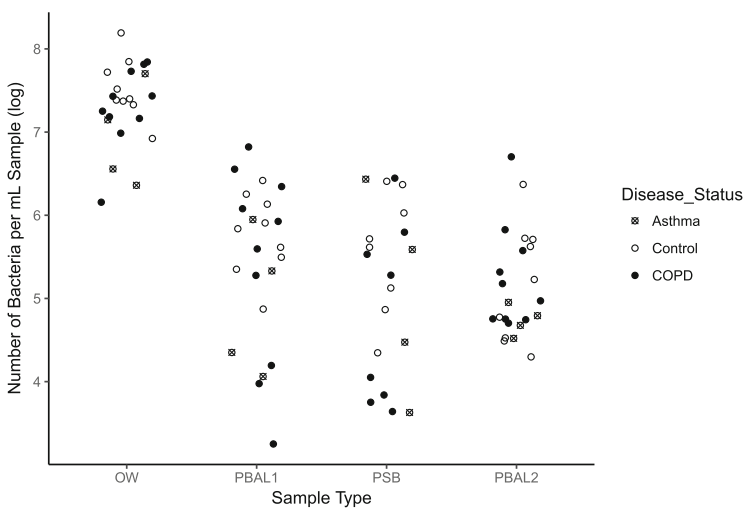


Fig. 1 Measured bacterial load in procedural samples (OW, PBAL1, PSB and PBAL2). The mean bacterial load in OW samples was approximately 30 fold higher than PBAL1, 50 fold higher than PSB and 70 fold higher than PBAL2. OW: oral wash (n = 23); PBAL1: first fraction of protected BAL from right middle lobe (n = 23); PSB: protected specimen brush from right lower lobe (n = 20); PBAL2: second fraction of protected BAL from right middle lobe (n = 23)

on a separate sequencing run (Run B), the molecular grade water samples would indicate whether the PBS was the main source of contamination. A sample of molecular grade water that was not processed through the DNA extraction protocol (PCR water) was also included on both sequencing runs (Run A and B). This later

sample would reflect contamination introduced during PCR and sequencing steps without interference from contamination introduced during sampling and DNA extraction steps.

The total number of sequences obtained from the procedural control samples (Run A) after quality filtering

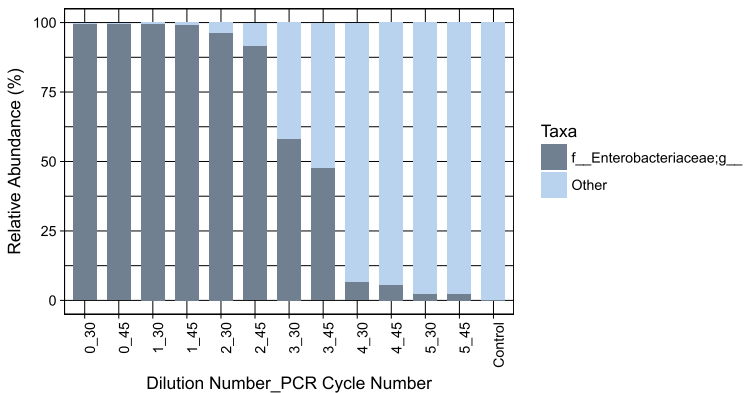


Fig. 2 Estimate of contaminant levels in ten-fold dilution series of *Salmonella* (SDS). The major operational taxonomic units (OTUs) observed in the initial *Salmonella* sample (10^7 bacteria/mL) were assigned to *f_Enterobacteriaceae:g__*. Using the NCBI nucleotide BLAST tool we confirmed that these OTUs (OTU821080, OTU813457 and OTU813217) matched to the genus *Salmonella*. With each successive dilution, the relative abundance of *f_Enterobacteriaceae:g__* decreased. By dilution 3 (45 PCR cycles), the percentage had reduced to 47.83%. For comparison, PCR amplification of the 16S rRNA gene was performed at both 30 and 45 cycles for all SDS samples. The control is a sample of PCR water processed through steps of PCR and sequencing alongside the SDS samples. Taxonomic rank is described using prefixes (*f__*: family, *g__*: genus)

and chimera removal was 4.8×10^6 . The mean number of sequences and operational taxonomic units (OTUs) obtained from the procedural controls were for phosphate buffered saline ($n = 10$): 64,745 sequences (123 OTUs); catheter rinse ($n = 10$): 98,379 sequences (131 OTUs); protected specimen brushes ($n = 10$): 106,853 sequences (132 OTUs); bronchoscope rinse ($n = 10$): 109,765 sequences (134 OTUs); cryotube ($n = 9$): 115,633 sequences (138 OTUs). The number of sequences obtained from the PCR water control sequenced on the same run (Run A) was lower than for the procedural control samples with only 43,433 sequences and 65 OTUs, suggesting that contamination was predominantly introduced prior to PCR steps of library preparation. The procedural control samples (Run A) showed a similar taxonomic distribution that was quite distinct from that of the PCR water sample (Run A) (Fig. 3). This indicated that contamination was either introduced with the phosphate buffered saline used for collection of all samples or during DNA extraction steps.

To differentiate between PBS and DNA extraction as contamination sources, we compared the molecular

grade water samples (Run B) to the corresponding PCR water sample sequenced on the same run. The molecular grade water ($n = 3$) (Run B) contained a mean number of 124,941 sequences and 107 OTUs, whereas the PCR water (Run B) contained 126,103 sequences and only 39 OTUs. Importantly, the taxonomic profile of the molecular grade water (Run B) resembled that of the procedural control samples (Run A), whereas the PCR water did not, indicating that the main source of contamination was the DNA extraction kit (Fig. 3).

Exploring in silico approaches to dealing with contamination in LRT samples

We began our analyses by looking at how the top 20 OTUs present in NCS were distributed in the procedural samples (OW, PBAL, PSB) in our 23 subjects (Fig. 4). The NCS were dominated by an OTU that mapped to the family *Enterobacteriaceae*. The *Ralstonia* OTU that dominated the procedural controls (Fig. 3) was the fourth most abundant OTU in the NCS with an average relative abundance of just 5.45%. This likely reflects differences in contamination introduced from different

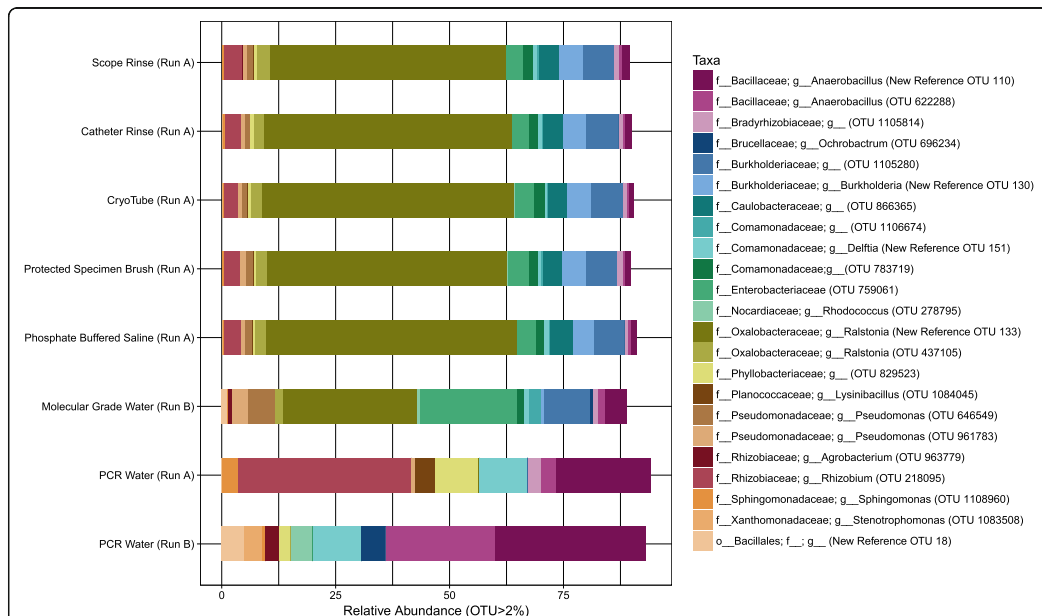


Fig. 3 Distribution of operational taxonomic units (OTUs) in procedural controls and PCR water samples. An OTU belonging to the genera *Ralstonia* dominated the procedural control samples with an average relative abundance of 51.81% in scope rinse ($n = 10$), 54.33% in catheter rinse ($n = 10$), 55.36% in cryotube ($n = 9$), 52.82% in protected specimen brushes ($n = 10$) and 54.93% in phosphate buffered saline ($n = 10$). The same *Ralstonia* OTU also dominated the molecular grade water samples ($n = 3$) at an average relative abundance of 29.42%. The PCR water control sample was dominated by *Rhizobium* (38.11%), *Anaerobacillus* New Reference OTU 110 (20.69%) and *Delftia* (10.65%) in run A and *Anaerobacillus* New Reference OTU 110 (32.93%), *Anaerobacillus* OTU 622288 (24.04%) and *Delftia* (10.68%) in run B. Taxonomic rank is described using prefixes (o__: order, f__: family, g__: genus). Data unrefined

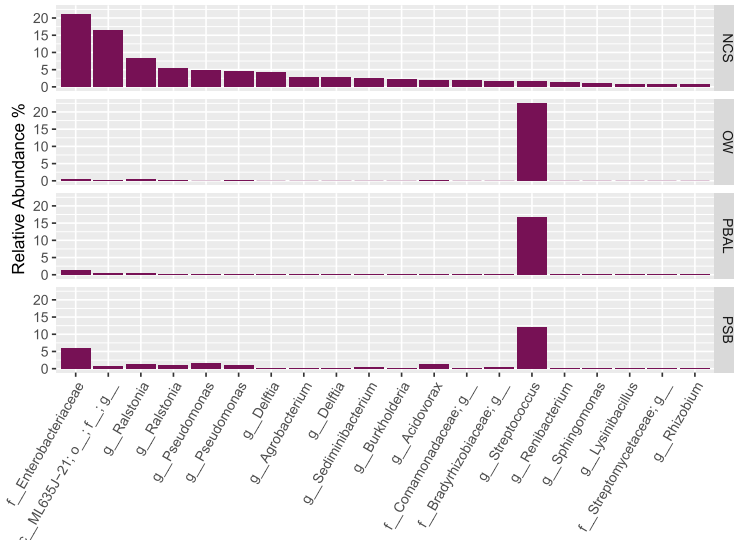


Fig. 4 Distribution of the 20 most abundant operational taxonomic units (OTUs) observed in negative control samples (NCS). The NCS were dominated by OTU 759061 assigned to the family *Enterobacteriaceae* (20.93%), OTU 4389128 assigned to a genus within the class *ML635J-21* (16.31%), OTU 437105 and New. Reference OTU 133 both assigned to the genus *Ralstonia* (8.30 and 5.45%, respectively). Taxonomic rank is described using prefixes (c__: class, o__: order, f__: family, g__: genus). Data presented as the average relative abundance. Data unrefined

lots of the FastDNA Spin Kit [5]. An OTU assigned to the *Streptococcus* genus was found in NCS at a relative abundance of just 1.51%; the same OTU was a major OTU in patient OW, PBAL and PSB samples. This is most likely not a contaminant and may be an important component of the bacterial lung microbiota. For a detailed presentation of the *Streptococcus* OTUs found in PSB and NCS samples, see Additional file 1: Figure S1 and Additional file 2: Figure S2.

Common in silico approaches to dealing with contamination include i) leaving the samples intact (i.e. do nothing), ii) removing all OTUs seen in NCS, and iii) correction based on statistical models (i.e. the Decontam R package). We next examined how the application of each approach would impact the taxonomic profiles of the procedural samples in our study (Fig. 5).

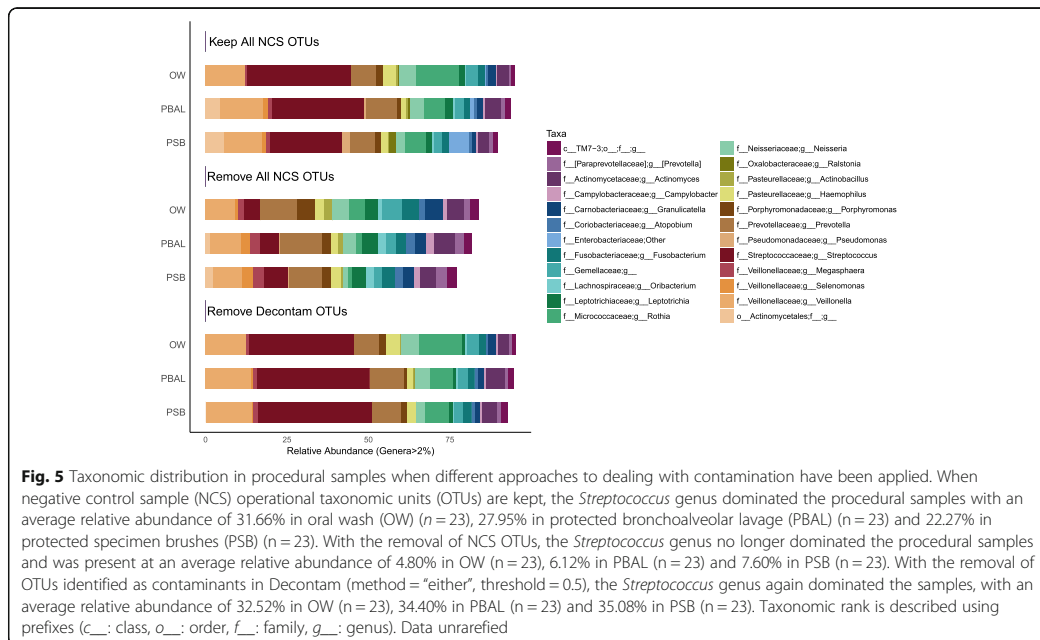
When leaving the procedural samples intact, the *Streptococcus* genus dominated all sample types. With the removal of OTUs seen in NCS, the relative abundance of the *Streptococcus* genus was significantly reduced in all sample types (Fig. 5), as was predicted from Fig. 4. With removal of OTUs identified as contaminants using Decontam [9], the *Streptococcus* genus again dominated the procedural samples. This approach thus appeared to provide a good balance between removing all OTUs found in the NCS and leaving intact OTUs present in both NCS and procedural samples.

Comparison of the frequency-based distribution plots for the top 4 OTUs observed in NCS and the *Streptococcus* OTU (Fig. 6), visually illustrate how Decontam (here frequency-based method) is able to differentiate between a contaminant OTU and a non-contaminant OTU.

Decontam performance test on the *Salmonella* dilution series (SDS)

In the Decontam introduction paper [9], the authors illustrate how Decontam is able to diminish the contaminant signal from the serially diluted *Salmonella* datasets published in the Salter paper [5]. As our study also included a *Salmonella* dilution series (SDS), we were able to test the Decontam package tools on sequencing data generated in the context of our laboratory setting after processing through our chosen bioinformatic pipeline.

The SDS in our study included seven samples of a successively ten-fold diluted *Salmonella* monoculture and a PBS negative control sample that went through DNA extraction and sequencing steps alongside the SDS (Fig. 7). As library preparation for sequencing of the SDS was performed at both 30 and 45 PCR cycles and the impact of varying number of PCR cycles was low (Fig. 2), the sequencing output for both sample sets were used as input in the Decontam analyses. We also included a PCR water control sample that was sequenced on the same sequencing run.



Using the *isContaminant* function in the Decontam R package, we compared three methods for identification of contaminant OTUs including i) the prevalence-based method, ii) the frequency-based method and iii) the either method. In the prevalence-based method, an OTU is marked as a contaminant based on a comparison of how often the OTU is observed in negative control samples compared to the samples under study. For testing the approach on the SDS, the final two samples in the SDS were assigned as negative control samples together with PBS and PCR water samples (as conducted by Decontam developers when testing the approach on the Salter dataset [13]). Figure 8 shows the taxonomic profile of the SDS samples after removal of contaminant OTUs identified using the prevalence-based approach. The impression was that many small OTUs were removed. In the frequency-based approach, the labelling of an OTU as a contaminant is based on the correlation between the DNA concentration measurements made for samples during steps of library preparation (in our lab using the Qubit instrument) and the relative abundance of the OTU across samples. Figure 9 shows the taxonomic profile of the SDS samples after removal of contaminant OTUs identified using the frequency-based approach. The impression was that the frequency-based approach removed fewer but more abundant OTUs compared to the prevalence-based approach. In the final approach tested in Decontam ("either"), all OTUs

marked as contaminants by either the prevalence or frequency-based methods are removed (Fig. 10).

Of the three approaches tested in Decontam, the "either" method was able to most effectively remove the contaminant signal from the bacterial community profiles of the samples; even in the most diluted sample over 50% of the sequences mapped to the *Salmonella* genus. Of concern is however that the PBS sample also consisted of over 50% *Salmonella*. Also present in the PBS sample was oral/lung specific genera including *Veillonella*, *Streptococcus* and *Neisseria* that are obvious contaminants from the procedural samples sequenced on the same run. The number of reads in the PBS sample after processing in Decontam was only 32. Therefore we learn that although effective, removal of contaminant OTUs identified in Decontam may also lead to the magnification of another type of noise in the sequencing data – particularly that from cross sample contamination during library preparation or index misassignment during MiSeq sequencing.

Discussion

In the current paper we illustrate an effective workflow for evaluating the quality of lower airway samples for amplicon-based analysis of bacterial composition. Our results show that the low bacterial load in samples from the lungs make them vulnerable to bacterial DNA contamination, which in our study mainly originated from

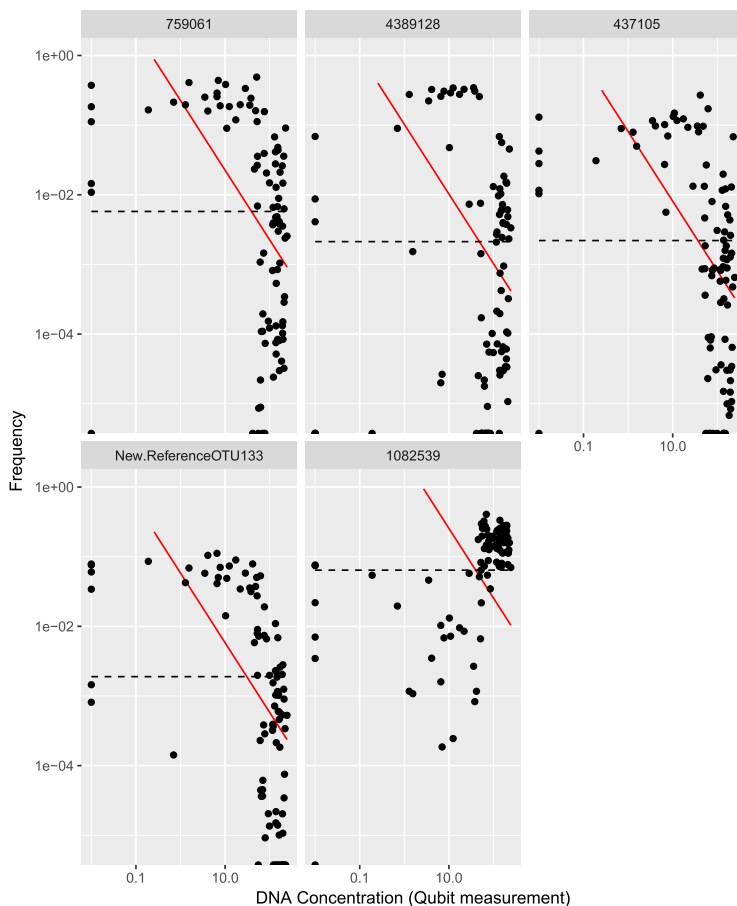
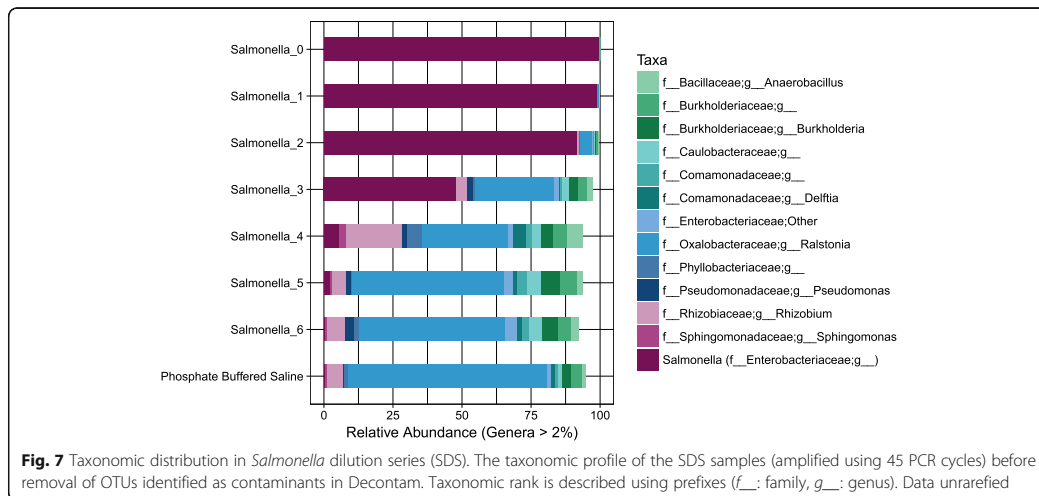


Fig. 6 Decontam frequency distribution plots distinguish contaminants from non-contaminants. A frequency distribution plot generated from samples with varying DNA concentration indicates whether a particular sequence fits the Decontam contaminant (red line) or non-contaminant (black stippled line) model. The first four plots represent the top four operational taxonomic units (OTUs) observed in negative control samples (NCS): OTU 759061 is assigned to the family *Enterobacteriaceae*; OTU 4389128 is assigned to a genus within the class *ML635J-21*; OTU 437105 and OTU New. Reference OTU 133 are both assigned to the genus *Ralstonia*. The final plot represents the *Streptococcus* OTU 1082539 that most likely is not a contaminant, although present among the top 20 OTUs found in NCS. Its frequency distribution pattern more closely fits the Decontam non-contaminant model in contrast to the others

DNA extraction kits. Even with contaminants representing an estimated 10–50% of the sequencing output for these samples, we demonstrate that most of the contaminating signal can be removed post sequencing using recently developed bioinformatic approaches.

Through the processing and sequencing of a serially diluted culture of *Salmonella* [5], we were able to define the threshold bacterial load for which contamination would begin to dominate the bacterial profile in our samples. At an input of between 10^3 and 10^4 *Salmonella*/mL, we observed that contaminants constituted

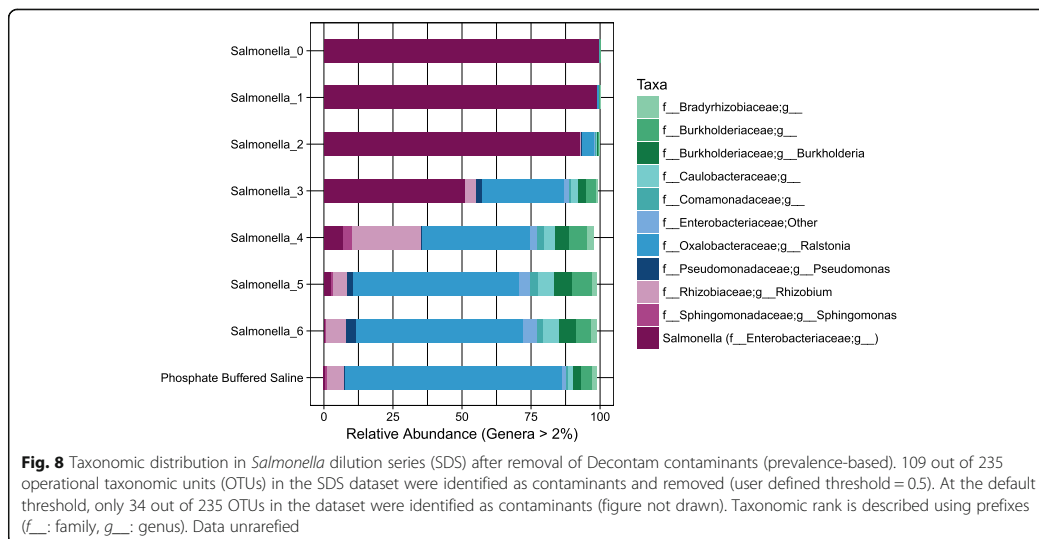
more than 50% of the bacterial profile of the sample. The use of alternative protocols for sample processing and sequencing can slide this defined threshold of bacterial load up or down and should therefore be determined independently in separate studies. Biesbroek et al. [14] for example show in their study how the choice of DNA extraction kit will affect the DNA yield and in turn the placement of samples above or below a defined threshold of bacterial load for which contamination becomes a problem. Despite differences in laboratory protocols, our results are in agreement with Salter and



colleagues [5] who in their study also recommend an input of more than 10^3 – 10^4 bacterial cells. The concordance of our results may partially be explained by the use of a DNA extraction kit from the same manufacturer (FastDNA Spin Kit, MP Biomedicals).

Using the *Salmonella* dilution series as a reference we were able to determine the degree of laboratory contamination in the various sample types (OW, PBAL1, PBAL2, PSB) collected from participants in the Micro-COPD study. The average bacterial load in the samples acquired from the lungs was highest for PBAL1 samples

(10^6 bacteria/mL) and approximately an order of magnitude lower for PSB and PBAL2 samples. This could mean that the first lavage fraction harvests a larger portion of the resident microbiota, but also a dilution effect, as lavage yield tends to increase in the second fraction. We used a sterile inner catheter for lavage sampling, to minimize contamination from BAL, something no other study has done to our knowledge. It is however possible that the first fraction of lavage (PBAL1) is more susceptible to contamination from the upper airways during sampling compared to PBAL2 and PSB samples [4].



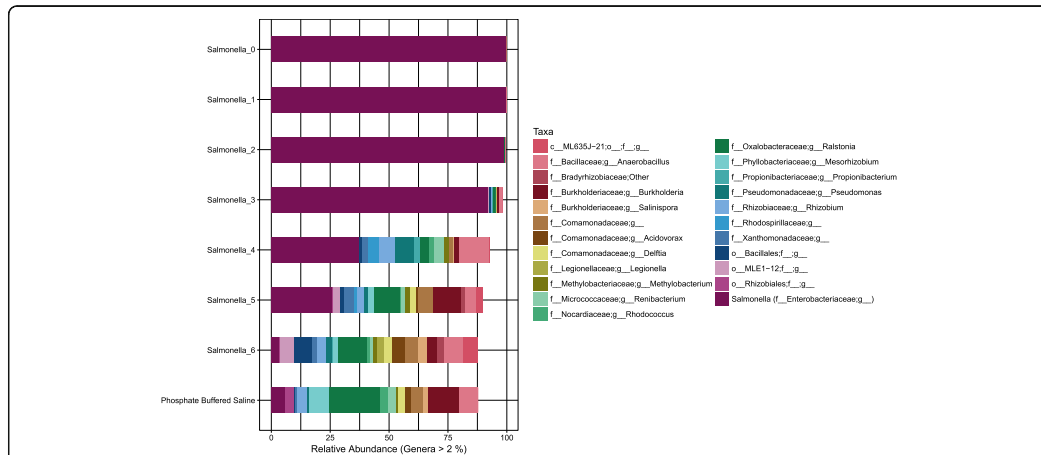


Fig. 9 Taxonomic distribution in *Salmonella* dilution series (SDS) after removal of Decontam contaminants (frequency-based). 58 out of 235 operational taxonomic units (OTUs) in the dataset were identified as contaminants and removed (user defined threshold = 0.5). At the default threshold, only 9 out of 235 OTUs in the dataset were identified as contaminants (figure not drawn). Taxonomic rank is described using prefixes (c_: class, o_: order, f_: family, g_: genus). Data unrarefied

Thus, the question remains as to whether PBAL1 with its higher bacterial load is a more representative sample compared to PBAL2 and PSB samples or if we are simply swapping contamination sources (contaminating bacterial DNA introduced from the upper airways during sampling versus contaminating bacterial DNA introduced during laboratory processing steps). The optimal sample type may thus be a question of which contamination source is easiest to identify and remove post sequencing.

Through the sequencing of procedural control samples and PCR negative control samples that were not processed through the DNA extraction protocol, we were able to trace the main source of contamination back to the DNA extraction kit. Our findings are in agreement with several other studies [5, 15, 16]. The difference in the microbiota readout for the procedural control samples and the negative control samples are likely explained by differences in lot number for the DNA extraction kits. Salter and colleagues report differences

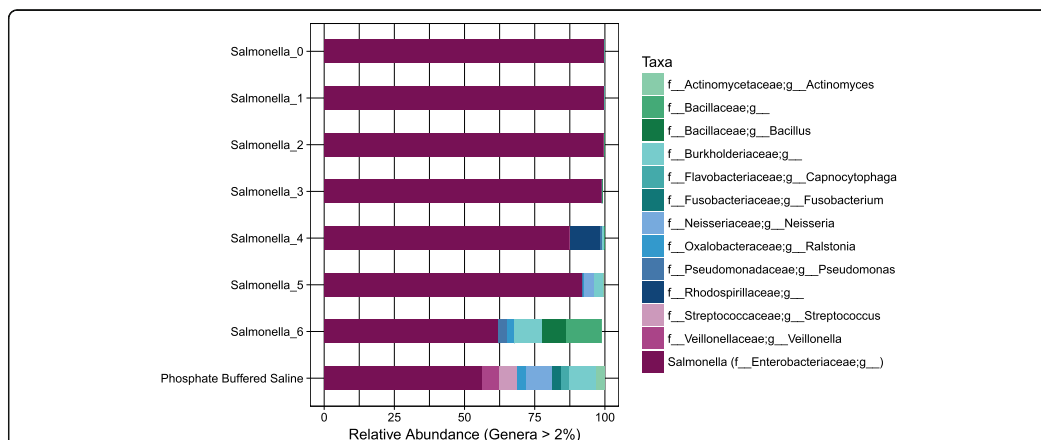


Fig. 10 Taxonomic distribution in *Salmonella* dilution series (SDS) after removal of Decontam contaminants (approach either). 136 out of 235 OTUs in the dataset were identified as contaminants and removed (user defined threshold = 0.5). Taxonomic rank is described using prefixes (f_: family, g_: genus). Data unrarefied

in contaminant profiles for three replicates of SDS extracted using different lots of the FastDNA Spin Kit for soil; similar to our results they also found that one SDS replicate was dominated by unclassified *Enterobacteriaceae*.

Publications such as that by Salter and colleagues have led to an increased awareness of the effects of contamination on microbiome studies of low biomass samples [5, 16]. Most studies now process negative control samples that allow for monitoring of the contaminant signal introduced from the laboratory. However, the inclusion of NCS only partly addresses the issue. In our study for example, we recognized that a major *Streptococcus* OTU found in procedural samples (OW, PBAL, PSB) was also among the top 20 most abundant OTUs found in NCS. A comparison of the relative abundance of the *Streptococcus* OTU in procedural samples and NCS indicated that the OTU was likely not a contaminant. However, the question of where to draw the line with regards to a set abundance threshold for which an OTU should be identified as a contaminant or not is not always as straightforward. The Decontam package in R has been developed to identify contaminants using statistical models [9]. The Decontam developers demonstrate the accuracy of their approach on the *Salmonella* dilution series datasets generated in the Salter publication. We show in the context of our laboratory setting that Decontam is efficient at removing the contaminant signal from the SDS also in our study. Using Decontam we were also able to confirm the identity of the *Streptococcus* OTU found in both procedural samples and the NCS as a non-contaminant.

We acknowledge that our study does not address all issues related to bacterial load in microbiome sequencing data. The serially diluted *Salmonella* monoculture does not provide insight into the effects of bacterial load on the relative abundance of bacteria in a more complex microbiota sample. Biesbroek et al. [14] show in their study examining the microbiota of a serially diluted saliva sample, an increase in the relative abundance of *Proteobacteria* and *Firmicutes* and a decrease in *Bacteroidetes* across the dilution series. *Proteobacteria* likely reflect contaminants as has been suggested in several papers [14, 17], again illustrating the inverse relationship between bacterial load and the influence of contamination as observed in our study. The observed increase in relative abundance of *Firmicutes* and concurrent decrease in *Bacteroidetes* is however of concern, as these phyla hold members often detected in studies of the lung microbiome (e.g. *Veillonella* and *Prevotella*). The field would benefit from studies addressing the potential effects of bacterial load on the measured relative abundance of taxa in a more complex sample, particularly those that are suspect core lung microbiota members. Secondly, we did not quantify the amount of human

DNA in the procedural samples. The presence of human DNA may affect the efficiency of the qPCR reaction [16], and thereby also the accuracy of the direct comparison to the SDS. Studies evaluating the impact of contamination might consider quantification of human DNA for an even more accurate estimate of contamination.

Conclusions

Measured amounts of bacteria will vary in lower airway samples collected with different bronchoscopic sampling techniques (e.g. PBAL1, PBAL2, PSB in the current study). These differences combined with the inverse relationship between bacterial load and bacterial DNA contamination will render some sampling modalities dominated by contaminating taxa.

Differences in protocols for sampling, laboratory processing and bioinformatics analysis across studies will require investigators to evaluate the impact of contamination in the context of their own laboratory setting. We encourage investigators to report an estimate of the degree of contamination in their datasets defined against a sample of known bacterial load as exemplified in the current study. We further suggest the use of contaminant identification tools (e.g. Decontam) based on statistical models for the objective removal of laboratory contaminants in lung microbiome sequencing data. Such measures will enable more accurate inter-study comparisons and may also resolve discrepancies between studies that have likely impeded understanding the potential relationship between microbiota and its role in chronic lung diseases.

Methods

Study samples

Study subjects ($n = 23$) were chosen from the Bergen COPD Microbiome Study (short name “MicroCOPD”) [12], to give an equal representation of healthy ($n = 9$) and diseased (asthma ($n = 4$), COPD ($n = 10$)) states. Details on data collection and the bronchoscopy procedures have been previously published [4, 12]. Briefly, adult subjects recruited from Western Norway with and without obstructive lung disease, underwent voluntary bronchoscopies between 2013 and 2015. All subjects were examined in the stable state, not having received antibiotics at least 2 weeks prior to the procedure. All bronchoscopies were performed by experienced chest physicians at the outpatient clinic at the Department of Thoracic Medicine, Haukeland University Hospital. The regional ethical committee (REK-Nord, case # 2011/1307) approved the study, and all patients gave written informed consent.

Sample types acquired per patient included the first and second fraction of 2×50 mL bronchoalveolar lavage (PBAL1 and PBAL2) sampled through a sterile inner catheter (Plastimed Combicath, Le Plessis Bouchard, France) of the bronchoscope while the scope itself was

wedged in the right middle lobe, and three protected specimen brushes subsequently sampled from the right lower lobe (rPSB), an oral wash (OW), and a negative control sample (NCS). Additional procedural control samples were collected after ten simulated bronchoscopy procedures (no patient) carried out over two days; samples included a bronchoscope rinse (BR), a catheter rinse (CR), a protected specimen brush (PSB), a sample of phosphate buffered saline (PBS) transferred to a cryotube (CT) and a sample of PBS used for collection of all samples. The PBS used for sample collection was sterilized by sterile filtration (0.22 µm) and autoclaving at 121 °C for 15 min. To study the relationship between bacterial load and the influence of contaminating bacterial DNA in our laboratory setting [5], we included a ten-fold dilution series of *Salmonella enterica serovar Typhimurium* (ATCC 14028) (ATCC, Manassas, VA, USA) (SDS).

Bacterial DNA extraction using enzymatic and mechanical lysis steps

Samples were treated with lytic enzymes mutanolysin, lysozyme and lysostaphin (all from Sigma-Aldrich, St. Louis, MO, USA) and subsequently processed through the FastDNA Spin Kit (MP Biomedicals, LLC, Solon, OH, USA) following the manufacturer's instructions. Procedural samples were processed using different lots of the DNA extraction kit (#79113, #84562, #57212, #62903). The procedural controls and the SDS were processed using a kit of same lot number (#93678). The sample volume used as input varied with sample type (for procedural samples: 450 µl for PSB and NCS and 1800 µl for OW, PBAL1, PBAL2; for procedural control samples: 450 µl for PBS and CT, 550 µl for PSB and 1800 µl for BR and CR; for samples in the SDS: 500 µl). DNA was eluted in a total volume of 100 µl.

Quantification of bacterial load by quantitative PCR (qPCR)

The bacterial load in the samples was determined by probe-based qPCR targeting the bacterial 16S rRNA gene (region V1 V2) using forward primer 5'-AGAGTTTGATCCTGGCTCAG-3', reverse primer 5'-CTGCTGCCTYCCGTA-3' and probe 5'-6-FAM-TAACATG-CAAGTCGA-BHQ-1-3' (locked nucleic acid bases are underlined; 6-FAM: 6-carboxyfluorescein; BHQ-1: Black Hole Quencher-1) [7, 18–20]. PCR reactions were carried out using the following cycling conditions: an initial cycle at 95 °C for 5 min followed by 45 cycles of 95 °C for 5 s, 60 °C for 20 s and 72 °C for 10 s and a final extension cycle of 72 °C for 2 min. A standard curve was constructed from genomic DNA from *E. coli* strain JM109 (Zymo Research, Irvine, CA, USA).

MiSeq sequencing of the bacterial 16S rRNA gene

The bacterial composition in the samples was determined by paired-end sequencing of the 16S rRNA gene (region V3 V4) following instructions provided in the Illumina 16S Metagenomic Sequencing Library Preparation guide (Part no. 15044223 Rev. B). PCR cycling conditions were modified from the commercial protocol and consisted of an initial cycle at 95 °C for 3 min followed by 45 cycles of 95 °C for 30 s, 55 °C for 30 s, 72 °C for 30 s and a final extension cycle at 72 °C for 5 min.

Bioinformatic sequence processing steps

Bioinformatic sequence processing steps were performed using tools provided within the Quantitative Insights into Microbial Ecology (QIIME) bioinformatic package, version 1.9.1. In short, raw sequences were retrieved from the MiSeq sequencer in the form of demultiplexed forward and reverse fastq files (paired end reads). Primer sequences were trimmed off and forward and reverse reads joined. Chimera sequences identified using the VSEARCH program [21] were subsequently removed. Remaining sequences were grouped into open-reference operational taxonomic units (OTUs) using UCLUST [22] and the GreenGenes reference database (v.13.8) [23]. Small OTUs, defined as those containing less than 0.005% of the total sequence count in the dataset were then filtered out [24]. Taxonomy was assigned to OTUs using the naïve bayesian RDP Classifier [25] together with the GreenGenes reference database (v.13.8) [23]. The resulting OTU table displaying the sequence count in each OTU for each sample was the starting point for all subsequent analyses. The QIIME commands used for generating the working OTU table are provided in the Additional file 3: Supplementary Methods.

In silico contaminant identification and removal

Two approaches to contaminant identification and subsequent removal were tested. In the first approach contaminant OTUs were identified through their presence in NCS. NCS OTUs were filtered out from the procedural samples (OW, PSB, PBAL) collected under the same procedure using QIIME commands (illustrated in the supplementary methods). In the second approach, contaminant OTUs were identified based on statistical models using the Decontam package [9] in R. Contaminant OTUs identified using the Decontam *isContaminant* function (method = either, user defined threshold = 0.5) were filtered out of the main OTU working table using QIIME commands.

For greater details on study design, sample collection, preparation of *Salmonella* samples, DNA extraction, qPCR, 16S rRNA gene sequencing and bioinformatics, please see the Additional file 3: Supplementary Methods.

Additional files

Additional file 1: Figure S1. Distribution of *Streptococcus* OTUs in Protected Specimen Brush (PSB) samples (n=23). (PDF 8 kb)

Additional file 2: Figure S2. Distribution of *Streptococcus* OTUs in Negative Control Samples (NCS) (n=23). (PDF 8 kb)

Additional file 3: Supplementary Methods. This file provides a detailed description of protocols for sample collection, preparation of *Salmonella* samples, DNA extraction, qPCR, 16S rRNA gene sequencing and bioinformatics. (DOCX 240 kb)

Abbreviations

COPD: Chronic obstructive pulmonary disease; NCS: Negative control sample; OTU: Operational taxonomic unit; OW: Oral wash; PBAL: Protected bronchoalveolar lavage; PSB: Protected specimen brush; QIIME: Quantitative Insights into Microbial Ecology; qPCR: Quantitative PCR

Acknowledgements

The authors wish to thank Marit Aardal, Kristina Apalseth, Hildegunn Bakke Fleten, Ane Aamli Gagnat, Ingvild Haaland, Tuyen Thi Van Hoang, Gunnar Husebø, Kristel Knudsen, Sverre Lehmann, Lise Østgård Monsen, Randi Sandvik, Ølstein Svanes for their contributions in the data collection and/or analyses.

Authors' contributions

TME, RN, HGW, EN and TK participated in the planning and collection of procedural samples in the MicroCOPD study. HGW planned the sequencing analyses. CD, TK, EN, TME and RN participated in planning and collection of procedural control samples. CD and TK performed DNA extraction and library preparation for sequencing. CD performed qPCR, bioinformatics analyses and drafted the manuscript. RN and TME participated in bioinformatics analyses and drafting of the manuscript. All authors participated in the revision of the manuscript and approved the final version for publication.

Funding

The MicroCOPD study was funded by unrestricted grants and fellowships from Helse Vest, Bergen Medical Research Foundation, the Endowment of timber merchant A. Delphin and wife through the Norwegian Medical Association and GlaxoSmithKline through the Norwegian Respiratory Society. The funding bodies had no role in the design of the study, data collection and analysis, interpretation of data, or in writing the manuscript.

Availability of data and materials

The fastq files and metadata needed to rerun the analyses performed in the current study will be available in the DRYAD repository upon publication (doi:<https://doi.org/10.5061/dryad.1v92t8b>). Bioinformatics analyses steps are described in the Additional file 3: Supplementary Methods. A detailed description of the protocols and laboratory materials used in the MicroCOPD Study are available at [dx.doi.org/10.17504/protocols.io.2sygef.v](https://doi.org/10.17504/protocols.io.2sygef.v).

Ethics approval and consent to participate

The study was approved by the Regional Committees for Medical and Health Research Ethics (REK-Nord, case # 2011/1307) and was conducted in accordance with the Declaration of Helsinki. All study subjects signed informed consent forms.

Consent for publication

Not applicable.

Competing interests

CD, HGW, TK, EN: The authors declare that they have no competing interests. TME: Reports bursary from Boehringer Ingelheim for educational meetings within the last three years, unrelated to the current study. RN: Reports grants from GlaxoSmithKline, during the conduct of the study; reports from Boehringer Ingelheim, grants and personal fees from AstraZeneca, grants from Novartis, grants from Boehringer Ingelheim, personal fees from GlaxoSmithKline, outside the submitted work.

Author details

¹Department of Thoracic Medicine, Haukeland University Hospital, Bergen, Norway. ²Department of Clinical Science, Faculty of Medicine, University of Bergen, Bergen, Norway. ³Department of Microbiology, Haukeland University Hospital, Bergen, Norway.

Received: 11 March 2019 Accepted: 31 July 2019

Published online: 14 August 2019

References

- Hilty M, Burke C, Pedro H, Cardenas P, Bush A, Bossley C, et al. Disordered microbial communities in asthmatic airways. *PLoS One*. 2010;5:e8578.
- Tangedal S, Aanerud M, Grønseth R, Drengenes C, Wiker HG, Bakke PS, et al. Comparing microbiota profiles in induced and spontaneous sputum samples in COPD patients. *Respir Res*. 2017;18:164.
- Leitao Filho FS, Alotaibi NM, Ngan D, Tam S, Yang J, Hollander Z, et al. Sputum microbiome is associated with 1-year mortality following COPD hospitalizations. *Am J Respir Crit Care Med*. 2018. <https://doi.org/10.1164/rccm.201806-1135OC>.
- Grønseth R, Drengenes C, Wiker HG, Tangedal S, Xue Y, Husebø GR, et al. Protected sampling is preferable in bronchoscopic studies of the airway microbiome. *ERJ Open Res*. 2017;3.
- Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol*. 2014;12:87.
- Einarsson GG, Comer DM, McIlreavey L, Parkhill J, Ennis M, Tunney MM, et al. Community dynamics and the lower airway microbiota in stable chronic obstructive pulmonary disease, smokers and healthy non-smokers. *Thorax*. 2016;71:795–803.
- Dickson RP, Erb-Downward JR, Freeman CM, Walker N, Scales BS, Beck JM, et al. Changes in the lung microbiome following lung transplantation include the emergence of two distinct *Pseudomonas* species with distinct clinical associations. *PLoS One*. 2014;9. <https://doi.org/10.1371/journal.pone.0097214>.
- Pragman AA, Lyu T, Baller JA, Gould TJ, Kelly RF, Reilly CS, et al. The Lung tissue microbiota of mild and moderate chronic obstructive pulmonary disease. *Microbiome*. 2018;6:7.
- Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*. 2018;6:226.
- Jervis-Bardy J, Leong LEX, Marri S, Smith RJ, Choo JM, Smith-Vaughan HC, et al. Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of Illumina MiSeq data. *Microbiome*. 2015;3:19.
- Lazarevic V, Gaia N, Girard M, Schrenzel J. Decontamination of 16S rRNA gene amplicon sequence datasets based on bacterial load assessment by qPCR. *BMC Microbiol*. 2016;16:73.
- Grønseth R, Haaland I, Wiker HG, Martinsen EMH, Leiten EO, Husebø G, et al. The Bergen COPD microbiome study (MicroCOPD): rationale, design, and initial experiences. *Eur Clin Respir J*. 2014;1:26196.
- Davis NM, Proctor D, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *bioRxiv*. 2018. <https://doi.org/10.1101/221499>.
- Biesbroek G, Sanders EAM, Roesslers G, Wang X, Caspers MPM, Trzciński K, et al. Deep sequencing analyses of low density microbial communities: working at the boundary of accurate microbiota detection. *PLoS One*. 2012;7:e32942.
- Lauder AP, Roche AM, Sherrill-Mix S, Bailey A, Laughlin AL, Bittinger K, et al. Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. *Microbiome*. 2016;4:29.
- Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathog*. 2016;8. <https://doi.org/10.1186/s13099-016-0103-7>.
- Willner D, Daly J, Whitley D, Grimwood K, Wainwright CE, Hugenholz P. Comparison of DNA extraction methods for microbial community profiling with an application to pediatric Bronchoalveolar lavage samples. *PLoS One*. 2012;7:e34605.
- Charlson ES, Bittinger K, Haas AR, Fitzgerald AS, Frank I, Yadav A, et al. Topographical continuity of bacterial populations in the healthy human respiratory tract. *Am J Respir Crit Care Med*. 2011;184:957–63.

19. Dickson RP, Erb-Downward JR, Freeman CM, McCloskey L, Falkowski NR, Huffnagle GB, et al. Bacterial topography of the healthy human lower respiratory tract. *mBio*. 2017;8:e02287-16.
20. Bassis CM, Erb-Downward JR, Dickson RP, Freeman CM, Schmidt TM, Young VB, et al. Analysis of the upper respiratory tract microbiotas as the source of the lung and gastric microbiotas in healthy individuals. *mBio*. 2015;6:e00037-15.
21. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*. 2016;4. <https://doi.org/10.7717/peerj.2584>.
22. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinforma Oxf Engl*. 2010;26:2460–1.
23. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J*. 2012;6:610–8.
24. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, et al. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat Methods*. 2013;10:57–9.
25. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Env Microbiol*. 2007;73:5261–7.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



1 **Exploring protocol bias in airway microbiome studies: one versus**
2 **two PCR steps and 16S rRNA gene region V3 V4 versus V4.**

3 **Authors:**

4 Christine Drengenes, MSc^{1,2}

5 *Christine.Drengenes@gmail.com*

6 Tomas ML Eagan, MD, PhD^{1,2}

7 *Tomas.Eagan@uib.no*

8 Ingvild Haaland, PhD^{1,2}

9 *Ingvild.Haaland@uib.no*

10 Harald G Wiker, MD, PhD^{2,3}

11 *Harald.Wiker@uib.no*

12 Rune Nielsen, MD, PhD^{1,2}

13 *Rune.Nielsen@uib.no*

14

15 **Author affiliation:**

16 ¹ Dept. of Thoracic Medicine, Haukeland University Hospital, Bergen, Norway

17 ² Dept. of Clinical Science, Faculty of Medicine, University of Bergen, Norway

18 ³ Dept. of Microbiology, Haukeland University Hospital, Bergen, Norway

19

20 Corresponding author: *Christine.Drengenes@gmail.com*

21

22 **Keywords:**

23 Microbiota, Contamination, Low Biomass, Respiratory, 16S rRNA gene

24

25 **Abstract**

26 **Background:** Studies on the airway microbiome have been performed using a wide range of
27 laboratory protocols for high-throughput sequencing of the bacterial 16S ribosomal RNA
28 (16S rRNA) gene. We sought to determine the impact of number of polymerase chain
29 reaction (PCR) steps (1- or 2- steps) and choice of target marker gene region (V3 V4 and V4)
30 on the presentation of the upper and lower airway microbiome. Our analyses included
31 Illumina MiSeq sequencing following three setups: Setup 1 (2-step PCR; V3 V4 region), Setup
32 2 (2-step PCR; V4 region), Setup 3 (1-step PCR; V4 region). Samples included oral wash,
33 protected specimen brushes and protected bronchoalveolar lavage (healthy and obstructive
34 lung disease), and negative controls.

35 **Results:** The number of sequences and amplicon sequence variants (ASV) decreased in order
36 setup1>setup2>setup3. This trend appeared to be associated with an increased taxonomic
37 resolution when sequencing the V3 V4 region (setup 1) and an increased number of small
38 ASVs in setups 1 and 2. The latter was considered a result of contamination in the two-step
39 PCR protocols as well as sequencing across multiple runs (setup 1). Although genera
40 *Streptococcus*, *Prevotella*, *Veillonella* and *Rothia* dominated, differences in relative
41 abundance were observed across all setups. Analyses of beta-diversity revealed that while
42 oral wash samples (high biomass) clustered together regardless of number of PCR steps,
43 samples from the lungs (low biomass) separated. The removal of contaminants identified
44 using the Decontam package in R, did not resolve differences in results between sequencing
45 setups.

46 **Conclusions:** Differences in number of PCR steps will have an impact of final bacterial
47 community descriptions, and more so for samples of low bacterial load. Our findings could
48 not be explained by differences in contamination levels alone, and more research is needed

49 to understand how variations in PCR-setups and reagents may be contributing to the
50 observed protocol bias.

51

52

53 **Background**

54 The bacterial airway microbiome has been studied using a wide range of protocols for high-
55 throughput sequencing of the bacterial 16S ribosomal RNA (16S rRNA) gene. Common to all
56 amplicon based protocols is the application of the polymerase chain reaction (PCR) for i)
57 amplification of the target marker gene to be sequenced and ii) the addition of index
58 sequences necessary for sample multiplexing. These steps can be performed in a single PCR
59 or in two separate PCRs. No study has addressed whether the increased number of
60 laboratory processing steps associated with a 2-step PCR protocol, will leave samples more
61 vulnerable to bacterial DNA contamination from the laboratory than when following a 1-step
62 PCR protocol. The inverse relationship between sample bacterial load and the impact of
63 contamination has been well documented in the literature by others [1, 2] and ourselves [3].
64 Thus, we predicted that while samples with a high bacterial load (i.e. upper airway samples)
65 would be able to buffer against protocol effects resulting from differences in contamination
66 levels, samples with a low bacterial load (i.e. lower airway samples) would not be resistant
67 to these effects.

68
69 In addition to number of PCR steps, sequencing protocols vary by choice of targeted marker
70 gene region. Several different 16S rRNA gene variable regions have been targeted in studies
71 of the lung microbiome, including V1 V2 [4, 5], V1 V3 [6–8], V3 V5 [7, 9–13], V3 [14, 15] and
72 V4 [16–20]. Choice of target marker gene region has been limited by the short length of DNA
73 that can be sequenced using current high-throughput sequencing technologies. The V4
74 region has increased in popularity as studies on estimates of alpha- [21] and beta- diversity
75 [22] (i.e. measures of diversity within and between samples, respectively) and taxonomic
76 assignments [23] have collectively indicated that this site generates the most accurate

77 descriptions. In addition, its relatively short length has allowed for the complete overlap of
78 the forward and reverse sequencing read; advantageous because correction of sequencing
79 errors is possible using the read with highest quality score [25]. The increased capacity of the
80 MiSeq sequencer to sequence longer DNA sequences coupled with the development of
81 novel denoising strategies (e.g. DADA2 [26]), has however led to an increased interest in the
82 targeting of the longer V3 V4 region. It is however unclear how these results compare to
83 earlier studies based on the shorter V4 region.

84

85 In the current study, we sought to evaluate the impact of number of PCR steps (1- or 2-
86 steps) and choice of target marker gene region (V3 V4 vs V4) on the presentation of the
87 upper and lower airway microbiome. To address these issues we processed samples of both
88 high and low bacterial load through three library preparation setups varying in the number
89 of PCR steps and target marker gene region: Setup 1 (2-step PCR; V3 V4 region), Setup 2 (2-
90 step PCR; V4 region), Setup 3 (1-step PCR; V4 region). The upper airways were represented
91 by oral wash (OW) samples and the lower airways by protected specimen brushes (PSB) and
92 protected bronchoalveolar lavages (PBAL) collected by bronchoscopy. Negative control
93 samples (NCS) consisting of saline used in the collection of all samples was processed
94 together with the clinical samples for assessment of contamination.

95

96

97

98

99

100 **Results**

101 **Study Participants**

102 The study included 23 subjects from the MicroCOPD study [27]. Subject characteristics are
103 provided in Table 1.

104

105 **Table 1. Subject characteristics.**

	Controls	COPD	Asthma
Subjects	9	10	4
Age, mean \pm SD years	63.0 \pm 6.7	68.2 \pm 5.2	63.6 \pm 3.1
Men	6 (66.7%)	8 (80.0%)	2 (50.0%)
Current-smokers	2 (22.2%)	1 (10.0%)	0
Former-smokers	5 (55.6%)	9 (90.0%)	3 (75.0%)
Never-smokers	2 (22.2%)	0	1 (25.0%)
Smoker pack years, mean \pm SD years	11.8 \pm 6.1	25.2 \pm 8.1	12.1 \pm 6.2
FEV₁ (% predicted), mean \pm SD	97.0 \pm 13.7	72.6 \pm 23.2	101.6 \pm 9.3
Inhaled corticosteroids	0	2 (20.0%)	3 (75.0%)
LABA	0	3 (30.0%)	1 (25.0%)
LAMA	0	4 (40.0%)	0

106 *COPD: chronic obstructive pulmonary disease; FEV₁: forced expiratory volume in 1 second; LABA: long-acting*
107 *beta-agonist; LAMA: long-acting muscarinic antagonist. 1 smoker pack year = 20 cigarettes (one pack) smoked*
108 *daily for 1 year. Age, smoker pack years and FEV₁ (% predicted) are presented as the mean \pm standard*
109 *deviation. SD: standard deviation.*

110

111 **Number of Sequences and Amplicon Sequence Variants (ASVs)**

112 We began our analyses with a comparison of the number of sequences and amplicon
113 sequence variants (ASVs) retained at each step when processing through the bioinformatic
114 pipeline (Figure 1). For sequencing setup 1, the procedural samples were dispersed across four
115 sequencing runs (I-IV). For sequencing setups 2 and 3, two separate sequencing runs (one per
116 setup) were conducted including all samples.

117

118 As the sequences were passed through the different bioinformatic filtering steps, the total
119 number of sequences and ASVs across the three setups became more similar. Denoising in
120 DADA2 (Figure 1, step 1) resulted in the greatest decrease in sequence number. The greatest
121 decrease in ASV number occurred after the removal of *small* ASVs, for which the number of
122 sequences was calculated to be less than 0.005% of the total number of sequences on the
123 same run (Figure 1, step 3). The drop in ASV number was greatest for sequencing setups 1 and
124 2, both of which are based on the longer 2-step PCR protocol.

125

126 After the final filtering step (Figure 1, step 6), the number of ASVs was significantly higher for
127 setup 1 compared to that observed for setups 2 and 3. When we restricted analyses to samples
128 from the largest sequencing run in setup 1 (14 participants, 56 samples) (Figure 2), the number
129 of ASVs for setup 1 was now more comparable to that observed for setups 2 and 3 (Figure 2,
130 step 6). The higher number of ASVs still observed for setup 1, was expected due to the greater
131 taxonomic resolution obtained when targeting a longer marker gene region (V3 V4).

132

133

134

135 **Protocol effects on mock community sample**

136 The mock community sample HM-783D, consisting of genomic DNA from 20 different
137 bacterial species (17 genera) was included on each sequencing run. For a detailed
138 presentation of the mock community, see Additional file 5: Supplementary Methods.

139 Because the protocols targeting different hypervariable regions result in different ASVs, we
140 describe ASVs obtained for setup 1 (V3 V4 target) and setups 2 and 3 (V4 target), separately.
141

142 When following setup 1 across four sequencing runs, we obtained the following number of
143 sequences and ASVs: run I: 128,413 (27 ASVs); run II: 109,709 (23 ASVs); run III: 110,492 (24
144 ASVs) and run IV: 84,909 (27 ASVs). As the number of sequences obtained for each run was
145 similar, ASV numbers were also comparable across the four runs. While most genera were
146 defined by a single ASV, genera *Escherichia*, *Staphylococcus*, *Streptococcus*, *Clostridium* and
147 *Rhodobacter* were defined by multiple ASVs. The major ASVs attributed to each genus (i.e.
148 those with the highest number of sequences) were the same across all four sequencing runs.
149 For a detailed presentation of the ASVs observed in the mock community following setup 1,
150 see Additional File 1: Table S.1.

151
152 When following setups 2 and 3, we obtained 103,409 sequences (31 ASVs) and 120,073
153 sequences (23 ASVs), respectively. The genera *Escherichia*, *Staphylococcus*, *Streptococcus*,
154 *Clostridium* and *Neisseria* were defined by multiple ASVs. The major ASVs attributed to each
155 genus were the same in both setups 2 and 3. For a detailed presentation of the ASVs
156 observed in the mock community following each setup, see Additional File 2: Table S.2. and
157 Additional File 3: Table S.3..

158

159 A summary of the expected and observed taxonomic distribution in the mock community
160 sample, obtained for each setup is presented in Figure 3 and Table 2. We found that the
161 three sequencing setups were for the most part equally efficient at recovering high
162 abundant mock community members. Sequencing setup 3, was least efficient at recovering
163 the low abundant members. Across all setups, we observed an increase in the relative
164 abundances of genera *Escherichia* and *Staphylococcus* and a significant decrease in
165 *Rhodobacter* compared to that expected. All setups generated low abundant ASVs that did
166 not match to any of the expected taxa in the mock community (i.e contaminants). Because
167 the mock community sample was included on each of the four sequencing runs I-IV
168 performed following setup 1, we were also able to show that mock community sequencing is
169 reproducible.

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

Genera	Expected	Setup 1	Setup 1	Setup 1	Setup 1	Setup 2	Setup 3
		(I)	(II)	(III)	(IV)		
<i>Escherichia</i>	21.91	27.68	23.99	25.20	26.65	22.54	32.90
<i>Rhodobacter</i>	21.91	5.98	9.52	9.23	8.94	11.00	8.77
<i>Staphylococcus</i>	24.10	29.02	29.27	29.66	30.56	29.88	24.98
<i>Streptococcus</i>	24.12	28.51	27.39	27.03	25.38	26.20	25.81
<i>Bacillus</i>	2.19	3.38	2.86	2.95	2.85	3.15	2.49
<i>Clostridium</i>	2.19	2.18	3.28	2.19	1.88	2.64	1.69
<i>Pseudomonas</i>	2.19	1.44	1.68	1.75	1.93	2.12	2.02
<i>Acinetobacter</i>	0.22	0.32	0.29	0.33	0.30	0.29	0.12
<i>Helicobacter</i>	0.22	0.36	0.49	0.44	0.38	0.61	0.26
<i>Lactobacillus</i>	0.22	0.22	0.20	0.23	0.24	0.35	0.18
<i>Listeria</i>	0.22	0.33	0.33	0.30	0.32	0.37	0.26
<i>Neisseria</i>	0.22	0.24	0.31	0.30	0.27	0.43	0.39
<i>Propionibacterium</i>	0.22	0.13	0.22	0.18	0.15	0.29	0.00
<i>Actinomyces</i>	0.02	0.01	0.01	0.00	0.00	0.01	0.00
<i>Bacteroides</i>	0.02	0.02	0.00	0.03	0.02	0.04	0.02
<i>Deinococcus</i>	0.02	0.02	0.04	0.03	0.02	0.03	0.02
<i>Enterococcus</i>	0.02	0.03	0.02	0.03	0.02	0.02	0.00
Other	0.00	0.13	0.11	0.10	0.09	0.03	0.08

186

187

188 **Table 2.** Expected and observed relative abundance (%) of genera in mock community

189 sample HM-783D. Setup 1 (2-step PCR; V3 V4 region); Setup 2 (2-step PCR; V4 region); Setup

190 3 (1-step PCR; V4 region).

191

192

193

194

195

196

197 **Protocol effects on contamination profiles**

198 Our working hypothesis linked protocol bias to differences in susceptibility to laboratory
199 contamination. We therefore proceeded with an examination of the average top 20 ASVs
200 found in NCS. Because the same DNA extracts were processed through each of the three
201 setups, any observed differences in taxonomic distribution would be attributed to library
202 preparation steps (post DNA extraction). We also examined PCR water samples included on
203 each sequencing run. In contrast to NCS, this later sample reflects contamination introduced
204 during library preparation steps without interference from contaminating DNA introduced
205 from the DNA extraction kit. ASVs obtained for setups 2 and 3, targeting the V4 region and
206 the single setup targeting the V3 V4 region are described separately.

207

208 The average top 20 ASVs observed in NCS in setups 2 and 3, are presented in Figure 4. The
209 samples were dominated by many of the same taxa, and most of these taxa were defined by
210 the same ASVs. The Decontam package (method=either, threshold= 0.5) applied
211 downstream of the presented data identified the majority of the top 20 ASVs presented in
212 NCS as contaminants. Exceptions included both ASVs mapping to the genus *Streptococcus* (in
213 line with our previous findings [3]) (using NCBI blastn these ASVs were determined to be
214 *Streptococcus oralis* (06f825b512d903b9230e1a55d87359ee) and *Streptococcus*
215 *thermophilus* (fd496fd32dc8c08ade2e8b6c9d8ee13d) and the single ASV mapping to the
216 family *Pasteurellaceae*.

217

218 The distribution of ASVs in NCS (Figure 4) differed the most between setups 2 and 3 for an
219 ASV belonging to the family *Enterobacteriaceae* (mapped to *Escherichia* using NCBI blastn),
220 with a significant increase observed in samples sequenced by setup 3 (0.02% observed for

221 setup 2 and 29.34% observed for setup 3). These findings were in accordance with the
222 results from the mock community analysis (Figure 3), for which the same *Escherichia* ASV
223 was also found at higher levels in the mock community sample sequenced by setup 3
224 (22.54% observed for setup 2 and 32.90% observed for setup 3). Its relatively high
225 abundance in the mock community processed through setup 2 compared to NCS was
226 expected as the *Escherichia* genus defined by this ASV constituted 21.91% of the expected
227 mock community profile; i.e. for this sample the ASV represented both a contaminant and a
228 non-contaminant.

229

230 We proceeded with a comparison of the taxonomic distribution in PCR water samples
231 sequenced following setups 2 and 3 (Table 3). A relatively low number of sequences and
232 ASVs were obtained (setup 2: 178 sequences (10 ASVs); setup 3: 130 sequences (6 ASVs)).
233 Importantly, the dominating ASV (35.38%) found in the PCR water samples sequenced
234 following setup 3, was the same ASV mapping to *Escherichia* discussed above. The same ASV
235 was not found in the PCR water sample sequenced by setup 2. Together these findings
236 indicate that the *Escherichia* ASV is a contaminant introduced during steps of library
237 preparation using a reagent that is exclusive to setup 3.

238

239 We next looked at the average top 20 ASVs observed in NCS when sequencing following
240 setup 1 (Figure 5). The taxonomic profiles obtained after sequencing the longer V3 V4 region
241 resulted in greater taxonomic resolution compared to that observed when sequencing the
242 V4 region in setups 2 and 3. Whereas the three ASVs belonging to the family
243 *Enterobacteriaceae* classified down to genus level *Gluconacetobacter* in setup 1, the
244 *Enterobacteriaceae* ASVs classified no lower than to family level in setups 2 and 3 (Figure 5).

245 The cumulative average relative abundance of the three ASVs mapping to *Gluconacetobacter*
246 when following setup 1 (22 %) was however the same as that found for the single ASV
247 mapping to the family *Enterobacteriaceae* when following setup 2 (23%). Thus, for these two
248 setups, the contamination profiles were similar although greater resolution was obtained
249 when sequencing a longer target gene region in setup 1 (V3 V4).

250

251

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

ASV	Lowest Classification	Setup 2	Setup 3
06f825b512d903b9230e1a55d87359ee ^Δ	f__Streptococcaceae; g__Streptococcus	35.39	20.77
ddfd49f939f92958b1ec816741055348	f__Oxalobacteraceae; g__Ralstonia; s__	12.36	0.00
394eda29c886632f514dd94b58381186	f__Pasteurellaceae	8.99	0.00
d32e579b3ae7b2aae8d5bf9f027c29af	f__Comamonadaceae	8.99	0.00
5648dccee530d68ceb3e4d7d22cf8756	f__Pseudomonadaceae; g__Pseudomonas	7.87	0.00
4f5efd25dacb5d639316e7291ff6ff8b	f__Neisseriaceae; g__Neisseria	7.87	7.69
85c44c83eddc5d3028261a1000b7d0e1	f__Gemellaceae	5.62	0.00
923f521b9cf313f1f95c9367e09bbc1c	f__Veillonellaceae; g__Veillonella; s__dispar	5.62	12.31
dcba105f35d8ebc9e22269c7491ad3a7	f__Xanthomonadaceae; g__Stenotrophomonas; s__geniculata	5.06	0.00
df8456a1abbfb4c8a2c450b44378d4cb	f__Actinomycetaceae; g__Actinomyces; s__	2.25	0.00
d46e2205f0c6ecf67b51f83d111cf509c*	f__Enterobacteriaceae	0.00	35.38
edc9e5c16e40aff1eadce6597940f08f	f__Streptococcaceae; g__Streptococcus; s__	0.00	13.85
65d43491988bfe557da4d86a5ba25dae	f__Staphylococcaceae; g__Staphylococcus	0.00	10.00

270

271

272 **Table 3.** Relative abundance (%) of ASVs observed in PCR water samples in setups 2 and 3.

273 The same *Escherichia* ASV (*) that differentiated mock community samples and NCS in
 274 setups 2 and 3, also caused the greatest difference observed in PCR water samples.

275 Bioinformatic processing steps were performed up until the removal of contaminants

276 identified using Decontam. Taxonomic rank is described using prefixes (*f*__: family, *g*__:

277 genus, *s*__: species).

278

279

280

281

282

283

284

285 **Protocol effects on procedural samples**

286 We next compared the sequencing output obtained for the procedural samples sequenced
287 following each of the three setups. Because we suspected that any differences observed
288 between sequencing setups could be explained by differences in susceptibility to laboratory
289 contamination, comparisons were made both before and after the removal of contaminants
290 identified in Decontam (Figure 1, Step 5).

291

292 Before the removal of Decontam contaminants (Figure 6), we found that across all three
293 sequencing setups, procedural samples (OW, PSB, PBAL) were dominated by many of the
294 same taxa. The most prominent taxa averaged across all samples in order of decreasing
295 relative abundance were genera *Streptococcus*, *Prevotella*, *Veillonella* and *Rothia*. We
296 interpreted these as representative of the authentic airway microbiota based on the growing
297 body of literature for which these same taxa have been consistently observed in airways.

298

299 Several less abundant taxa for which we interpreted as contaminants, based on their
300 dominance in NCS were also observed in the data. We previously learned that ASVs
301 attributed to the family *Enterobacteriaceae* dominated the NCS and that an ASV mapping to
302 *Escherichia* had a discriminating impact on NCS and mock communities processed through
303 setup 3. We were therefore particularly interested in understanding whether
304 *Enterobacteriaceae* would also have a discriminating impact on procedural samples
305 processed through the different sequencing setups. Across all three sequencing setups we
306 found that the levels of *Enterobacteriaceae* was highest in samples from the lower airways
307 (PSB>PBAL) and nearly undetected in OW samples (Figure 6). The higher levels of
308 *Enterobacteriaceae* in PSB samples compared to PBAL, was expected as less sample volume

309 was used as input to the DNA extraction protocol (450 μ l PSB vs 1800 μ l PBAL) thereby
310 securing a lower bacterial load in PSB compared to PBAL. Across all sample types, the
311 relative abundance of *Enterobacteriaceae* was highest when sequencing following setup 3;
312 this was also in accordance with our results when sequencing the mock community and
313 likely due to the additional *Escherichia* contamination introduced during library preparation
314 following setup 3 (Figure 4). By analysis of beta diversity using the unweighted UniFrac
315 metric, we were able to confirm that there was greater overlap or similarity between the
316 bacterial communities found in NCS and procedural samples from the lungs when
317 sequencing following setup 3 (Additional file 4: Figure S.1).

318

319 After the removal of Decontam contaminants, the less abundant taxa that we predicted as
320 representative of contaminants had been filtered out (Figure 7). Although the dominating
321 taxa across all samples were now mainly expected core airway microbiota members, the
322 relative abundances of these taxa still varied across the three setups.

323

324 A direct comparison of the bacterial communities recovered when sequencing by a 1 or 2
325 steps PCR protocol was achieved by analysis of beta-diversity on samples processed through
326 each of setups 2 and 3. Before the removal of Decontam contaminants, OW and NCS
327 clustered together regardless of whether they had been processed through setups 2 or 3
328 (Figure 8). The samples from the lungs however clustered separately according to the
329 protocol for which they were processed. When Decontam contaminants were removed, the
330 samples from the lungs processed by setups 2 and 3 became more similar in bacterial
331 community composition, as indicated by a greater degree of overlap in PCoA space (Figure
332 9). The separation of the lower airway samples based on the setup for which they were

333 processed was however still apparent. This indicated that mechanisms related to the low
334 bacterial load, other than differences in contamination were driving the observed protocol
335 bias.

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

353

354

355

356

357 **Discussion**

358 We have shown that choice of library preparation protocol for high-throughput amplicon-
359 based sequencing of the 16S rRNA gene (1-step PCR vs 2-step PCR) will have an impact on
360 final bacterial community descriptions for airway samples - and more so for samples of low
361 bacterial load. Differences observed when sequencing the different target regions (V3 V4
362 and V4) appeared to be relatively small in comparison, and mainly attributed to differences
363 in taxonomic resolution. Using bioinformatic filtering parameters, we were able to reduce
364 but not completely remove the differences in sequencing output observed for the three
365 sequencing setups: Setup 1 (2-step PCR; V3 V4), Setup 2 (2-step PCR; V4) and Setup 3 (1-step
366 PCR; V4). We propose that protocol bias in studies of the lung microbiome are related not
367 only to differences in susceptibility to contamination but also to less understood (and largely
368 ignored) mechanisms of PCR bias.

369

370 Beginning with a comparison of the number of sequences and ASVs retained at each
371 bioinformatic processing step, we gained insight into the differences in the sequencing
372 output generated for each of the three setups. We found that the removal of small ASVs
373 resulted in the greatest decrease in total ASV number across all three setups - with greatest
374 impact on data generated from the two sequencing setups based on the 2-step PCR protocol
375 (setup 1 and 2). Our interpretation was that the small ASVs likely represent low abundant
376 contamination and that the observed higher frequencies in data generated when processing
377 through longer laboratory workflows was as predicted. Interestingly, this filtering step was
378 originally recommended for filtering out spurious operational taxonomic units (OTUs)
379 derived from PCR and sequencing error [28], and therefore not regarded as necessary after
380 denoising to ASVs [29]. The total number of ASVs after the removal of small ASVs, was still

381 markedly higher when sequencing was performed following setup 1, for which samples were
382 spread across four different sequencing runs. We can expect contamination profiles to vary
383 across sequencing runs, thereby adding to the number of ASVs in the data set, and we
384 therefore interpreted the higher number of ASVs as contamination that had not been
385 filtered out. When analyses were conducted on the subset of samples sequenced on the
386 same run, we still observed a slight increase in ASV count in setup 1; this likely attributed to
387 the greater taxonomic resolution obtained when sequencing a larger gene region. Based on
388 the raw sequencing data, the take home message is therefore that researchers need to pay
389 particular attention to small ASVs when making comparisons across datasets sequenced
390 following different protocols. The observed inflation of ASVs when sequencing across
391 multiple sequencing runs also needs to be accounted for.

392

393 By sequencing of a mock community sample, we were able to show that the three
394 sequencing setups were for the most part equally efficient at recovering the high abundant
395 mock community members. For reasons that are unclear to us, we found that sequencing
396 setup 3, was least efficient at recovering the low abundant members. Together with the
397 observation that the total number of ASVs recovered following setup 3 was lower than for
398 Setups 1 and 2, we concluded that the 1 step-PCR protocol may be less apt for detecting rare
399 but potentially significant taxa [30, 31]. Berry *et al.* [32] also compared sequencing data
400 generated when processing samples through PCR protocols that differed in the number of
401 PCR steps (1-step PCR vs 2-step PCR). In accordance with our findings, they observed
402 reduced richness when processing samples through the 1-step PCR protocol. Thus, it could
403 be that although the 1-step PCR protocol may generate data less influenced by small
404 contaminating ASVs, measures of alpha diversity may be underestimated.

405

406 To further explore the potential impact of contamination, we compared the contamination
407 profiles (based on NCS) obtained for the three sequencing setups. We were surprised to find
408 that the NCS samples processed through setup 3 were dominated by an ASV mapping to
409 *Escherichia coli* (family *Enterobacteriaceae*). It was unexpected because we have previously
410 traced the main source of contamination in the MicroCOPD study to the DNA extraction kit
411 [3]. Because the same DNA extracts were used as input into the sequencing setup 3, we
412 expected that the lower number of laboratory processing steps compared to setups 1 and 2,
413 would secure a contaminant profile representative of that introduced during DNA
414 extraction. We however learned that a contaminant introduced during library preparation
415 was enough to overwhelm the contamination profile of the entire sequencing run. We
416 immediately suspected that the DNA polymerase, manufactured in *Escherichia coli* and used
417 exclusively in the PCR amplification step when sequencing following setup 3, was the main
418 contamination source. Our findings emphasize the fact that researchers must be meticulous
419 in their choice of PCR reagents and also aware of these effects when comparing data
420 generated using different protocols.

421

422 We have previously estimated that contaminants will represent 10-50% of the sequencing
423 output for lower airway samples when sequencing by setup 1 [3]. We found that the
424 *Enterobacteriaceae* family represented less than 10% of the taxonomy profiles for the
425 procedural samples in all three setups and recognized that a significant fraction of the
426 contaminants, were likely also represented by small ASVs and other taxa. For a more
427 accurate assessment of the impact of contamination, we therefore also relied on the
428 Decontam R package [33] for the identification of contaminants. We predicted that if

429 contamination was the main distinguishing factor causing the separation in sequencing
430 output across sequencing setups, the removal of Decontam contaminants would close this
431 gap. By analysis of unweighted Unifrac distances in PCoA space, both before and after the
432 removal of Decontam contaminants, we observed that while the high biomass OW samples
433 clustered together, the low biomass samples from the lungs (PBAL,PSB) separated according
434 to the setup 2 or 3, for which they had been processed. We concluded that factors related to
435 bacterial load, other than contamination must also be contributing to the observed protocol
436 bias.

437

438 The polymerase chain reaction (PCR) lies at the core of all amplicon-based sequencing
439 protocols. The impact of PCR related bias (i.e. all mechanisms that may lead to the
440 preferential amplification of particular sequences or taxa) on studies involving samples
441 holding a low bacterial load is however not well understood. This despite that recent papers
442 as well as research dating back even two decades has documented that PCR related bias
443 appears to increase with decreasing template DNA concentration [1, 34–36]. Kennedy *et al.*
444 [36] observed that bacterial community profiles of replicate soil samples decreased in
445 similarity after sample dilution. The authors attributed these observations to an increased
446 impact of stochastic fluctuations in PCR amplifications at lower bacterial loads. Biesbroek *et*
447 *al.* [1] observed an increase in *Firmicutes* and decrease in *Bacteroidetes* across a serially
448 diluted saliva sample, but were unable to explain the direct mechanism behind their
449 observations. Our study contributes to the literature addressing these issues by
450 demonstrating that samples of high bacterial load (OW) appear to be able to buffer against
451 protocol bias (i.e. differences in number of PCR steps), while samples of low bacterial load

452 (PSB, PBAL) are directly impacted. More research is needed in order to understand the
453 extent to which these mechanisms are responsible for our observations.

454

455 The results presented in the current study have several important implications. Because the
456 upper respiratory tract represents both i) a major potential source of contamination under
457 sampling and ii) the main source community for the lung microbiota, most studies include
458 representative samples from this site (e.g. OW samples) [4, 17, 19, 37, 38]. Our findings
459 demonstrate that the observed overlap between the bacterial communities of the upper and
460 lower respiratory tract may be protocol dependent. Of concern is also that similar
461 community descriptions obtained for upper respiratory tract samples across protocols may
462 mistakenly be interpreted as evidence that datasets are comparable also for lower
463 respiratory tract samples. Our findings also lead us to question the conclusions made in
464 studies where similar PCR reagents have been used. Dickson *et al.* [12] have for example
465 suggested that *Escherichia coli* may be a significant lung pathogen that has previously gone
466 undetected using culture-based techniques. Our results open for interpreting the bacterium
467 as a contaminant introduced with the recombinant DNA polymerase used in the PCR.

468

469 **Conclusion**

470 Our findings show that choice of protocol for library preparation and sequencing (1- or 2-
471 steps of PCR) will have an impact on the analyses of the airway microbiome. Upper airway
472 samples (high biomass) were less impacted than lower airway samples (low biomass),
473 indicating that protocol bias is related to sample biomass. This did not appear to be
474 associated with differences in contamination levels when following a longer or shorter

475 protocol, but rather to mechanisms related to the PCR, for which more research is required.
476 These methodological limitations likely explain the variable conclusions across studies of the
477 airway microbiome (e.g. for comparisons of upper and lower airway samples). Differences in
478 targeted amplicon region (16S rRNA gene V3 V4 versus V4) did not appear have a great
479 impact on final bacterial community descriptions, although greater taxonomic resolution
480 was observed when targeting the longer V3 V4 region.

481

482 **Methods**

483 **Study Samples**

484 The 23 study subjects were chosen from the Bergen COPD Microbiome Study (short name
485 “MicroCOPD”) for representation of both healthy (n=9) and diseased (asthma (n=4), COPD
486 (n=10)) states. Out of the 350 study subjects included in the MicroCOPD study (with samples
487 dispersed across over 30 sequencing runs), the subset of subjects included in the current
488 investigation were chosen in order to minimize the spread of samples across multiple runs.
489 Details on the MicroCOPD study design and bronchoscopy procedures have been
490 previously published [27]. The MicroCOPD study was approved by the regional ethical
491 committee (REK-Vest, case # 2011-1307), and all subjects signed written informed consent.

492

493 In brief, voluntary bronchoscopies were performed on adult subjects (with and without
494 obstructive lung disease) recruited from Western Norway between 2013 and 2015, at the
495 Department of Thoracic Medicine, Haukeland University Hospital. Subjects were examined in
496 the stable state and were not to have received antibiotics at minimum 2 weeks prior to the
497 procedure. Samples collected under each procedure included the first and second fraction of

498 2 x 50 mL protected (through a sterile inner catheter passed through the scope channel)
499 bronchoalveolar lavage (PBAL1 and PBAL2) from the right middle lobe, three protected
500 specimen brushes sampled from the right lower lobe (PSB), an oral wash (OW) sample, and a
501 negative control sample (NCS) taken from the sterile bottle of phosphate buffered saline
502 directly; the same fluid used for BAL sampling, OW, and dissolution of the PSBs.
503 We also included a mock community sample, obtained through BEI Resources NIAID, NIH as
504 part of the Human Microbiome Project: Genomic DNA from Microbial Mock Community B
505 (Staggered, Low Concentration), v5.2L, for 16S rRNA Gene Sequencing, HM783D.

506

507 **Bacterial DNA Extraction**

508 Bacterial DNA extraction was performed first by treatment with lytic enzymes mutanolysin,
509 lysozyme and lysostaphin (all from Sigma-Aldrich, St. Louis, MO, USA) and subsequently by
510 processing through the Fast DNA Spin Kit (MP Biomedicals, LLC, Solon, OH, USA) following
511 the manufacturer's instructions. The sample volume used as input into the DNA extraction
512 protocol varied with sample type; 450 μ l for PSB and NCS and 1800 μ l for OW and PBAL.

513

514 **Library Preparation for MiSeq Sequencing**

515 We processed the same DNA extracts through three different library preparation setups for
516 MiSeq sequencing of the bacterial 16S rRNA marker gene: Setup 1 (2-step PCR; 16S rRNA
517 gene region V3 V4); Setup 2 (2-step PCR; 16S rRNA gene region V4); Setup 3 (1-step PCR; 16S
518 rRNA gene region V4). Setups 1 and 2, were based on the 2-step PCR protocol described in
519 the Illumina 16S Metagenomic Sequencing Library Preparation guide (Part no. 15044223
520 Rev. B). In the first PCR, the 16S rRNA gene regions V3 V4 (setup 1) and V4 (setup 2) were

521 targeted using primers (gene specific sequences are underlined):

522 Setup 1:

523 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG-3' and

524 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC-3'

525 Setup 2:

526 5'TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTGCCAGCMGCCGCGGTAA3' and

527 5'GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGACTACHVGGGTWTCTAAT3'

528

529 PCR cycling was performed with an initial cycle at 95 °C for 3 min followed by 45 cycles of 95

530 °C for 30s, 55 °C for 30 s (setup1)/ 50 °C (setup 2), 72 °C for 30 s and a final extension cycle at

531 72 °C for 5 min. In the second PCR (8 cycles), index sequences were added to the ends of the

532 amplicons generated in the first PCR, using primers from the Nextera XT Index Kit (Illumina

533 Inc., San Diego. CA, USA). Amplifications were performed using the Kappa HiFi HotStart

534 ReadyMix (KAPA Biosystems, USA). Setup 3 was based on the 1-step PCR protocol described

535 in Kozich *et al.* [25], with modifications (see Additional file 5: Supplementary Methods). The

536 primers used targeted the 16S rRNA gene region V4 and consisted of both gene specific

537 sequences (underlined) and index sequences (N):

538 Setup 3:

539 5'AATGATACGGCGACCACCGAGATCTACACNNNNNNNTATGGTAATTGTGTGCCAGCMGCCGCGGTAA3'

540 5'CAAGCAGAAGACGGCATACGAGATNNNNNNNAGTCAGTCAGCCGACTACHVGGGTWTCTAAT3'

541 PCR cycling was performed with an initial cycle at 95 °C for 2 min followed by 45 cycles of 95

542 °C for 20 s, 55 °C for 15 s, 72 °C for 5 min and a final extension cycle at 72 °C for 5 min.

543 Amplifications were performed using the recombinant DNA polymerase Accuprime Pfx Super
544 Mix (Thermo Fisher Scientific, USA).

545 **Bioinformatics**

546 **General Steps.** Sequences were processed using *plugin* tools available within the
547 Quantitative Insights Into Microbial Ecology (QIIME2) bioinformatic package (release
548 2019.1). Two fastq-files per sample (demultiplexed, paired-end reads) were imported into
549 the QIIME2 environment. Using the *dada2 denoise-paired* plugin i) primer sequences and
550 low quality bases at read-ends were trimmed off, ii) paired-end reads were joined, iii)
551 chimeras discarded and iv) amplicon sequence variants (ASVs) inferred [26, 29]. Additional
552 chimera filtering was performed using the *vsearch uchime-denovo* plugin [39]. ASVs with
553 fewer sequences than 0.005% of the total number of sequences and ASVs not found in at
554 least two samples were then discarded [28]. Taxonomy was assigned using the *feature-*
555 *classifier classify-sklearn* plugin together with a Naïve Bayes classifier that had been pre-
556 trained on the full-length Greengenes 13_8 99% OTU reference database (available on
557 qiime2.org). ASVs classified as mitochondria, chloroplasts or archaea were discarded
558 together with classifications that ended above the phylum level. Contaminant ASVs
559 identified using the Decontam package in R were then discarded [40]. The Decontam
560 method “either” (threshold=0.5) was chosen based on our previous work [3]. As the study
561 samples were found across multiple sequencing runs, bioinformatics processing of samples
562 was performed in batches according to run number. Samples not included in the study, but
563 present on the same run were also included in the pipeline to optimize performance of run
564 specific algorithms (e.g. DADA2 and Decontam). **Analyses.** Analysis on taxonomic
565 composition was performed in Excel on ASV tables generated at various stages of the

566 bioinformatic pipeline. Analyses on procedural samples (PSB, PBAL, OW) were performed on
567 the ASV table processed through all general steps described above. Analyses on the top 20
568 ASVs found in NCS and in PCR water controls, were based on the ASV table processed
569 through all steps in the pipeline except removal of contaminants identified in Decontam. For
570 analyses on mock community samples, processing steps were limited to DADA2, VSEARCH
571 and removal of ASVs not classified at minimum to phylum level. Analyses of beta-diversity
572 were conducted using PCoA on unweighted UniFrac distances. The unweighted UniFrac
573 metric scores samples with bacterial communities found at similar positions within the
574 phylogenetic tree, as more similar than samples with bacterial communities found at
575 different positions within the tree. The (dis)similarity between samples is visualized in
576 principal coordinates of analysis (PCoA) space, with samples similar in bacterial composition
577 plotted closer together. The *unweighted* UniFrac metric was chosen to ensure that the less
578 abundant ASVs would have equal impact on the clustering pattern as the high abundant
579 ASVs.

580

581 **Additional Files**

582 **Additional file 1: Table S.1:** The table presents an overview of the sequence count per ASV
583 obtained after V3 V4 sequencing of mock community sample HM-783D following setup 1.

584 **Additional file 2: Table S.2:** The table presents an overview of the sequence count per ASV
585 obtained after V4 sequencing of mock community sample HM-783D following setup 2.

586 **Additional file 3: Table S.3:** The table presents an overview of the sequence count per ASV
587 obtained after V4 sequencing of mock community sample HM-783D following setup 3.

588 **Additional file 4: Figure S.1:** Principal coordinates analysis on unweighted UniFrac distances
589 for procedural samples sequenced following each setup before the removal of Decontam
590 contaminants.

591 **Additional file 5: Supplementary Methods.** The file provides a detailed description of the
592 mock community HMD 783-D, protocols for sequencing. (DOCX)

593

594 **List of abbreviations**

595 ASV: Amplicon sequence variant; COPD: Chronic obstructive pulmonary disease; EMP: Earth
596 Microbiome Project; HMP: Human Microbiome Project; NCS: Negative control sample; OTU:

597 Operational taxonomic unit; OW: Oral wash; PBAL: Protected bronchoalveolar lavage; PSB:
598 Protected specimen brushes; QIIME: Quantitative Insights into Microbial Ecology
599

600 **Declarations**

601

602 **Ethics approval and consent to participate**

603 The study was approved by the Regional Committees for Medical and Health Research Ethics
604 (REK-Nord, case # 2011/1307) and was conducted in accordance with the Declaration of
605 Helsinki. All study subjects signed informed consent forms.

606

607 **Consent for publication**

608 Not applicable.

609

610 **Availability of data and material**

611 Fastq files and metadata will be available in the DRYAD repository upon publication (doi:).
612 Protocols and laboratory materials used in the MicroCOPD study are available at
613 dx.doi.org/10.17504/protocols.io.2sygefww.

614

615 **Competing interests**

616 CD, HGW: The authors declare that they have no competing interests. TME: Reports bursary
617 from Boehringer Ingelheim for educational meetings within the last three years, unrelated to
618 the current study. RN: Reports grants from GlaxoSmithKline, during the conduct of the
619 study; grants from Boehringer Ingelheim, grants and personal fees from AstraZeneca, grants
620 from Novartis, grants from Boehringer Ingelheim, personal fees from GlaxoSmithKline,
621 outside the submitted work.

622 **Funding**

623 The MicroCOPD study was funded by unrestricted grants and fellowships from Helse Vest,
624 Bergen Medical Research Foundation, the Endowment of timber merchant A. Delphin and
625 wife through the Norwegian Medical Association and GlaxoSmithKline through the
626 Norwegian Respiratory Society. The funding bodies had no role in the design of the study,
627 data collection and analysis, interpretation of data, or in writing the manuscript.

628

629 **Authors' contributions**

630 TME, RN, IH and HGW participated in the planning and collection of procedural samples in
631 the MicroCOPD study. HGW and CD planned the sequencing analyses. CD performed DNA
632 extraction, library preparation for sequencing, bioinformatics analyses and drafted the
633 manuscript. RN and TME participated in bioinformatics analyses and drafting of the
634 manuscript. All authors participated in the revision of the manuscript and approved the final
635 version for publication.

636

637 **Acknowledgements**

638 The authors wish to thank Marit Aardal, Kristina Apalseth, Hildegunn Bakke Fleten, Ane
639 Aamli Gagnat, Tuyen Thi Van Hoang, Gunnar Husebø, Tharmini Kalanathan, Kristel

640 Knudsen, Sverre Lehmann, Lise Østgård Monsen, Eli Nordeide, Randi Sandvik, Øistein Svanes
641 for their contributions in the data collection and/or analyses.

642

643 **Authors' information**

644 ¹Department of Thoracic Medicine, Haukeland University Hospital, Bergen, Norway.

645 ²Department of Clinical Science, Faculty of Medicine, University of Bergen, Bergen, Norway.

646 ³Department of Microbiology, Haukeland University Hospital, Bergen, Norway.

647

648 **References**

- 649 1. Biesbroek G, Sanders EAM, Roeselers G, Wang X, Caspers MPM, Trzciński K, et al. Deep
650 Sequencing Analyses of Low Density Microbial Communities: Working at the Boundary of
651 Accurate Microbiota Detection. *PLOS ONE*. 2012;7:e32942.
- 652 2. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and
653 laboratory contamination can critically impact sequence-based microbiome analyses. *BMC*
654 *Biology*. 2014;12:87.
- 655 3. Drengenes C, Wiker HG, Kalanathan T, Nordeide E, Eagan TML, Nielsen R. Laboratory
656 contamination in airway microbiome studies. *BMC Microbiology*. 2019;19:187.
- 657 4. Charlson ES, Bittinger K, Haas AR, Fitzgerald AS, Frank I, Yadav A, et al. Topographical
658 Continuity of Bacterial Populations in the Healthy Human Respiratory Tract. *Am J Respir Crit*
659 *Care Med*. 2011;184:957–63.
- 660 5. Einarsson GG, Comer DM, McIlreavey L, Parkhill J, Ennis M, Tunney MM, et al. Community
661 dynamics and the lower airway microbiota in stable chronic obstructive pulmonary disease,
662 smokers and healthy non-smokers. *Thorax*. 2016;71:795–803.
- 663 6. Erb-Downward JR, Thompson DL, Han MK, Freeman CM, McCloskey L, Schmidt LA, et al.
664 Analysis of the Lung Microbiome in the “Healthy” Smoker and in COPD. *PLoS One*. 2011;6.
665 doi:10.1371/journal.pone.0016384.
- 666 7. Morris A, Beck JM, Schloss PD, Campbell TB, Crothers K, Curtis JL, et al. Comparison of the
667 Respiratory Microbiome in Healthy Nonsmokers and Smokers. *Am J Respir Crit Care Med*.
668 2013;187:1067–75.
- 669 8. Beck JM, Schloss PD, Venkataraman A, Twigg H, Jablonski KA, Bushman FD, et al.
670 Multicenter Comparison of Lung and Oral Microbiomes of HIV-infected and HIV-uninfected
671 Individuals. *Am J Respir Crit Care Med*. 2015;192:1335–44.
- 672 9. Bassis CM, Erb-Downward JR, Dickson RP, Freeman CM, Schmidt TM, Young VB, et al.
673 Analysis of the Upper Respiratory Tract Microbiotas as the Source of the Lung and Gastric
674 Microbiotas in Healthy Individuals. *mBio*. 2015;6:e00037-15.
- 675 10. Dickson RP, Erb-Downward JR, Freeman CM, Walker N, Scales BS, Beck JM, et al. Changes
676 in the Lung Microbiome following Lung Transplantation Include the Emergence of Two
677 Distinct *Pseudomonas* Species with Distinct Clinical Associations. *PLoS One*. 2014;9.
678 doi:10.1371/journal.pone.0097214.
- 679 11. Dickson RP, Erb-Downward JR, Prescott HC, Martinez FJ, Curtis JL, Lama VN, et al. Cell-
680 associated bacteria in the human lung microbiome. *Microbiome*. 2014;2:28.
- 681 12. Dickson RP, Erb-Downward JR, Prescott HC, Martinez FJ, Curtis JL, Lama VN, et al.
682 Analysis of Culture-Dependent versus Culture-Independent Techniques for Identification of
683 Bacteria in Clinically Obtained Bronchoalveolar Lavage Fluid. *J Clin Microbiol*. 2014;52:3605–
684 13.

- 685 13. Venkataraman A, Bassis CM, Beck JM, Young VB, Curtis JL, Huffnagle GB, et al.
686 Application of a Neutral Community Model To Assess Structuring of the Human Lung
687 Microbiome. *mBio*. 2015;6:e02284-14.
- 688 14. Pragman AA, Kim HB, Reilly CS, Wendt C, Isaacson RE. The lung microbiome in moderate
689 and severe chronic obstructive pulmonary disease. *PLoS ONE*. 2012;7:e47305.
- 690 15. Pragman AA, Lyu T, Baller JA, Gould TJ, Kelly RF, Reilly CS, et al. The lung tissue
691 microbiota of mild and moderate chronic obstructive pulmonary disease. *Microbiome*.
692 2018;6:7.
- 693 16. Lozupone C, Cota-Gomez A, Palmer BE, Linderman DJ, Charlson ES, Sodergren E, et al.
694 Widespread colonization of the lung by *Tropheryma whipplei* in HIV infection. *Am J Respir*
695 *Crit Care Med*. 2013;187:1110–7.
- 696 17. Dickson RP, Erb-Downward JR, Freeman CM, McCloskey L, Beck JM, Huffnagle GB, et al.
697 Spatial Variation in the Healthy Human Lung Microbiome and the Adapted Island Model of
698 Lung Biogeography. *Ann Am Thorac Soc*. 2015;12:821–30.
- 699 18. Dickson RP, Erb-Downward JR, Freeman CM, McCloskey L, Falkowski NR, Huffnagle GB,
700 et al. Bacterial Topography of the Healthy Human Lower Respiratory Tract. *mBio*.
701 2017;8:e02287-16.
- 702 19. Segal LN, Clemente JC, Tsay J-CJ, Koralov SB, Keller BC, Wu BG, et al. Enrichment of the
703 lung microbiome with oral taxa is associated with lung inflammation of a Th17 phenotype.
704 *Nat Microbiol*. 2016;1:16031.
- 705 20. Dickson RP, Erb-Downward JR, Freeman CM, McCloskey L, Falkowski NR, Huffnagle GB,
706 et al. Bacterial Topography of the Healthy Human Lower Respiratory Tract. *mBio*.
707 2017;8:e02287-16.
- 708 21. Youssef N, Sheik CS, Krumholz LR, Najjar FZ, Roe BA, Elshahed MS. Comparison of Species
709 Richness Estimates Obtained Using Nearly Complete Fragments and Simulated
710 Pyrosequencing-Generated Fragments in 16S rRNA Gene-Based Environmental Surveys. *Appl*
711 *Environ Microbiol*. 2009;75:5227–36.
- 712 22. Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R. Short pyrosequencing reads suffice
713 for accurate microbial community analysis. *Nucleic Acids Res*. 2007;35:e120.
- 714 23. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naïve Bayesian Classifier for Rapid Assignment
715 of rRNA Sequences into the New Bacterial Taxonomy. *Appl Environ Microbiol*.
716 2007;73:5261–7.
- 717 24. Liu Z, DeSantis TZ, Andersen GL, Knight R. Accurate taxonomy assignments from 16S
718 rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res*.
719 2008;36:e120.
- 720 25. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a Dual-
721 Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on
722 the MiSeq Illumina Sequencing Platform. *Appl Environ Microbiol*. 2013;79:5112–20.
- 723 26. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-
724 resolution sample inference from Illumina amplicon data. *Nature Methods*. 2016;13:581.
- 725 27. Grønseth R, Haaland I, Wiker HG, Martinsen EMH, Leiten EO, Husebø G, et al. The Bergen
726 COPD microbiome study (MicroCOPD): rationale, design, and initial experiences. *Eur Clin*
727 *Respir J*. 2014;1.
- 728 28. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, et al. Quality-filtering
729 vastly improves diversity estimates from Illumina amplicon sequencing. *Nat Methods*.
730 2013;10:57–9.
- 731 29. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace

- 732 operational taxonomic units in marker-gene data analysis. *The ISME Journal*. 2017;11:2639–
733 43.
- 734 30. Jousset A, Bienhold C, Chatzinotas A, Gallien L, Gobet A, Kurm V, et al. Where less may
735 be more: how the rare biosphere pulls ecosystems strings. *ISME J*. 2017;11:853–62.
- 736 31. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, et al. Microbial diversity
737 in the deep sea and the underexplored “rare biosphere.” *PNAS*. 2006;103:12115–20.
- 738 32. Berry D, Mahfoudh KB, Wagner M, Loy A. Barcoded Primers Used in Multiplex Amplicon
739 Pyrosequencing Bias Amplification. *Appl Environ Microbiol*. 2011;77:7846–9.
- 740 33. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical
741 identification and removal of contaminant sequences in marker-gene and metagenomics
742 data. *Microbiome*. 2018;6:226.
- 743 34. Chandler DP, Fredrickson JK, Brockman FJ. Effect of PCR template concentration on the
744 composition and distribution of total community 16S rDNA clone libraries. *Mol Ecol*.
745 1997;6:475–82.
- 746 35. Polz MF, Cavanaugh CM. Bias in Template-to-Product Ratios in Multitemplate PCR. *Appl*
747 *Environ Microbiol*. 1998;64:3724–30.
- 748 36. Kennedy K, Hall MW, Lynch MDJ, Moreno-Hagelsieb G, Neufeld JD. Evaluating Bias of
749 Illumina-Based Bacterial 16S rRNA Gene Profiles. *Appl Environ Microbiol*. 2014;80:5717–22.
- 750 37. Grønseth R, Drengenes C, Wiker HG, Tangedal S, Xue Y, Husebø GR, et al. Protected
751 sampling is preferable in bronchoscopic studies of the airway microbiome. *ERJ Open Res*.
752 2017;3.
- 753 38. Segal LN, Alekseyenko AV, Clemente JC, Kulkarni R, Wu B, Chen H, et al. Enrichment of
754 lung microbiome with supraglottic taxa is associated with increased pulmonary
755 inflammation. *Microbiome*. 2013;1:19.
- 756 39. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool
757 for metagenomics. *PeerJ*. 2016;4. doi:10.7717/peerj.2584.
- 758 40. Davis NM, Proctor D, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification
759 and removal of contaminant sequences in marker-gene and metagenomics data. *bioRxiv*.
760 2017;:221499.

761

762 **Figure Legends**

763 **Figure 1.** Comparison of the number of sequences and amplicon sequence variants (ASVs),
764 retained at each bioinformatic filtering step for procedural samples (PSB, PBAL, OW, NCS)
765 collected from 23 participants (n=92 samples). Setup 1 (2-step PCR; V3 V4 region), Setup 2 (2-
766 step PCR; V4 region), Setup 3 (1-step PCR; V4 region).

767

768 **Figure 2.** Comparison of the number of sequences and amplicon sequence variants (ASVs),
769 retained at each bioinformatic filtering step for procedural samples (PSB, PBAL, OW, NCS)

770 collected from 14 participants (n=56). Setup 1 (2-step PCR; V3 V4 region), Setup 2 (2-step PCR;
771 V4 region), Setup 3 (1-step PCR; V4 region).

772

773 **Figure 3.** Analysis of mock community HM-783D. The expected relative abundances of
774 genera in the mock community sample is presented next to that observed in the sequencing
775 output across the three setups. The *Escherichia* genus consisted of ASVs classified to family
776 level (*Enterobacteriaceae*); ASV ffc36e27c82042664a16bcd4d380b286 dominated Setup 1
777 targeting the 16S rRNA gene V3 V4 region and ASV d46e2205f0c6ecf67b51f83d111c509c
778 dominated Setups 2 and 3 targeting the V4 region. Using the NCBI blastn tool we were able
779 to confirm that these ASVs belonged to the *Escherichia coli* genus. Bioinformatics processing
780 steps were limited to DADA2, VSEARCH, taxonomy assignment and removal of features not
781 classified at minimum to phylum level. Setup 1 (2-step PCR; V3 V4 region), Setup 2 (2-step
782 PCR; V4 region), Setup 3 (1-step PCR; V4 region).

783

784 **Figure 4.** Comparison of the 20 most abundant amplicon sequence variants (ASVs) observed
785 in negative control samples (NCS) after sequencing following setups 2 and 3. Taxa presented
786 according to decreasing abundance for ASVs observed following setup 2. Bioinformatic
787 processing steps were performed up the removal of contaminants identified using
788 Decontam. Taxonomic rank is described using prefixes (*c__*: class, *o__*: order, *f__*: family,
789 *g__*: genus). Setup 2 (2-step PCR; V4 region); Setup 3 (1-step PCR; V4 region). Data is
790 presented as the average relative abundance. Data unrarefied.

791

792 **Figure 5.** The 20 most abundant amplicon sequence variants (ASVs) observed in negative
793 control samples (NCS) after sequencing following setup 1. Multiple ASVs mapped to genera

794 *Gluconacetobacter*, belonging to family *Enterobacteriaceae* (cumulative 22%).
 795 Bioinformatic processing steps were performed up until the removal of contaminants
 796 identified using Decontam. Taxonomic rank is described using prefixes (c__: class, o__:
 797 order, f__: family, g__: genus). Data is presented as the average relative abundance. Data
 798 unrarefied.
 799

800 **Figure 6.** Taxonomic distribution obtained for procedural samples before the removal of
 801 Decontam contaminants. ASVs attributed to the family *Enterobacteriaceae*, had dominated
 802 the NCS across all setups. In the procedural samples, *Enterobacteriaceae* was observed with
 803 the following relative abundances in setups 2 and 3: setup 2 (OW: 0%; PBAL: 0.83%; PSB:
 804 5.23%); setup 3 (OW: 0.01%; PBAL: 1.87%; PSB: 7.51%). ASVs attributed to the genus
 805 *Gluconacetobacter* within the family *Enterobacteriaceae* was observed in procedural
 806 samples with the following relative abundances in setup 1 (OW: 0%; PBAL: 1.42%; PSB: 6.32
 807 %). Samples with fewer than 1000 sequences had been omitted from the analyses leaving
 808 the following number of samples in each setup: Setup 1 (OW: n=22; PBAL: n=23; PSB: n=23);
 809 Setup 2 (OW: n=23; PBAL: n=23; PSB: n=23); Setup 3 (OW: n=23; PBAL: n=21; PSB: n= 22).
 810 Setup 1 (2-step PCR; V3 V4 region); Setup 2 (2-step PCR; V4 region); Setup 3 (1-step PCR; V4
 811 region). Taxonomic rank is described using prefixes (p__: phyla; c__: class; o__: order; f__:
 812 family; g__: genus).

813
 814 **Figure 7.** Taxonomic distribution obtained for procedural samples after the removal of
 815 Decontam contaminants. Samples with fewer than 1000 sequences were omitted. Number
 816 of samples in each setup: V3V4 protocol A (OW: n=22; PBAL: n=22; PSB: n=21), V4 protocol A
 817 (OW: n=23; PBAL: n=22; PSB: n=20); V4 protocol B (OW: n=23; PBAL: n=21; PSB: n= 21).

818 Setup 1 (2-step PCR; V3 V4 region); Setup 2 (2-step PCR; V4 region); Setup 3 (1-step PCR; V4
819 region). Taxonomic rank is described using prefixes (p__: phyla; c__: class; o__: order; f__:
820 family; g__: genus).

821

822 **Figure 8.** Principal coordinates analysis on unweighted UniFrac distances for procedural
823 samples sequenced following setup 2 (sphere) and 3 (diamond) *before* the removal of
824 Decontam contaminants. Rarefaction depth: 1066 sequences. Setup 2 samples include OW:
825 n=23; PBAL: n=23; PSB: n= 23; NCS: n=21 and setup 3 samples include OW: n=23; PBAL:
826 n=21; PSB: n=22; NCS: n=18. Oral Wash (OW): blue; Protected bronchoalveolar lavage
827 (PBAL): green; Protected specimen brushes (PSB): purple; Negative control samples (NCS):
828 red. Setup 2 (2-step PCR; V4 region), Setup 3 (1-step PCR; V4 region).

829

830 **Figure 9.** Principal coordinates analysis on unweighted UniFrac distances for procedural
831 samples sequenced following setup 2 (sphere) and 3 (diamond) *after* the removal of
832 Decontam contaminants. Rarefaction depth: 1139 sequences. Setup 2 samples include OW:
833 n=23; PBAL: n=21; PSB: n= 20 and setup 3 samples include OW: n=23; PBAL: n=21; PSB:
834 n=21. Oral Wash (OW): blue; Protected bronchoalveolar lavage (PBAL): green; Protected
835 specimen brushes (PSB): purple. Setup 2 (2-step PCR; V4 region), Setup 3 (1-step PCR; V4
836 region).

837

838

839

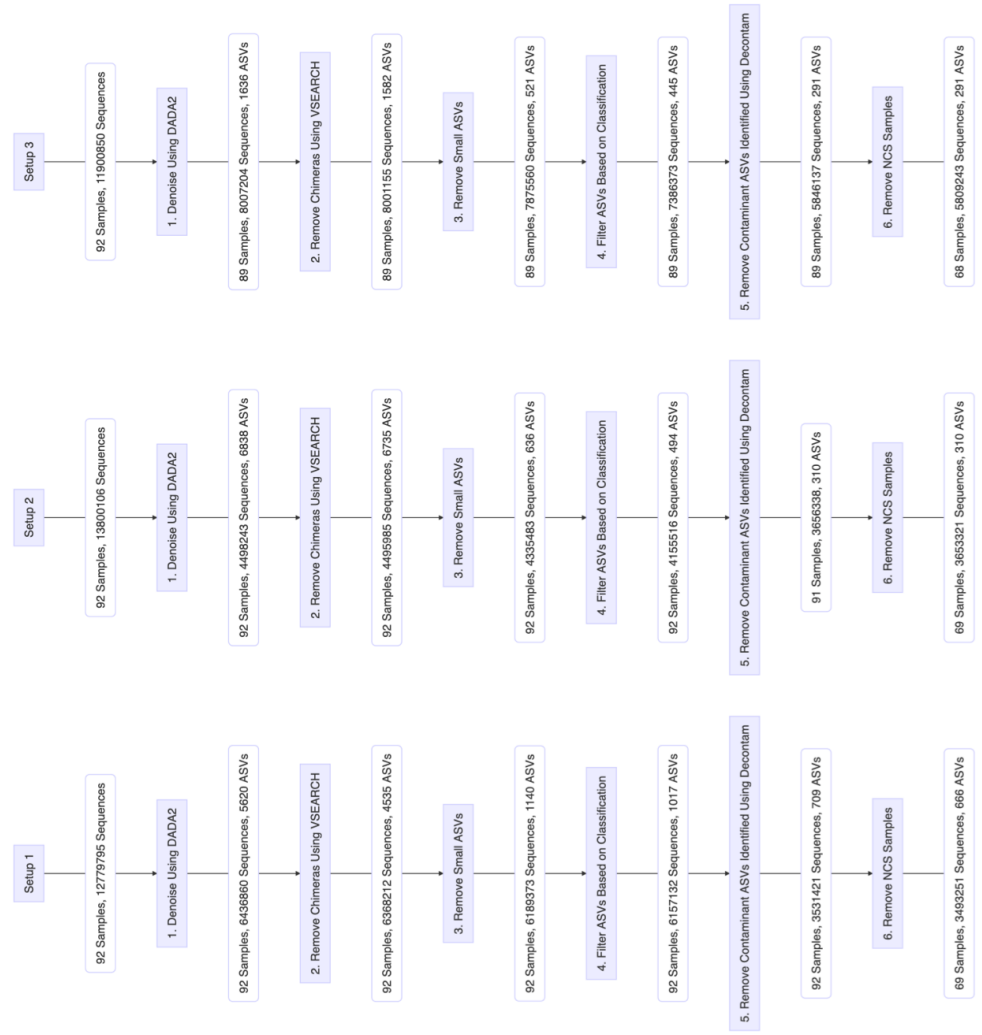
840

841

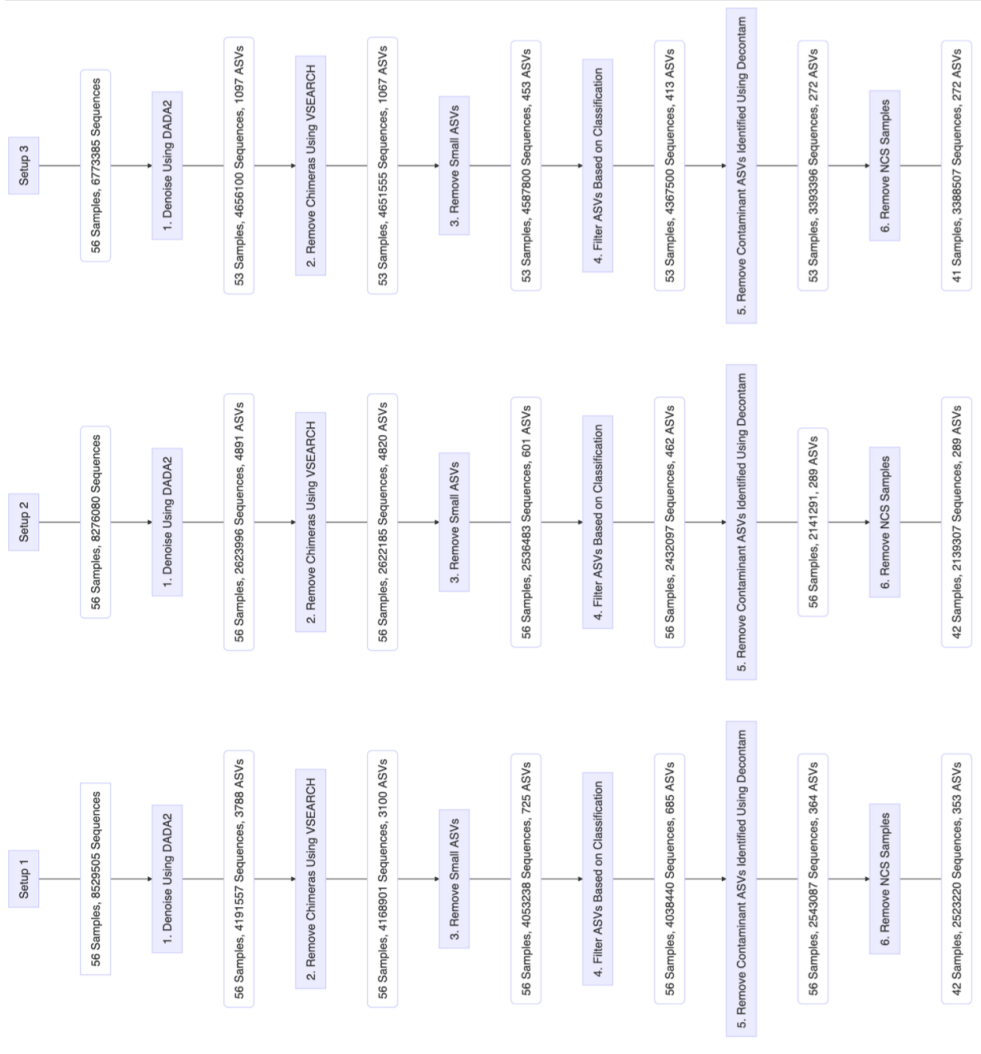
842

843

844 Figure_1



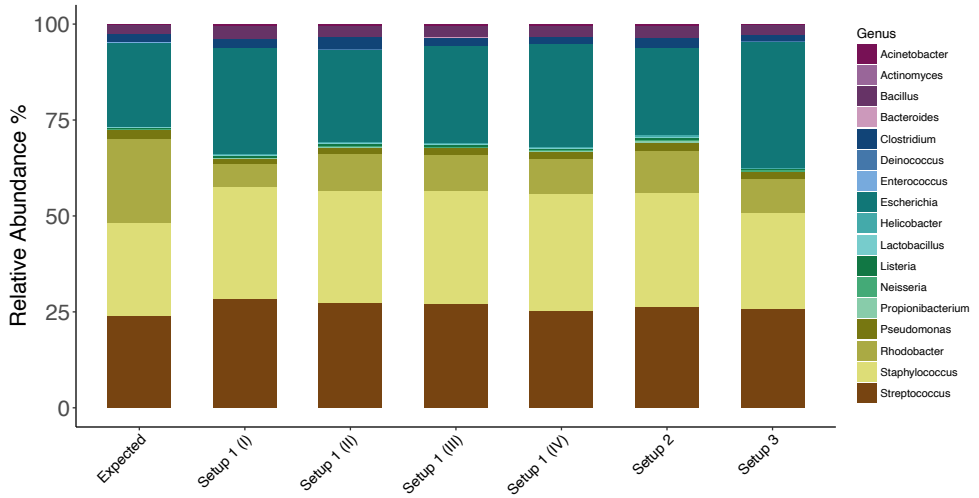
845
846
847



852 **Figure_3**

853

854



855

856

857

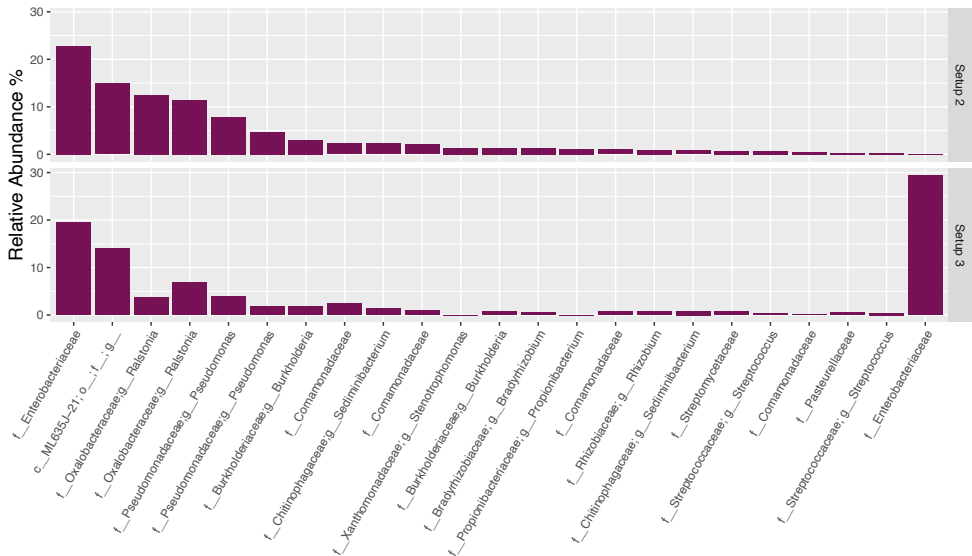
858

859 **Figure_4**

860

861

862



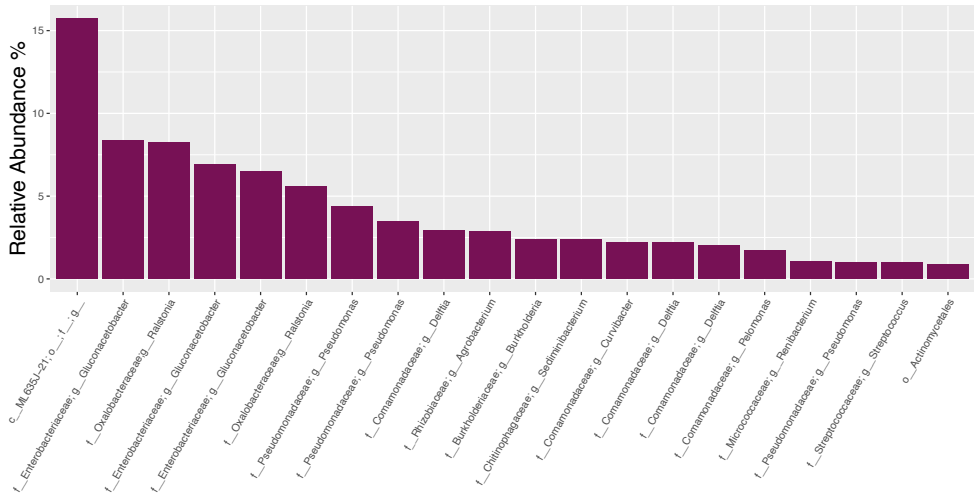
863

864

865

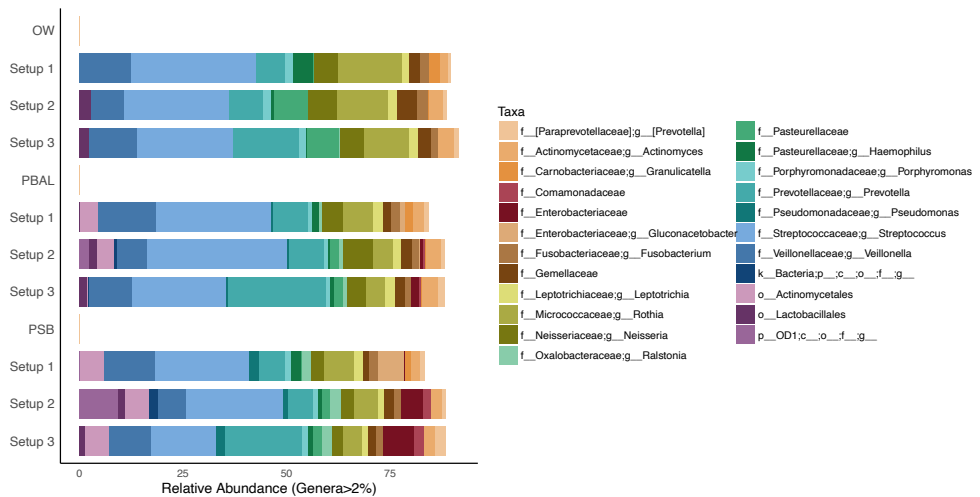
866 **Figure_5**

867
868
869



870
871
872
873
874
875
876
877

Figure_6



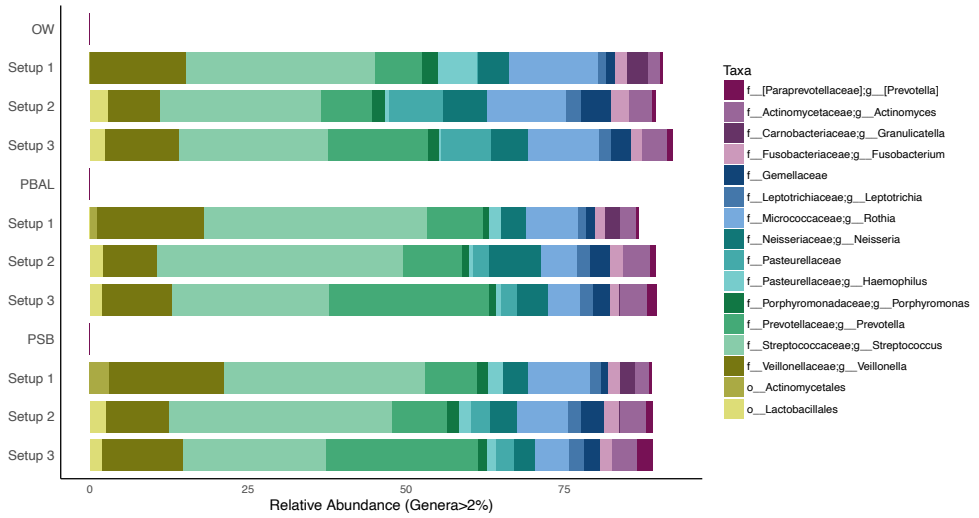
878
879
880
881
882

883 **Figure_7**

884

885

886



887

888

889

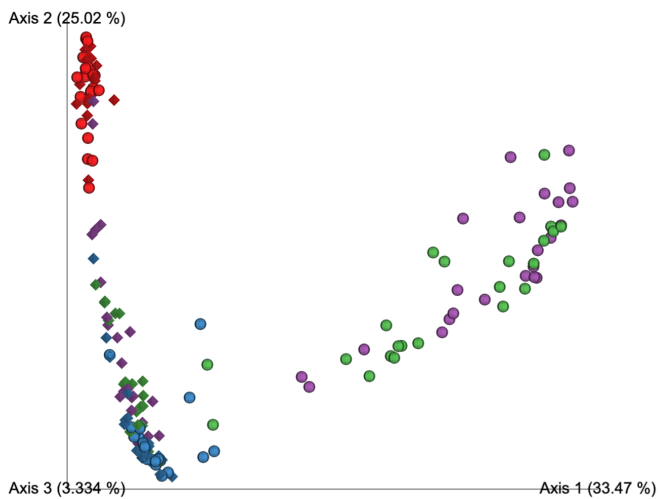
890

891 **Figure_8**

892

893

894



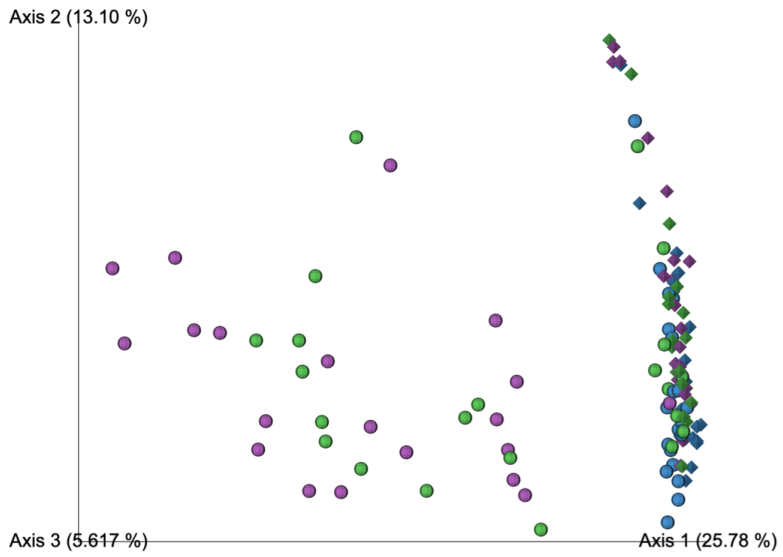
895

896 **Figure_9**

897

898

899



900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

918

919

920

Additional file 1

921
922
923
924
925

Mock Community HM-783D, Sample ID MKOLS1848, Setup 1 (I)

Mock Community Member (L6)	QIIME 2 lowest classification level	ASV	Sequence Count	Relative Abundance %	L6 Relative Abundance %
<i>Escherichia</i>	f__Enterobacteriaceae	ffc36e27c82042664a16bcd4d380b286	35542	27.67788308	27.68
<i>Staphylococcus</i>	f__Staphylococcaceae; g__Staphylococcus	5497318e515a8c328a68f95975d9c7d4	29124	22.67994673	29.02
	f__Staphylococcaceae; g__Staphylococcus	908e9b387f6b9ce7d3f794e658fba37e	5133	3.997258845	
	f__Staphylococcaceae; g__Staphylococcus	04be702b895fe5ed0568344daf564276	3004	2.339237015	
<i>Streptococcus</i>	f__Streptococcaceae; g__Streptococcus; s__	e5af48ec2b6c023c5de28c59cb08a40	32219	25.09013885	28.51
	f__Streptococcaceae; g__Streptococcus; s__agalactiae	c3a3a503752209bc052b3995236b079f	4377	3.408533404	
	f__Streptococcaceae; g__Streptococcus	a3725fbb7f4a76528d54d283e88cad8	13	0.010123586	
	f__Streptococcaceae; g__Streptococcus	2a025812c0b5b9ce4c6d4ab4c692bed7d	4	0.003114949	
<i>Rhodobacter</i>	f__Rhodobacteraceae; g__Rhodobacter; s__sphaeroides	332b70897316f7f62b81dfc53f41ca52	7676	5.977587939	5.98
<i>Bacillus</i>	f__Bacillaceae; g__Bacillus	8cb24777cb48dde0aac60dfeca125d10	4343	3.382056334	3.38
<i>Clostridium</i>	f__Clostridiaceae; g__Clostridium; s__butyricum	318669d5d926e9b81ca6911da00a14ea	2577	2.006806164	2.18
	f__Clostridiaceae; g__Clostridium; s__butyricum	88a3a8e95e3605bd051054937cde102	228	0.177552117	
<i>Pseudomonas</i>	f__Pseudomonadaceae	052ba7abaaa968c4f79e3f97d1f0a2f	1850	1.440664107	1.44
<i>Helicobacter</i>	f__Helicobacteraceae; g__Helicobacter; s__pylori	e8de432f2ada1078a2fda56ba92675a	464	0.361334133	0.36
<i>Neisseria</i>	f__Neisseriaceae; g__Neisseria; s__cinerea	f71e5aac4c976a1833958c2870b1d8b	306	0.238293631	0.24
<i>Listeria</i>	f__Listeriaceae; g__Listeria	b77b151f7481fd080bedbf415e736539	430	0.334857063	0.33
<i>Lactobacillus</i>	f__Lactobacillaceae; g__Lactobacillus; s__	9b23053de8f4269feb5ce286dfbf3c	279	0.217267722	0.22
<i>Acinetobacter</i>	f__Moraxellaceae; g__Acinetobacter; s__guillouiae	c49cc7c2c45bd7a87913453e515ea14f	416	0.32395474	0.32
<i>Propionibacterium</i>	f__Propionibacteriaceae; g__Propionibacterium; s__acnes	b02a8d33d018119dedb2db15db887bfd	166	0.129270401	0.13
<i>Bacteroides</i>	f__Bacteroidaceae; g__Bacteroides; s__	b6635d67cb594473d8ba9f8cbsa5d13d	26	0.020247171	0.02
<i>Deinococcus</i>	f__Deinococcaceae; g__Deinococcus; s__	e4c0868dfecdf2037ab5f7a071e4cc7	21	0.016353484	0.02
<i>Enterococcus</i>	f__Enterococcaceae; g__Enterococcus; s__	892a20bbdc3ce599dc0c5d9f0866c352	34	0.026477407	0.03
<i>Actinomyces</i>	f__Actinomycetaceae; g__Actinomyces; s__	c42488aff4cc842bf285a401dba39cc4	8	0.006292899	0.01
<i>Other</i>	p__OD1; c__o__f__g__s__	3be770f858d48a8c64d91c71fd000951	111	0.086439846	0.13
	p__OD1; c__o__f__g__s__	787fccde8e499d021ddb015a190b3b0	16	0.012459798	
	f__Comamonadaceae; g__Curvibacter; s__	9e4fcd82a0d6ac6109dac3d4f1a3409	25	0.019468434	
	f__Oxalobacteraceae; g__Ralstonia; s__	d026ba8391312c44726993268770b541	21	0.016353484	
			128413	100	100.00

Mock Community HM-783D, Sample ID MKOLS1946, Setup 1 (II)

Mock Community Member (L6)	QIIME 2 lowest classification level	ASV	Sequence Count	Relative Abundance %	L6 Relative Abundance %
<i>Escherichia</i>	f__Enterobacteriaceae	ffc36e27c82042664a16bcd4d380b286	26321	23.99165064	23.99
<i>Staphylococcus</i>	f__Staphylococcaceae; g__Staphylococcus	5497318e515a8c328a68f95975d9c7d4	24930	22.72375101	29.27
	f__Staphylococcaceae; g__Staphylococcus	908e9b387f6b9ce7d3f794e658fba37e	4529	4.128193676	
	f__Staphylococcaceae; g__Staphylococcus	04be702b895fe5ed0568344daf564276	2651	2.416392456	
<i>Streptococcus</i>	f__Streptococcaceae; g__Streptococcus; s__	e5af48ec2b6c023c5de28c59cb08a40	26478	24.13376549	27.39
	f__Streptococcaceae; g__Streptococcus; s__agalactiae	c3a3a503752209bc052b3995236b079f	3549	3.249241154	
	f__Streptococcaceae; g__Streptococcus	a3725fbb7f4a76528d54d283e88cad8	23	0.020964572	
<i>Rhodobacter</i>	f__Rhodobacteraceae; g__Rhodobacter; s__sphaeroides	332b70897316f7f62b81dfc53f41ca52	10448	9.523375475	9.52
<i>Bacillus</i>	f__Bacillaceae; g__Bacillus	8cb24777cb48dde0aac60dfeca125d10	3142	2.86394006	2.86
<i>Clostridium</i>	f__Clostridiaceae; g__Clostridium; s__butyricum	318669d5d926e9b81ca6911da00a14ea	3333	3.038036989	3.28
	f__Clostridiaceae; g__Clostridium; s__butyricum	88a3a8e95e3605bd051054937cde102	261	0.237902086	
<i>Pseudomonas</i>	f__Pseudomonadaceae	052ba7abaaa968c4f79e3f97d1f0a2f	1842	1.678987139	1.68
<i>Helicobacter</i>	f__Helicobacteraceae; g__Helicobacter; s__pylori	e8de432f2ada1078a2fda56ba92675a	534	0.4867422	0.49
<i>Neisseria</i>	f__Neisseriaceae; g__Neisseria; s__cinerea	f71e5aac4c976a1833958c2870b1d8b	339	0.308999262	0.31
<i>Listeria</i>	f__Listeriaceae; g__Listeria	b77b151f7481fd080bedbf415e736539	357	0.325406302	0.33
<i>Lactobacillus</i>	f__Lactobacillaceae; g__Lactobacillus; s__	9b23053de8f4269feb5ce286dfbf3c	218	0.19870749	0.20
<i>Acinetobacter</i>	f__Moraxellaceae; g__Acinetobacter; s__guillouiae	c49cc7c2c45bd7a87913453e515ea14f	319	0.290769217	0.29
<i>Propionibacterium</i>	f__Propionibacteriaceae; g__Propionibacterium; s__acnes	b02a8d33d018119dedb2db15db887bfd	240	0.218760539	0.22
<i>Bacteroides</i>	NA	NA	0	0.00	0.00
<i>Deinococcus</i>	f__Deinococcaceae; g__Deinococcus; s__	e4c0868dfecdf2037ab5f7a071e4cc7	39	0.035548588	0.04
<i>Enterococcus</i>	f__Enterococcaceae; g__Enterococcus; s__	892a20bbdc3ce599dc0c5d9f0866c352	23	0.020964552	0.02
<i>Actinomyces</i>	f__Actinomycetaceae; g__Actinomyces; s__	c42488aff4cc842bf285a401dba39cc4	7	0.006380516	0.01
<i>Other</i>	p__OD1; c__o__f__g__s__	3be770f858d48a8c64d91c71fd000951	108	0.98442243	0.11
	p__OD1; c__o__f__g__s__	787fccde8e499d021ddb015a190b3b0	18	0.01647047	
			109709	100	100.00

Mock Community HM-783D, Sample ID MKOLS2042, Setup 1 (III)

Mock Community Member (L6)	QIIME 2 lowest classification level	ASV	Sequence Count	Relative Abundance %	L6 Relative Abundance %
<i>Escherichia</i>	f__Enterobacteriaceae	ffc36e27c82042664a16bcd4d380b286	27844	25.20001448	25.20
<i>Staphylococcus</i>	f__Staphylococcaceae; g__Staphylococcus	5497318e515a8c328a68f95975d9c7d4	25520	23.09669478	29.66
	f__Staphylococcaceae; g__Staphylococcus	908e9b387f6b9ce7d3f794e658fba37e	4622	4.183108279	
	f__Staphylococcaceae; g__Staphylococcus	04be702b895fe5ed0568344daf564276	2634	2.383882996	
<i>Streptococcus</i>	f__Streptococcaceae; g__Streptococcus; s__	e5af48ec2b6c023c5de28c59cb08a40	26375	23.87050664	27.03
	f__Streptococcaceae; g__Streptococcus; s__agalactiae	c3a3a503752209bc052b3995236b079f	3454	3.126018173	
	f__Streptococcaceae; g__Streptococcus	a3725fbb7f4a76528d54d283e88cad8	36	0.032581544	
<i>Rhodobacter</i>	f__Rhodobacteraceae; g__Rhodobacter; s__sphaeroides	332b70897316f7f62b81dfc53f41ca52	10200	9.23143757	9.23
<i>Bacillus</i>	f__Bacillaceae; g__Bacillus	8cb24777cb48dde0aac60dfeca125d10	3257	2.94724722	2.95
<i>Clostridium</i>	f__Clostridiaceae; g__Clostridium; s__butyricum	318669d5d926e9b81ca6911da00a14ea	2250	2.036346523	2.19
	f__Clostridiaceae; g__Clostridium; s__butyricum	88a3a8e95e3605bd051054937cde102	172	0.155667379	
<i>Pseudomonas</i>	f__Pseudomonadaceae	052ba7abaaa968c4f79e3f97d1f0a2f	1939	1.754878181	1.75
<i>Helicobacter</i>	f__Helicobacteraceae; g__Helicobacter; s__pylori	e8de432f2ada1078a2fda56ba92675a	489	0.442565978	0.44
<i>Neisseria</i>	f__Neisseriaceae; g__Neisseria; s__cinerea	f71e5aac4c976a1833958c2870b1d8b	327	0.295949028	0.30
<i>Listeria</i>	f__Listeriaceae; g__Listeria	b77b151f7481fd080bedbf415e736539	337	0.304999457	0.30
<i>Lactobacillus</i>	f__Lactobacillaceae; g__Lactobacillus; s__	9b23053de8f4269feb5ce286dfbf3c	258	0.233501068	0.23
<i>Acinetobacter</i>	f__Moraxellaceae; g__Acinetobacter; s__guillouiae	c49cc7c2c45bd7a87913453e515ea14f	369	0.33396083	0.33

926
927

<i>Propionibacterium</i>	f__Propionibacteriaceae; g__Propionibacterium; s__acnes	b02a8d3d018119dedb2db15db887bfd	203	0,183723709	0,18
<i>Bacteroides</i>	f__Bacteroidaceae; g__Bacteroides; s__	b6635d67cb594473ddb9f8cfba5d13d	38	0,03439163	0,03
<i>Deinococcus</i>	f__Deinococcaceae; g__Deinococcus; s__	e4c0868fdefcd2037ab5f7a071e4cc7	31	0,0205633	0,03
<i>Enterococcus</i>	f__Enterococcaceae; g__Enterococcus; s__	892a20bbdc3ce599d0c5d9f0866c352	28	0,025341201	0,03
<i>Actinomyces</i>	NA	NA	0	0	0,00
<i>Other</i>	p__OD1; c__o__f__g__s__	3be770f858d48a8c64d91c71fd000951	94	0,085074033	0,10
	p__OD1; c__o__f__g__s__	787efccdb8e499d021ddb015a190b3b0	11	0,009955472	
	p__OD1; c__o__f__g__s__	35801c031a5311c7d870432585668de7	4	0,003620172	
			110492	100	100,00

Mock Community HM-783D, Sample ID MKOLS2138, Setup 1 (IV)

Mock Community Member (L6)	QIIME 2 lowest classification level	ASV	Sequence Count	Relative Abundance %	L6 Relative Abundance %
<i>Escherichia</i>	f__Enterobacteriaceae	ffc36e27c82042664a16bcc4d380b286	22060	25,98075587	26,65
	f__Enterobacteriaceae	b0728b5f5f391ce7f6f2c7f944a6afcd	570	0,671306929	
<i>Staphylococcus</i>	f__Staphylococcaceae; g__Staphylococcus	5497318e515a8c328a68f95975d9c7d4	19807	23,3273269	30,56
	f__Staphylococcaceae; g__Staphylococcus	908e9b387f6b9ce7d3f794e658fba37e	3471	4,087990576	
	f__Staphylococcaceae; g__Staphylococcus	04be702b895fe5ed0568344daf564276	2163	2,547433134	
	f__Staphylococcaceae; g__Staphylococcus	caf603c773ad5283866c506359242c7	505	0,594754384	
<i>Streptococcus</i>	f__Streptococcaceae; g__Streptococcus; s__	e5af48ec2b6c023c5de28c59cb08a40	18608	21,91522689	25,38
	f__Streptococcaceae; g__Streptococcus; s__agalactiae	c3a3a503752209bc052b3995236b079f	2477	2,917240811	
	f__Streptococcaceae; g__Streptococcus; s__	6fbb-c6275f0f1a6bd91628337b85b090	464	0,546467395	
<i>Rhodobacter</i>	f__Rhodobacteraceae; g__Rhodobacter; s__sphaeroides	332b70897316f7f62b81dfc53f41ca52	7404	8,719923683	8,94
	f__Rhodobacteraceae; g__Rhodobacter; s__sphaeroides	c5f93c01baff5679295e847f9c6b259	183	0,215524856	
<i>Bacillus</i>	f__Bacillaceae; g__Bacillus	8cb24777cb48d6e0aac60dfeca125d10	2417	2,846576924	2,85
<i>Clostridium</i>	f__Clostridiaceae; g__Clostridium; s__butyricum	31866945d926e9b81ca6911da0a14ea	1491	1,755997597	1,88
	f__Clostridiaceae; g__Clostridium; s__butyricum	88a3a8e95e3605bd051054f937cde102	108	0,127194997	
<i>Pseudomonas</i>	f__Pseudomonadaceae	052ba7abaea968c4f79e3f9d1f0a2f	1641	1,932657315	1,93
<i>Helicobacter</i>	f__Helicobacteraceae; g__Helicobacter; s__pylori	e8de9432f2ada1078a2fda56ba92675a	324	0,381584991	0,38
<i>Neisseria</i>	f__Neisseriaceae; g__Neisseria; s__cinerea	f71e5aac4c976a1833958c2870b1d8b	230	0,270878234	0,27
<i>Listeria</i>	f__Listeriaceae; g__Listeria	b77b151f7481fd080bedbf415e736539	274	0,322698418	0,32
<i>Lactobacillus</i>	f__Lactobacillaceae; g__Lactobacillus; s__	9b23053d68f4269f6b5ce286dfef3c	201	0,236724022	0,24
<i>Acinetobacter</i>	f__Moraxellaceae; g__Acinetobacter; s__guillouiae	c49cc7c2c45bd7a87913453e51ea14f	257	0,302676984	0,30
<i>Propionibacterium</i>	f__Propionibacteriaceae; g__Propionibacterium; s__acnes	b02a8d3d018119dedb2db15db887bfd	124	0,1460387	0,15
<i>Bacteroides</i>	f__Bacteroidaceae; g__Bacteroides; s__	b6635d67cb594473ddb9f8cfba5d13d	18	0,021199166	0,02
<i>Deinococcus</i>	f__Deinococcaceae; g__Deinococcus; s__	e4c0868fdefcd2037ab5f7a071e4cc7	19	0,022376898	0,02
<i>Enterococcus</i>	f__Enterococcaceae; g__Enterococcus; s__	892a20bbdc3ce599d0c5d9f0866c352	13	0,015310509	0,02
<i>Actinomyces</i>	NA	NA	0	0	0,00
<i>Other</i>	p__OD1; c__o__f__g__s__	3be770f858d48a8c64d91c71fd000951	58	0,068308424	0,09
	f__Enterobacteriaceae; g__Gluconacetobacter; s__	e8165c825d679874a9c71c16408fbbfd	11	0,012955046	
	c__ML6351-21; o__f__g__s__	6de3d71f0b5574f91e4569ad3168d64c	11	0,012955046	
			84909	100	100,00

Additional file 2

Mock Community HM-783D, Sample ID MKOLS2491, Setup 2

Mock Community Member (L6)	QIIME 2 lowest classification level	ASV	Sequence Count	Relative Abundance %	L6 Relative Abundance %
Escherichia	f_Enterobacteriaceae	d46e2205f0c6ecf67b51f83d111c509c	23274	22.50674506	22.54
	f_Enterobacteriaceae	6e93a9f573486663e97117b62fad86f2	34	0.03287915	
Staphylococcus	f_Staphylococcaceae; g_Staphylococcus	65d43491988bf5e57da4d86a5ba25dae	30851	29.83396029	29.88
	f_Staphylococcaceae; g_Staphylococcus	4008f0a6a397740091ad145f784080e5c	36	0.034813217	
Streptococcus	f_Staphylococcaceae; g_Staphylococcus	1776e0004f8ad79443ce2b037c69741	10	0.009670338	
	f_Streptococcaceae; g_Streptococcus	e7cfd084265c4df4856ca07b1c9b24ee	24279	23.47861405	26.20
Streptococcus	f_Streptococcaceae; g_Streptococcus; s_agalactiae	e9055f1b3a2ef5fe239567f02e0e758	2744	2.653540794	
	f_Streptococcaceae; g_Streptococcus	e9e8d4512233533218f96c9b2ecc2d8	35	0.033864184	
	f_Streptococcaceae; g_Streptococcus	02df8598ac39f1ef54609afb19c8c450	23	0.022141778	
	f_Streptococcaceae; g_Streptococcus; s__	5195aa3753b9257988f5339baa424e3	8	0.007736271	
	f_Streptococcaceae; g_Streptococcus	2082fcd6d3ed62054d6730e77350f1f8	2	0.001934068	
Rhodobacter	f_Rhodobacteraceae; g_Rhodobacter; s_sphaeroides	dfcc86cfa76e3e3d93e7eea450e6807	11372	10.99710857	11.00
Bacillus	f_Bacillaceae; g_Bacillus	bdf8a26094624622d68509a87fa75ba7	3261	3.153497278	3.15
Clostridium	f_Clostridiaceae; g_Clostridium; s_butyricum	4e8d7a662640b90817f015280cf5713	2549	2.46499622	2.64
	f_Clostridiaceae; g_Clostridium; s_butyricum	cb97fb83d4c8cc6cedd352a4ca3f8f	181	0.175033121	
Pseudomonas	f_Pseudomonadaceae; g_Pseudomonas	ff9d93d7b7e6787568f2d241caef3b	2194	2.121672195	2.12
Helicobacter	f_Helicobacteraceae; g_Helicobacter; s_pylori	e832be098a5318684958d14305267752	634	0.61309944	0.61
Neisseria	f_Neisseriaceae; g_Neisseria	8224351b2abd16dd4d58c3015f5e795	449	0.434198184	0.43
Listeria	f_Listeriaceae; g_Listeria	8ae518db29595b3f79214be0b589066	380	0.367427851	0.37
Lactobacillus	f_Lactobacillaceae; g_Lactobacillus; s__	0df6c802966e8670279671824da4f10a	357	0.345231023	0.35
Acinetobacter	f_Moraxellaceae; g_Acinetobacter; s_guillouiae	ea03646ed22d679fa586263d8fc32f	300	0.290110145	0.29
Propionibacterium	f_Propionibacteriaceae; g_Propionibacterium; s_acnes	5a7b179b154b50fe2282f260b07f360	302	0.292044213	0.29
Bacteroides	f_Bacteroidaceae; g_Bacteroides; s__	99deb3c5ecb022ec05609ebb1112a557	45	0.043516522	0.04
Deinococcus	f_Deinococcaceae; g_Deinococcus; s__	2385fe1c2dd5a3f8327272a6644088b	28	0.02076947	0.02
Enterococcus	f_Enterococcaceae; g_Enterococcus	9908ffab7ed4f3bec44cda2f5084d49	22	0.021274744	0.03
Actinomyces	f_Actinomycetaceae; g_Actinomyces; s__	df84561abbf84c8a2ca450b44378d4cb	7	0.006769237	0.01
Other	p_OD1; c__; f__; g__; s__	4479c551f476e1599f18ae69523c5395	18	0.017406609	0.03
	p_OD1	cb68de6534f0ba0aac84a0e027862c9	3	0.002901101	
	p_OD1; c__; f__; g__; s__	9cb241738bfe5cc7ed67aa803cddf70	2	0.001934068	
	f_mitochondria; g__; s__	7859d5f3e16e53f08178cb43bf95802	5	0.004835169	
	o_Actinomycetales	29f83e66700f19358051530ce2f68e96	4	0.003868135	
			103490	100	100.00

Additional file 3

Mock Community HM-783D, Sample ID MKOLS2830, Setup 3

Mock Community Member (L6)	QIIME 2 lowest classification level	ASV	Sequence Count	Relative Abundance %	L6 Relative Abundance %
Escherichia	f_Enterobacteriaceae	d46e2205f0c6ecf67b51f83d111c509c	39507	32.90248432	32.90
Staphylococcus	f_Staphylococcaceae; g_Staphylococcus	65d43491988bf5e57da4d86a5ba25dae	29989	24.97563982	24.98
Streptococcus	f_Streptococcaceae; g_Streptococcus	e7cfd084265c4df4856ca07b1c9b24ee	27912	23.24585877	25.81
	f_Streptococcaceae; g_Streptococcus; s_agalactiae	e3055f1b3a2ef5fe239567f02e0e758	2945	2.452674623	
Streptococcus	f_Streptococcaceae; g_Streptococcus	06f825b512d903b9230e1a55d87359ee	65	0.054133735	
	f_Streptococcaceae; g_Streptococcus; s__	edc9e5c16e40aff1eadce659794f0f8f	46	0.038310028	
Streptococcus	f_Streptococcaceae; g_Streptococcus; s__	fd496fd32dc8c08ade2e86c9d8ee13d	26	0.021653494	
	f_Rhodobacteraceae; g_Rhodobacter; s_sphaeroides	dfcc86cfa76e3e3d93e7eea450e6807	10534	8.772996427	8.77
Bacillus	f_Bacillaceae; g_Bacillus	bdf8a26094624622d68509a87fa75ba7	2992	2.491814778	2.49
Pseudomonas	f_Pseudomonadaceae; g_Pseudomonas	ff9d93d7b7e6787568f2d241caef3b	2430	2.023768874	2.02
Clostridium	f_Clostridiaceae; g_Clostridium; s_butyricum	4e8d7a662640b90817f015280cf5713	1923	1.601525739	1.69
Clostridium	f_Clostridiaceae; g_Clostridium; s_butyricum	cb97fb83d4c8cc6cedd352a4ca3f8f	110	0.091610937	
	f_Neisseriaceae; g_Neisseria	8224351b2abd16dd4d58c3015f5e795	442	0.3681094	0.39
Neisseria	f_Neisseriaceae; g_Neisseria	4f5ef25dacb5d639316e7291ff6f8b	21	0.017489361	
	f_Helicobacteraceae; g_Helicobacter; s_pylori	e832be098a5318684958d14305267752	317	0.264006063	0.26
Helicobacter	f_Helicobacteraceae; g_Helicobacter; s_pylori	e832be098a5318684958d14305267752	317	0.264006063	0.26
Listeria	f_Listeriaceae; g_Listeria	8ae518db29595b3f79214be0b589066	309	0.257343449	0.26
Listeria	f_Listeriaceae; g_Listeria	8ae518db29595b3f79214be0b589066	309	0.257343449	0.26
Lactobacillus	f_Lactobacillaceae; g_Lactobacillus; s__	0df6c802966e8670279671824da4f10a	220	0.183221873	0.18
Lactobacillus	f_Lactobacillaceae; g_Lactobacillus; s__	0df6c802966e8670279671824da4f10a	220	0.183221873	0.18
Acinetobacter	f_Moraxellaceae; g_Acinetobacter; s_guillouiae	ea03646ed22d679fa586263d8fc32f	144	0.119927044	0.12
Propionibacterium	NA	NA	0	0	0.00
Bacteroides	f_Bacteroidaceae; g_Bacteroides; s__	99deb3c5ecb022ec05609ebb1112a557	30	0.024984801	0.02
Deinococcus	f_Deinococcaceae; g_Deinococcus; s__	2385fe1c2dd5a3f8327272a6644088b	19	0.015823707	0.02
Enterococcus	NA	NA	0	0	0.00
Actinomyces	NA	NA	0	0	0.00
Other	f_Pasteurellaceae	394eda29c886632f514dd94b58381186	36	0.029981761	0.08
	f_Prevotellaceae; g_Prevotella; s_melaninogenica	32f8fd11d2bee78d609a1d4ab767554	36	0.029981761	
	f_Veillonellaceae; g_Veillonella; s_parvula	cd9401a6bce4a63af316d06d2a843f9d	20	0.016656534	
			120073	100	100.00

Additional file 4: Figure S.1

929
930
931
932
933

A. Setup 1

934

935

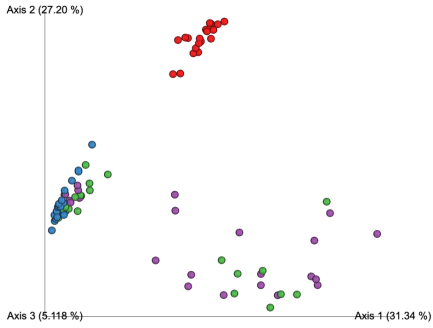
936

937

938

939

940



941

B. Setup 2

942

943

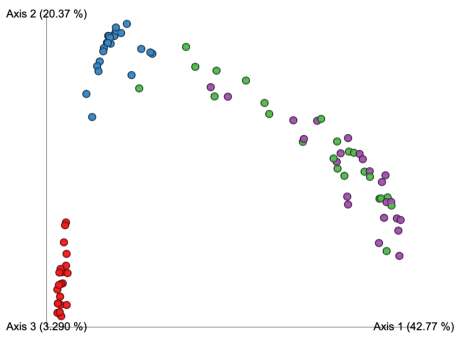
944

945

946

947

948



949

950

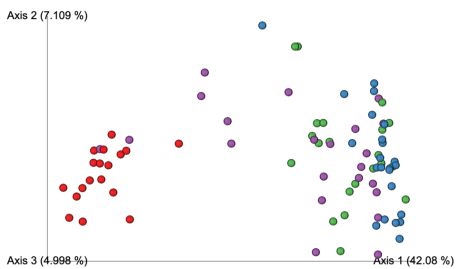
C. Setup 3

951

952

953

954



955

956 **Supplementary Figure 1.** Principal coordinates analysis on unweighted UniFrac distances for
957 procedural samples sequenced following each setup before the removal of Decontam
958 contaminants. A. Setup 1 (OW: n=22; PBAL: n=23; PSB: n=23; NCS: n=20). B. Setup 2 (OW:
959 n=23; PBAL: n=23; PSB: n= 23; NCS: n=21). C. Setup 3 (OW: n=23; PBAL: n=21; PSB: n=22;
960 NCS: n=18). Setup 1 (2-step PCR; V3 V4 region); Setup 2 (2-step PCR; V4 region); Setup 3 (1-
961 step PCR; V4 region). Rarefaction depth: 1066 sequences. Oral Wash (OW): blue; Protected
962 bronchoalveolar lavage (PBAL): green; Protected specimen brushes (PSB): purple; Negative
963 control samples (NCS): red.

964

965

966

967

968

969

970

971

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

990

991

992

993
994
995

996

997

998

999

1000

1001

1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

1021

1022

Additional file 5: Supplementary Methods

Online supplement for

Exploring protocol bias in airway microbiome studies: one versus two

PCR steps and 16S rRNA gene regions V3 V4 versus V4

Christine Drengenes, Tomas ML Eagan, Ingvild Haaland, Harald G Wiker, Rune Nielsen

1023

1024

1025 **Mock Community Sample HM-783D**

1026 The mock community sample was obtained through BEI Resources, NIAID, NIH, as part of the

1027 Human Microbiome Project: Genomic DNA from Microbial Mock Community B (Staggered,

1028 Low Concentration), v5.2L, for 16S rRNA Gene Sequencing, HM-783D. The input number of

1029 16S rRNA gene operons was given on the certificate of analysis provided by the BEI

1030 Resources, and used to calculate the relative abundance of the different bacteria in the

1031 sample (Table S.1).

1032

1033 Table S.1. Mock community HM-783D

Species	Number of operons	Relative abundance (%)
<i>Acinetobacter baumannii</i>	10000	0.22 %
<i>Actinomyces odontolyticus</i>	1000	0.02 %
<i>Bacillus cereus</i>	100000	2.19 %
<i>Bacteroides vulgatus</i>	1000	0.02 %
<i>Clostridium beijerinckii</i>	100000	2.19 %
<i>Deinococcus radiodurans</i>	1000	0.02 %
<i>Enterococcus faecalis</i>	1000	0.02 %
<i>Escherichia coli</i>	1000000	21.91 %
<i>Helicobacter pylori</i>	10000	0.22 %
<i>Lactobacillus gasseri</i>	10000	0.22 %
<i>Listeria monocytogenes</i>	10000	0.22 %
<i>Neisseria meningitidis</i>	10000	0.22 %
<i>Propionibacterium acnes</i>	10000	0.22 %
<i>Pseudomonas aeruginosa</i>	100000	2.19 %
<i>Rhodobacter sphaeroides</i>	1000000	21.91 %
<i>Staphylococcus aureus</i>	100000	2.19 %
<i>Staphylococcus epidermidis</i>	1000000	21.91 %
<i>Streptococcus agalactiae</i>	100000	2.19 %
<i>Streptococcus mutans</i>	1000000	21.91 %

<i>Streptococcus pneumoniae</i>	1000	0.02 %
---------------------------------	------	--------

1034

1035 **Library Preparation for MiSeq Sequencing (Setups 1, 2 and 3)**

1036 We compare three different library preparation setups for MiSeq sequencing of the bacterial
 1037 16S rRNA gene. The three setups vary with regards to the number of PCR steps (one or two)
 1038 and the target marker gene region sequenced (16S rRNA gene region V3 V4 or V4): Setup 1
 1039 (2-step PCR; region V3 V4); Setup 2 (2-step PCR; region V4); Setup 3 (1-step PCR; region V4).

1040 Setups 1 and 2

1041 Setups 1 and 2 were performed according to the the Illumina 16S Metagenomic Sequencing
 1042 Library Preparation guide (Part no. 15044223 Rev. B). The protocol consists of two PCR
 1043 steps; the first for amplification of the target marker gene region to be sequenced and the
 1044 second for the addition of index sequences required for sample multiplexing.

1045

1046 *Setup 1.* In the first PCR step, the 16S rRNA gene V3 V4 region was targeted using primers:

1047 5'-*TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG*CCTACGGGNGGCWGCAG-3' (forward) and

1048 5'-*GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG*GGACTACHVGGGTATCTAATCC-3' (reverse).

1049 Illumina overhang adapter sequences are *italicized*. Gene specific sequences (underlined)

1050 are taken from Klindworth *et al.* [1]. Each reaction consisted of 5 µl sample, 12.5 µl KAPA

1051 HiFi HotStart ReadyMix (2X) (KAPA Biosystems, USA), 0.5 µl of each primer (10 µM) and 6.5

1052 µl RT-PCR grade water (Thermo Fisher Scientific, USA) for a total volume of 25 µl. PCR cycling

1053 was performed using the following program: an initial cycle at 95 °C for 3 minutes, followed

1054 by 45 cycles of 95 °C for 30 seconds, 55 °C for 30 seconds, 72 °C for 30 seconds, and a final

1055 extension cycle at 72 °C for 5 minutes.

1056

1057 *Setup 2.* In the first PCR step, the 16S rRNA gene V4 region was targeted using primers:

1058 5'-*TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG*GTGCCAGCMGCCGCGGTAA-3' (forward) and

1059 5'-*GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG*GGACTACHVGGGTWTCTAAT-3' (reverse).

1060 Illumina overhang adapter sequences are *italicized*. Gene specific sequences (underlined)

1061 are taken from Caporaso *et al.* [2]. Each reaction consisted of 5 µl sample, 12.5 µl KAPA HiFi

1062 HotStart ReadyMix (2X), 1.25 µl of each primer (10 µM), and 5 µl RT-PCR grade water

1063 (Thermo Fisher Scientific, USA) for a total volume of 25 μ l. PCR cycling was performed using
1064 the following program: an initial cycle at 95 $^{\circ}$ C for 3 minutes, followed by 45 cycles of 95 $^{\circ}$ C
1065 for 30 seconds, 50 $^{\circ}$ C for 30 seconds, 72 $^{\circ}$ C for 30 seconds, and a final extension cycle at 72
1066 $^{\circ}$ C for 5 minutes.

1067

1068 For both setups 1 and 2, the second PCR step was performed using primers from the Nextera
1069 XT Index kit (Illumina Inc., USA). Each reaction consisted of 5 μ l amplicons from PCR step
1070 one, 25 μ l KAPA HiFi HotStart ReadyMix (2X), 5 μ l of each forward and reverse index primer
1071 (Nextera XT Kit), and 10 μ l RT-PCR grade water (Thermo Fisher Scientific, USA) for a total
1072 volume of 50 μ l. PCR cycling was performed using the following program: an initial cycle of
1073 95 $^{\circ}$ C for 3 minutes, followed by 8 cycles of 95 $^{\circ}$ C for 30 seconds, 55 $^{\circ}$ C for 30 seconds, 72 $^{\circ}$ C
1074 for 30 seconds, and a final extension cycle at 72 $^{\circ}$ C for 5 minutes.

1075

1076 Amplicon libraries were quantified using the Qubit dsDNA HS Assay Kit (Life Technologies,
1077 USA), normalized to 4 nM and pooled together. The pooled library was denatured with
1078 NaOH and diluted to 10 pM. The library was then spiked (15%) with PhiX from the PhiX
1079 Control Kit (Illumina). Paired-end sequencing was performed using 2x300 cycles (setup
1080 1)/2x275 cycles (setup 2) on the Illumina MiSeq using reagents from the MiSeq reagent kit
1081 v3 (Illumina).

1082

1083 Setup 3

1084 Setup 3 was based on the 1-step PCR protocol described by Kozich *et al.* [3]. The protocol
1085 consists of just one PCR step using primers that contain gene targeting sequences, index
1086 sequences and illumina sequencing adapter sequences.

1087

1088 The 16S rRNA gene V4 region was targeted using primers 5'-
1089 AATGATACGGCGACCACCGAGATCTA CACNNNNNNNTATGGTAATTGTGTGCCAGCMGCCGCGTAA-3'
1090 and 5'-CAAGCAGAAGACGGCATACGA GATNNNNNNNAGTCAGTCAGCCGGACTACHVGGGTWTCTAAT-
1091 3'. As detailed in Kozich *et al.* [3], the primers consist of different regions including: the Illumina
1092 sequencing adapter sequence, index sequence (NNNNNNNN), pad and linker sequence
1093 (reading 5'-3'). The gene specific sequences (underlined) are the same as for the primers

1094 used in the sequencing setup 2. Each reaction consisted of 5 µl sample, 18 µl AccuPrime Pfx
1095 SuperMix (Thermo Fisher Scientific, USA) and 1 µl of each primer (10 µM) for a total volume
1096 of 25 µl. PCR cycling was performed using the following program: an initial cycle at 95 °C for
1097 2 minutes, followed by 45 cycles of 95 °C for 20 seconds, 55 °C for 15 seconds, 72 °C for 5
1098 minutes, and a final extension cycle at 72 °C for 5 minutes. PCR clean-up was performed
1099 using Agencourt AMPure XP beads (Beckman Coulter, USA).

1100

1101 Amplicon libraries were quantified using the Qubit dsDNA HS Assay Kit, normalized to 4 nM
1102 and pooled together. The pooled library was denatured with NaOH and diluted to 10 pM.
1103 The library was spiked (15%) with PhiX from the PhiX Control Kit (Illumina). Paired-end
1104 sequencing was performed using 2x250 cycles on the Illumina MiSeq using reagents from
1105 the MiSeq reagent kit v3.

1106

1107

1108

1109 **References**

- 1110 1. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, et al. Evaluation of general
1111 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based
1112 diversity studies. *Nucleic Acids Res.* 2013;41:e1.
- 1113 2. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al.
1114 Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc*
1115 *Natl Acad Sci USA.* 2011;108 Suppl 1:4516–22.
- 1116 3. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a Dual-Index
1117 Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the
1118 MiSeq Illumina Sequencing Platform. *Appl Environ Microbiol.* 2013;79:5112–20.

1119

1120

1121

1122

1123

1124

1125



Graphic design: Communication Division, UIB / Print: Skjipes Kommunikasjon AS



uib.no

ISBN: 9788230847725 (print)
9788230843536 (PDF)