UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Lung microbiome of chronic obstructive pulmonary disease patients with and without HIV infection in Pretoria, South Africa

By

Tanweer Goolam Mahomed

Submitted in partial fulfilment for the degree

DOCTOR OF PHILOSOPHY
PhD (Medical Microbiology)

Department of Medical Microbiology
Faculty of Health Sciences
University of Pretoria
South Africa

December 2020

I, undersigned, declare that the dissertation hereby submitted to the University of Pretoria for the degree PhD (Medical Microbiology) and the work contained herein is my own original work and has not previously, in its entirety or in part, been submitted to any university for a degree. I further declare that all sources cited are acknowledged by means of a list of references

Signed_____this_____day of_____2020

In the name of Allah, the Most Beneficent, the Most Merciful.

ACKNOWLEDGEMENTS

I would like to thank the Lord for giving me strength and patience to finish my PhD degree.

I would like to kindly acknowledge and thank the following people:

My parents, especially my mother for support and continuing guidance, without you none of this would be possible.

My brothers, Ayaaz and Abdullah and my sister, Nusrat for keeping me sane and for the "encouragement".

My supervisor Prof MM Ehlers for her continuous guidance, support and steadfastness throughout the project

My co-supervisor Prof RPH Peters for his professional guidance and assistance throughout the project

Prof MM Kock, Prof V Ueckermann, Prof A Goolam Mahomed, Prof GHJ Pretorius, Dr M Allam, Dr A Ismail and the late Prof A Stoltz, for their contributions to the project and their support.

I would like to thank all my friends and colleagues, especially Arifa, Didi, Essa, Thabang, Humaira, Mpho, Lerato, Michelle, Hyun-Sul, Rashmika, Barend, Johnie, Sam and Karen for never letting me back down. I do not know what I would have done without you guys.

Lastly, I would like to thank all my other friends. There are too many of you for me to mention individually but I appreciate each and every one of you

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

SYMBOLS

α	Alpha
β-	Beta
%-	Percentage
°	Degrees
°C	degrees Celsius
µ	Micro

ABBREVIATIONS

µg	Microgram
µL	Microlitre
µm	Micrometre
µM	Micromolar
16S rRNA	16S ribosomal ribonucleic acid
23S rRNA	23S ribosomal ribonucleic acid
ACE	Abundance-base coverage
AECI/ ATI	Alveolar type I
AECII/ ATII	Alveolar type II
AECOPD	Acute exacerbation of chronic obstructive pulmonary disease
ANOSIM	Analysis of group similarities
ART	Antiretroviral therapy
ASL	Airway surface liquid
ATS	American Thoracic Society
BOLD	Burden of obstructive lung disease
CA	Correspondence analysis
CAT	Chronic obstructive pulmonary disease assessment test
CCA	Canonical correspondence analysis
CCorA	Canonical correlation analysis
cDNA	Complementary deoxyribonucleic acid
CDQ	Chronic obstructive pulmonary disease questionnaire
CF	Cystic fibrosis
CO	Carbon monoxide
CoNS	Coagulase-negative staphylococci
COPD	Chronic obstructive pulmonary disease
CT	Computed tomography
DCA	Detrended correspondence analysis

DFA/ LDA	Discriminatory function analysis
*dfr*A	Dihydrofolate reductase gene
DGGE	Denaturing gradient gel electrophoresis
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
dsDNA	Double-stranded deoxyribonucleic acid
dsRNA	Double-stranded ribonucleic acid
DTT	Dithiothreitol
ECM	Extracellular matrix
EDTA	Ethylenediaminetetraacetic acids
ERS	European Respiratory Society
FEF 25/75	The forced expiratory flow at 25% to 75%;
FEV-1%	Percentage of the forced vital capacity
FISH	Fluorescence in situ hybridisation
*fnb*A	Fibronectin binding protein A gene
FVC%	Forced vital capacity
GMP	Guanosine monophosphate
GOLD	Global Initiative for Obstructive Lung Disease
h	Hour
HCA	Hierarchical clustering
HCl	Hydrochloric acid
HIV	Human immunodeficiency virus
HS	Hotstart
HSRC	Human Sciences Research Council
IS	Intergenic spacer
IS-Pro	Intergenic spacer profiling
ITS	Internal transcribed spacer
J	Joining (regions)
kb	Kilo-base pair
kDA	Kilodalton
LABA	Long-acting inhaled beta-agonists
LAMA	Long-acting muscarinic antagonist
LLN	Lower limits of normal
LRT	Lower respiratory tract
Ltd	Limited
Mb	Mega-base pairs
mg	Milligram
MG-RAST	Metagenomics-rapid annotation using subsystems technology
min	Minute(s)
mL	Millilitre

mmol	Millimole
MMP1	Matrix metallopeptidase 1
MMP12	Matrix metallopeptidase 12
MMP9	Matrix metallopeptidase 9
mMRC	Modified Medical Research Council
M-PCR	Multiplex polymerase chain reaction
MST	Minimum spanning tree
NGS	Next-generation sequencing
NHLS	National Health Laboratory Services
NICD	National Institute of Communicable Disease
NK	Natural killer
nm	Nanometer
NMDS	Nonmetric multidimensional scaling
NRF	National Research Foundation
NTM	Non-tuberculous mycobacteria
OPLS-DA	Orthogonal projections to latent structure discriminant analysis
ORF	Open reading frame
OTU	Operational taxonomic unit
PA	Procrustes analysis
PCA	Principal component analysis
PCoA	Principal coordinate analysis
PCR	Polymerase chain reaction
PD	Phylogenetic diversity
PDE-4	Phosphodiesterase-4
PERMANOVA	Multivariate analysis of variance with permutation
pH	Power of hydrogen
QC	Quality control
QIIME	Quantitative insights into microbial ecology
QIIME 2	Quantitative insights into microbial ecology 2
Q-PCR	Quantitative polymerase chain reaction
RDA	Redundancy analysis
RDP	Ribosomal database project
RESCOM	Research Committee
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
RNS	Reactive nitrogen species
ROS	Reactive oxygen species
rRNA/ Rrn	Ribosomal ribonucleic acid
rRNA	Ribosomal ribonucleic acid
RSV	Respiratory syncytial virus

RT-PCR	Reverse transcriptase polymerase chain reaction
SABA	Short-acting inhaled beta-agonists
SAMA	Short-acting muscarinic antagonist
sec	Second
SISPA	Sequence-independent single primer amplification
SP-A	Surfactant protein A
SP-B	Surfactant protein B
SP-C	Surfactant protein C
spp.	Species (plural)
ssDNA	Single-stranded deoxyribonucleic acid
ssRNA	Single-stranded ribonucleic acid
SSU	Small subunit
TB	Tuberculosis
TBE	Tris(hydroxymethyl)aminomethane -borate- Ethylenediaminetetraacetic acid
TE	Tris(hydroxymethyl)aminomethane - Ethylenediaminetetraacetic acid
tet	Tetracycline resistance gene
*tet*K	Tetracycline efflux protein gene
T-RFLP	Terminal restriction fragment length polymorphism
UK	United Kingdom
UP	University of Pretoria
UPGMA	Unweighted pair group method with arithmetic mean
URT	Upper respiratory tract
USA	United States of America
VAMPS	Visualization and analysis of microbial population structures
W.A.T.E.R.S	Workflow for the alignment, taxonomy and ecology of ribosomal sequences
WGS	Whole-genome sequence
WHO	World Health Organization

LIST OF PUBLICATIONS AND CONFERENCE CONTRIBUTIONS

Publications

1. **Goolam Mahomed T,** Peters RPH and Ehlers MM (2020) Basic overview of the methods used in the statistical analysis of microbiome studies. Submitted for publication in *Applied and Environmental Ecology* Journal
2. **Goolam Mahomed T**, Peters RPH, Goolam Mahomed A, Ueckermann V, Kock MM and Ehlers MM (2020) Lung microbiome of stable and exacerbated COPD patients in Pretoria, South Africa. Submitted for publication to the *Scientific Reports* Journal
3. **Goolam Mahomed T**, Peters RPH, Pretorius GHJ, Goolam Mahomed A, Ueckermann V, Kock MM and Ehlers MM (2020) Comparison of targeted metagenomics and the IS-Pro method for analysing the lung microbiome. Submitted for publication to the *BMC Microbiology* Journal

Conference Presentation

1. **Goolam Mahomed T,** Peters RPH, Goolam Mahomed A, Ueckermann V, Kock MM and Ehlers MM (2019) The lung microbiome of stable and exacerbated COPD patients in Pretoria, South Africa. Presented at the 8th Federation of Infectious Diseases Societies of Southern Africa (FIDSSA) Congress 2019 from 7 November to 9 November in Johannesburg, South Africa (ePoster presentation)
2. **Goolam Mahomed T,** Peters RPH, Pretorius GHJ, Goolam Mahomed A, Ueckermann V, Stoltz A, Kock MM and Ehlers MM (2020) Determining the lung microbiome of chronic obstructive pulmonary disease patients from hospitals in Pretoria, South Africa using IS-Pro method and 16S rDNA sequencing. Presented at European Society of Clinical Microbiology and Infectious Diseases (ECMIDD) Congress from 18 April to 21 April 2020 in Paris. France (Poster presentation; conference was cancelled due to the COVID2019 virus and abstract was published in abstract book)

LUNG MICROBIOME OF CHRONIC OBSTRUCTIVE PULMONARY DISEASE PATIENTS WITH AND WITHOUT HIV INFECTION IN PRETORIA, SOUTH AFRICA

By

Tanweer Goolam Mahomed

DEGREE: **PhD (Medical Microbiology)**

PROMOTER: **Prof Marthie M Ehlers (Department of Medical Microbiology)**
DEPARTMENT: **Medical Microbiology, Faculty of Health Sciences, University of Pretoria/Tshwane Academic Division, National Health Laboratory Service**
CO-PROMOTER: **Prof Remco PH Peters**
DEPARTMENT: **Medical Microbiology, Faculty of Health Sciences, University of Pretoria/Department of Medical Microbiology, CAPHRI School for Public Health and Primary Care, Maastricht University Medical Centre+, Maastricht, The Netherlands/ Foundation for Professional Development, Research Unit, East London, South Africa**

SUMMARY

Chronic obstructive pulmonary disease (COPD) is a leading cause of death and is highly prevalent in South Africa (19% in adults over the age of 40 years). Inflammation of the lungs in COPD impairs the immune response and allows colonisation and infection with bacteria and viruses, that may cause exacerbations of the disease.

Culture-independent technologies have greatly increased the understanding of the lung microbiome. The most widely used method for targeted metagenomics is 16S rRNA sequencing. The IS-Pro (intergenic spacer profiling) method provides an alternative targeted metagenomics approach; however, the two methods have not been compared.

There is limited data on the microbiome in the lungs of COPD patients in Africa. Due to local environmental conditions, immunological differences and clinical comorbidities, such as HIV, the microbiome may be different from that reported in studies from other countries. The purpose of this study was to identify the lung microbiome and lung virome in COPD patients in South Africa and to determine if the COPD disease states result in differences in its composition. Next-generation sequencing was used to determine the microbiome and virome of COPD patients from hospitals in Pretoria, South Africa and the IS-Pro method was compared to targeted metagenomics.

Twenty-four patients over the age of 40 years with a confirmed COPD diagnosis and no *Mycobacterium tuberculosis* infection were included; eighteen were in the stable state of diseases and six were in the exacerbation state of disease. Sputum specimens were collected from all consenting participants and DNA and RNA were extracted directly from the specimens using commercial kits. The extracted bacterial DNA was sent for targeted metagenomics and the IS-Pro method and the extracted viral DNA and RNA were sent for shotgun metagenomics sequencing.

The lung of the COPD participants showed a diverse microbiome with over 77 genera identified and the *Firmicutes* phylum predominating. When the stable and exacerbation states of COPD disease were compared, no significant differences in the alpha and beta diversity between the disease states were observed. However, during exacerbation state of the disease, the abundance of key phyla had decreased. Analysis of the virome showed a high prevalence of BeAn 58058, a close relative of the smallpox virus, with bacteriophages being the second most prevalent viruses.

When comparing the IS-Pro method to targeted metagenomics, an increased relative abundance of *Proteobacteri*a with the IS-Pro method was observed, which was attributed to known lung pathogens, such as *Burkholderia.* The IS-Pro method was able to classify more operational taxonomic units (OTUs) to a species level, however, the unclassified OTUs from the IS-Pro method could only be classified to a phylum level.

To conclude, a diverse COPD microbiome was observed, with a virome that was dominated by the BeAn 58058 virus. The COPD disease states showed no variations in terms of diversity, however, the relative abundances of key phyla differed between disease states for the bacterial

microbiome. Future studies should focus on longitudinal studies of the sputum microbiome in an African setting as well as functional metatranscriptomics studies with a focus on antibiotic resistance and virulence factors.

506/500 words

CHAPTER 1

INTRODUCTION

1.1 Introduction

Chronic obstructive pulmonary disease (COPD) is a lung disease that is characterised by progressive airflow limitation (Simpson *et al.*, 2016). This disease is one of the world's leading causes of death, with the vast majority (90%) of deaths occurring in low- and middle-income countries (Lalloo *et al.*, 2016). Most of these deaths could be attributed to the South Asia region, with 81.2 deaths per 100 000 individuals attributed to COPD (Soriano *et al.*, 2020). South Africa ranks amongst the countries with the highest prevalence of COPD (>19% in adults over the age of 40 years), however, this information is over ten years old and as only one city was studied, the prevalence is not representative of the entire country (Buist *et al.*, 2007; Viviers and Van Zyl-Smit, 2015). The increased incidence of COPD risk factors in this city i.e. Cape Town suggest that this prevalence is higher than the general South African prevalence (Abdool-Gaffar *et al.*, 2019). Regardless, data suggest that the worldwide prevalence may increase in the coming years due to increased exposure to risk factors, such as smoking (not as important in South Africa; fewer people are smoking), indoor air pollution and genetic factors (van Gemert *et al.*, 2011). Additionally, in South Africa, other factors contribute to COPD prevalence such as tuberculosis (TB), exposure to mining and human immunodeficiency virus (HIV) (Allwood and van Zyl-Smit, 2015). In South Africa, the high burden of HIV (20.4% amongst adults between the ages 15 and 49 years old) increases the risk of TB and is associated with a decline in lung health (Lalloo *et al.*, 2016; UNAIDS, 2020). With the increased use of antiretroviral therapy (ART), HIV-positive individuals live longer and has a higher lifetime exposure risk to factors that contribute to COPD (Lalloo *et al.*, 2016). In South Africa, approximately 3.7 million people are using ART (PEPFAR, 2020).

Chronic obstructive pulmonary disease (COPD) is characterised by progressive airway obstruction (Lee *et al.*, 2016; Macnee *et al.*, 2016). Diagnosis of COPD is done using spirometry (to determine lung function) (Global Initiative for Obstructive Lung Disease, 2019). Spirometry is a method whereby the volume of air that a patient can expel from the lungs (after inhalation) is measured (Global Initiative for Obstructive Lung Disease, 2019). The forced expiratory volume in one second (FEV_1)/forced vital capacity (FVC) ratio, with a value below 0.7 is used to establish COPD diagnosis (Vogelmeier *et al.*, 2017). The spirometry is used to classify the

different stages of COPD as follows: i) mild/Global Initiative for Obstructive Lung Disease (GOLD) 1 (FEV_1% ≥80), ii) moderate/ GOLD 2 (FEV_1% between 50 and 79), iii) severe/ GOLD 3 (FEV_1% between 30 and 49) and iv) very severe/ GOLD 4 (FEV_1% <30) (Vogelmeier *et al.*, 2017). Differential diagnosis between COPD and other lung diseases is usually done through the use of chest computed tomography (CT) (Global Initiative for Obstructive Lung Disease, 2019).

One of the key features of COPD is the inflammation of the airways (Cullen and McClean, 2015; Fan *et al.*, 2016). Like other diseases causing airway inflammation, such as cystic fibrosis (CF), this inflammation facilitates colonisation of the lungs by microorganisms such as bacteria and viruses, partially due to impaired local immune response (Molyneaux *et al.*, 2013; Cullen and McClean, 2015). Inflammation of the lungs in COPD can cause bronchiolitis (by affecting the small airways), chronic bronchitis (by affecting the large airways) or emphysema (by affecting lung parenchyma) (MacNee, 2006; Macnee *et al.*, 2016; Global Initiative for Obstructive Lung Disease, 2019).

During COPD, there are points where the patients experience a worsened state of disease (Miravitlles and Anzueto, 2015). This worsened state can present as either respiratory or non-respiratory symptoms (such as fatigue and malaise) and is referred to as an exacerbation (Pavord *et al.*, 2016). These exacerbations are often triggered by a bacterial infection, viral infection or bacterial-viral co-infection (Aaron, 2014; Shimizu *et al.*, 2015; Bellinghausen *et al.*, 2016). The exacerbations caused by bacteria are often due to the acquisition of a new strain of the colonising bacteria entering the lung, e.g. a new *Pseudomonas aeruginosa* strain enters the lung that is already colonised with *P. aeruginosa*, causing an exacerbation (Aaron, 2014). Bacteria and viruses have been detected in stable COPD patient as well, however, the role that these microorganisms play in stable state COPD is unclear (Doring *et al.*, 2011; D'Anna *et al.*, 2016). To provide clarity on the issue of colonisation (the roles of bacteria in disease have not yet been elucidated) vs infection (cause inflammation and damage) in these patients (as well as other chronic lung diseases), Leung *et al.* (2017) have defined colonisation as the presence of microorganisms in the absence of infective symptoms.

While some viruses have been detected during the stable state of COPD, the majority of viruses have been detected as aetiological agents during exacerbations (D'Anna *et al.*, 2016). The most commonly isolated viruses (during exacerbations) are the rhinoviruses, however, other viruses

such as adenovirus, coronavirus, influenza viruses, metapneumovirus, parainfluenza virus and respiratory syncytial virus have been detected (Doring *et al.*, 2011; Cullen and McClean, 2015; D'Anna *et al.*, 2016). The majority of these viruses have been identified using virus-specific targeted polymerase chain reaction (PCR)-based techniques; these can only detect known viruses and as such the true viral community within the COPD lung may be unknown (Willner *et al.*, 2009).

The lung microbiome has been studied in a variety of patient groups including those with asthma, CF and HIV infection as well as in healthy individuals (Sze *et al.*, 2014; Boutin *et al.*, 2015; Huang and Boushey, 2015; Twigg *et al.*, 2017). The COPD lung microbiome has been investigated as well and studies have shown that the *Proteobacteria* phylum predominates in COPD lung, while the *Bacteroidetes* phylum predominates in healthy individuals (Sze *et al.*, 2014; Dickson and Huffnagle, 2015; Huang and Boushey, 2015). Studies of the lung microbiome of HIV infected individuals have shown an increased prevalence (53.7% of 82 HIV infected individuals across six research sites) of *Tropheryma whipplei* (a microorganism associated with the gastrointestinal tract), compared to HIV uninfected individuals (23.4% of 77 HIV uninfected individuals across six research sites) (Lozupone *et al.*, 2013; Twigg *et al.*, 2017). Studies that compared the lung microbiome in "healthy" HIV-positive and HIV-negative individuals in the absence of lung disease observed that the lung microbiome was indistinguishable between the two groups (Twigg *et al.*, 2017). However, as HIV infection progresses (in the absence of ART) a decrease in microbial diversity has been noted (Twigg *et al.*, 2017). This pattern of decreased microbial diversity in advanced stages of the disease has also been seen in advanced CF and COPD (Mammen and Sethi, 2016). Studies done on the exacerbation state of COPD infection have shown that there is no significant change in alpha diversity of the bacterial population in the COPD lung (Dickson *et al.*, 2014; Sze *et al.*, 2014; Mammen and Sethi, 2016). However, a change in abundance of certain phyla such as *Proteobacteria* was noted (Dickson *et al.*, 2014; Sze *et al.*, 2014; Sze *et al.*, 2015; Mammen and Sethi, 2016).

The microbiome in COPD and other disease states has been elucidated using technologies such as real-time PCR assays, restriction fragment length polymorphism and sequencing (both Sanger and next-generation sequencing) (Zakharkina *et al.*, 2013; D'Anna *et al.*, 2016). The majority of these methods target the 16S rRNA gene, a gene which is conserved across all bacteria (Williams, 2013; Zakharkina *et al.*, 2013; Dickson *et al.*, 2014; D'Anna *et al.*, 2016).

The hypervariable regions of the 16S rRNA gene are used, with hypervariable regions V1-V3 and V3-V5 being used most often (Cui *et al.*, 2014).

Another region which has been targeted in microbiome studies is the intergenic spacer (IS) region between the 16S rRNA and 23S rRNA genes (Budding *et al.*, 2016). This IS region is polymorphic and as it is present in all bacterial species, it makes it the target of choice for methods such as the IS-Pro (intergenic spacer profiling) method (Budding *et al.*, 2016). The IS-Pro method is a bacterial profiling method, which is based on the polymorphism in length and sequences of the IS region and can identify bacteria by comparing the profile generated against a reference database (Budding *et al.*, 2010; Budding *et al.*, 2016). This method can detect and identify bacterial species regardless if there are single-species or if part of a complex microbiome (Budding *et al.*, 2016). The advantage of this method is that it has a faster turnaround time and is less technically complex (more user-friendly) than targeted metagenomics (16S rRNA sequencing) (Budding *et al.*, 2016). The IS-Pro method has not been used to study the lung microbiome but has been used successfully to study faecal, intestinal, urogenital and vaginal microbiomes (Daniels *et al.*, 2014; de Meij *et al.*, 2016; Koedooder *et al.*, 2018; Koedooder *et al.*, 2019).

While, bacteria, fungi and viruses have been shown to influence the course of diseases, such as COPD, the majority of studies done on the "microbiome" have focused solely on the bacterial microbiome and not the fungal (mycobiome) and viral microbiome (also known as the virome) (Cabrera-Rubio *et al.*, 2012; Molyneaux *et al.*, 2013; Williams, 2013; Zakharkina *et al.*, 2013; Huang *et al.*, 2015; Sze *et al.*, 2015; D'Anna *et al.*, 2016).

While studies have been done on detecting viruses in COPD, most of these studies have been done using PCR and on only a select few viruses (Molyneaux *et al.*, 2013). The reason for this is that PCR requires prior knowledge of the sequence of the intended target (Wylie, 2017). The use of next-generation sequencing (NGS) removes this bias and can identify previously unknown viruses (Wylie, 2017). However, even with NGS, studying the virome can be challenging. One of the reasons is that viruses are more challenging to identify, in part due to a lack of a consensus sequence that can be used as a target for amplification (of the viral sequences); both bacteria and fungi have the 16S rDNA and the internal transcribed spacer (ITS) regions, respectively that are universal sequences that are present in all organisms (Williams, 2013; Wang, 2020). Two ways have been used to overcome this problem: i) shotgun

metagenomic sequencing i.e. sequencing all DNA within a sample, no matter its origin (whether bacterial, viral or human) and ii) purifying the viruses (before extraction) through size-filtration or density-screening (Williams, 2013). Another challenge is the diversity of the viruses that are capable of infecting humans, as viruses can be either DNA or RNA viruses (International Committee on Taxonomy of Viruses, 2011; Cadwell, 2015). The DNA viruses can be either double-stranded (dsDNA) viruses, single-stranded (ssDNA) viruses or reverse transcriptase DNA viruses (International Committee on Taxonomy of Viruses, 2011). The RNA viruses are complex that in addition to double-stranded (dsRNA) and reverse transcriptase RNA viruses, the single-stranded (ssRNA) viruses can be either positive sense or negative sense viruses (International Committee on Taxonomy of Viruses, 2011). The majority of viruses that infect COPD patients belong to the positive and negative sense ssRNA virus groups (Buss and Hurst, 2015). To ensure that all ssRNA viruses are sequenced using NGS, the ssRNA is converted to complementary DNA (cDNA) and then to double-stranded DNA (Lysholm *et al.*, 2012). While these methods have not been used to study the virome in COPD, they have been successfully used in CF (Lim *et al.*, 2013).

In South Africa, there is no data on the lung microbiome composition of COPD patients. Studies have been done on the COPD microbiome in countries such as Spain and the USA, however, the microbiome found in these other countries may not be the same as in South Africa (Cabrera-Rubio *et al.*, 2012; Dickson *et al.*, 2014; Sze *et al.*, 2015). Variables such as the local environmental conditions and other clinical comorbidities such as HIV and TB infection have the potential to affect the microbiome composition (Cabrera-Rubio *et al.*, 2012; Dickson *et al.*, 2014; Sze *et al.*, 2015). Several studies have focused on the lung microbiome in HIV patients, however, none of these studies recruited COPD patients, none of these studies were conducted on the African continent and none of them compared HIV-positive and HIV-negative patients (Williams *et al.*, 2016). Thus, it is unknown what potential effect the HIV status of a patient can have on the microbiome in COPD patients. The purpose of this novel study is to determine the effect that HIV status has on the lung microbiome during stable and exacerbation states of COPD and to determine the effect that viruses (the virome) have on the composition of the bacterial lung microbiome in these COPD patients.

1.2. Aim

The aim of this study was to identify and determine the variations in the lung microbiome (using next-generation sequencing and the IS-Pro method) and virome (using next-generation sequencing) in COPD patients with and without HIV infection in Pretoria, South Africa.

1.3. Objectives

The objectives of this research study were:

- To collect sputum specimens from COPD patients with and without HIV infection in stable and exacerbation states of disease from lung and HIV clinics at a tertiary academic hospital (20 individuals in each of the four groups)
- To determine and compare the composition of the bacteria present in the lung microbiome of COPD patients in the four groups, using a subset of a minimum of five patients per group, using next-generation sequencing
- To determine and compare the composition of the bacteria present in the lung microbiome of COPD patients in the four groups, using the IS-Pro method
- To determine if the virome has an effect on the lung microbiome composition in stable and exacerbation states of disease and in the context of HIV infection, using a subset of a minimum of five patients per group

References

Aaron, SD (2014). Management and prevention of exacerbations of COPD. *BMJ*, **349:** g5237.

Abdool-Gaffar, MS, Calligaro, G, Wong, ML, Smith, C, Lalloo, UG, Koegelenberg, CFN, Dheda, K, Allwood, BW, Goolam-Mahomed, A & Van Zyl-Smit, RN 2019. Management of chronic obstructive pulmonary disease-a position statement of the South African Thoracic Society: 2019 update. *J Thorac Dis*, **11:** 4408-4427.

Allwood, B & Van Zyl-Smit, RN (2015). Chronic obstructive pulmonary disease in South Africa: Under-recognized and undertreated. *S Afr Med J*, **105:** 785.

Bellinghausen, C, Gulraiz, F, Heinzmann, AC, Dentener, MA, Savelkoul, PH, Wouters, EF, Rohde, GG & Stassen, FR (2016). Exposure to common respiratory bacteria alters the airway epithelial response to subsequent viral infection. *Respir Res*, **17:** 68.

Boutin, S, Graeber, SY, Weitnauer, M, Panitz, J, Stahl, M, Clausznitzer, D, Kaderali, L, Einarsson, G, Tunney, MM, Elborn, JS, Mall, MA & Dalpke, AH (2015). Comparison of microbiomes from different niches of upper and lower airways in children and adolescents with cystic fibrosis. *PLoS One*, **10:** e0116029.

Budding, AE, Grasman, ME, Lin, F, Bogaards, JA, Soeltan-Kaersenhout, DJ, Vandenbroucke-Grauls, CM, Van Bodegraven, AA & Savelkoul, PH (2010). IS-Pro: High-throughput molecular fingerprinting of the intestinal microbiota. *FASEB J*, **24:** 4556-4564.

Budding, AE, Hoogewerf, M, Vandenbroucke-Grauls, CM & Savelkoul, PH (2016). Automated broad-range molecular detection of bacteria in clinical samples. *J Clin Microbiol*, **54:** 934-943.

Buist, AS, Mcburnie, MA, Vollmer, WM, Gillespie, S, Burney, P, Mannino, DM, Menezes, AMB, Sullivan, SD, Lee, TA, Weiss, KB, Jensen, RL, Marks, GB, Gulsvik, A & Nizankowska-Mogilnicka, E (2007). International variation in the prevalence of COPD (the BOLD study): A population-based prevalence study. *The Lancet*, **370:** 741-750.

Buss, L & Hurst, JR (2015). Viruses and exacerbations of chronic obstructive pulmonary disease: Unmet clinical need. *J Virus Erad*, **1:** 208-210.

Cabrera-Rubio, R, Garcia-Nunez, M, Seto, L, Anto, JM, Moya, A, Monso, E & Mira, A (2012). Microbiome diversity in the bronchial tracts of patients with chronic obstructive pulmonary disease. *J Clin Microbiol,* **50:** 3562-3568.

Cadwell, K (2015). The virome in host health and disease. *Immunity,* **42:** 805-813.

Cui, L, Morris, A, Huang, L, Beck, JM, Twigg, HL, 3rd, Von Mutius, E & Ghedin, E (2014). The microbiome and the lung. *Ann Am Thorac Soc,* **11 Suppl 4:** S227-232.

Cullen, L & Mcclean, S (2015). Bacterial adaptation during chronic respiratory infections. *Pathogens,* **4:** 66-89.

D'Anna, SE, Balbi, B, Cappello, F, Carone, M & Di Stefano, A (2016). Bacterial-viral load and the immune response in stable and exacerbated COPD: Significance and therapeutic prospects. *Int J Chron Obstruct Pulmon Dis,* **11:** 445-453.

Daniels, L, Budding, AE, De Korte, N, Eck, A, Bogaards, JA, Stockmann, HB, Consten, EC, Savelkoul, PH & Boermeester, MA (2014). Fecal microbiome analysis as a diagnostic test for diverticulitis. *Eur J Clin Microbiol Infect Dis,* **33:** 1927-1936.

de Meij, TG, Budding, AE, De Groot, EF, Jansen, FM, Frank Kneepkens, CM, Benninga, MA, Penders, J, Van Bodegraven, AA & Savelkoul, PH (2016). Composition and stability of intestinal microbiota of healthy children within a Dutch population. *FASEB J,* **30:** 1512-1522.

Dickson, RP & Huffnagle, GB (2015). The lung microbiome: New principles for respiratory bacteriology in health and disease. *PLoS Pathog,* **11:** e1004923.

Dickson, RP, Martinez, FJ & Huffnagle, GB (2014). The role of the microbiome in exacerbations of chronic lung diseases. *The Lancet,* **384:** 691-702.

Doring, G, Parameswaran, IG & Murphy, TF (2011). Differential adaptation of microbial pathogens to airways of patients with cystic fibrosis and chronic obstructive pulmonary disease. *FEMS Microbiol Rev,* **35:** 124-146.

Fan, VS, Gharib, SA, Martin, TR & Wurfel, MM (2016). COPD disease severity and innate immune response to pathogen-associated molecular patterns. *Int J Chron Obstruct Pulmon Dis,* **11:** 467-477.

Global Initiative for Obstructive Lung Disease. (2019). *2019 global strategy for the prevention, diagnosis and management of COPD* [Online]. [Accessed 07 January 2019].

Huang, YJ & Boushey, HA (2015). The microbiome in asthma. *J Allergy Clin Immunol,* **135:** 25-30.

Huang, YJ, Nariya, S, Harris, JM, Lynch, SV, Choy, DF, Arron, JR & Boushey, H (2015). The airway microbiome in patients with severe asthma: Associations with disease features and severity. *J Allergy Clin Immunol,* **136:** 874-884.

International Committee on Taxonomy of Viruses (2011). Virus taxonomy: Ninth report of the international committee on taxonomy of viruses, Elsevier Science.

Koedooder, R, Singer, M, Schoenmakers, S, Savelkoul, PHM, Morre, SA, De Jonge, JD, Poort, L, Cuypers, W, Beckers, NGM, Broekmans, FJM, Cohlen, BJ, Den Hartog, JE, Fleischer, K, Lambalk, CB, Smeenk, J, Budding, AE & Laven, JSE (2019). The vaginal microbiome as a predictor for outcome of in vitro fertilization with or without intracytoplasmic sperm injection: A prospective study. *Hum Reprod,* **34:** 1042-1054.

Koedooder, R, Singer, M, Schoenmakers, S, Savelkoul, PHM, Morre, SA, De Jonge, JD, Poort, L, Cuypers, WSS, Budding, AE, Laven, JSE & ReceptIVFity Study Group (2018). The ReceptIVFity cohort study protocol to validate the urogenital microbiome as predictor for IVF or IVF/ICSI outcome. *Reprod Health,* **15:** 202.

Lalloo, UG, Pillay, S, Mngqibisa, R, Abdool-Gaffar, S & Ambaram, A (2016). HIV and COPD: A conspiracy of risk factors. *Respirology,* **21:** 1166-1172.

Lee, SW, Kuan, CS, Wu, LS & Weng, JT (2016). Metagenome and metatranscriptome profiling of moderate and severe COPD sputum in Taiwanese Han males. *PLoS One,* **11:** e0159066.

Leung, JM, Tiew, PY, Mac Aogain, M, Budden, KF, Yong, VF, Thomas, SS, Pethe, K, Hansbro, PM & Chotirmall, SH (2017). The role of acute and chronic respiratory colonization and infections in the pathogenesis of COPD. *Respirology,* **22:** 634-650.

Lim, YW, Schmieder, R, Haynes, M, Willner, D, Furlan, M, Youle, M, Abbott, K, Edwards, R, Evangelista, J, Conrad, D & Rohwer, F (2013). Metagenomics and metatranscriptomics: Windows on CF-associated viral and microbial communities. *J Cyst Fibros,* **12:** 154-164.

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

Lysholm, F, Wetterbom, A, Lindau, C, Darban, H, Bjerkner, A, Fahlander, K, Lindberg, AM, Persson, B, Allander, T & Andersson, B (2012). Characterization of the viral microbiome in patients with severe lower respiratory tract infections, using metagenomic sequencing. *PLoS One*, **7:** e30875.

Macnee, W (2006). Pathology, pathogenesis, and pathophysiology. *BMJ*, **332:** 1202-1204.

Macnee, W, Vestbo, J & Agusti, A (2016). COPD: Pathogenesis and natural history. *In:* Broaddus, V. C., Mason, R. J., Ernst, J. D., King, T. E., Lazarus, S. C., Murray, J. F., Nadel, J. A., Slutsky, A. S. & Gotway, M. B. (eds.) Murray and Nadel's textbook of respiratory medicine (sixth edition). Sixth ed. Philadelphia: W.B. Saunders.

Mammen, MJ & Sethi, S (2016). COPD and the microbiome. *Respirology*, **21:** 590-599.

Miravitlles, M & Anzueto, A (2015). Antibiotic prophylaxis in COPD: Why, when, and for whom? *Pulm Pharmacol Ther*, **32:** 119-123.

Molyneaux, PL, Mallia, P, Cox, MJ, Footitt, J, Willis-Owen, SA, Homola, D, Trujillo-Torralbo, MB, Elkin, S, Kon, OM, Cookson, WO, Moffatt, MF & Johnston, SL (2013). Outgrowth of the bacterial airway microbiome after rhinovirus exacerbation of chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*, **188:** 1224-1231.

Pavord, ID, Jones, PW, Burgel, PR & Rabe, KF (2016). Exacerbations of COPD. *Int J Chron Obstruct Pulmon Dis*, **11 Spec Iss:** 21-30.

PEPFAR. (2020). *PEPFAR panarama spotlight* [Online]. Available: https://data.pepfar.gov/dashboards [Accessed 03 October 2020].

Shimizu, K, Yoshii, Y, Morozumi, M, Chiba, N, Ubukata, K, Uruga, H, Hanada, S, Saito, N, Kadota, T, Ito, S, Wakui, H, Takasaka, N, Minagawa, S, Kojima, J, Hara, H, Numata, T,

Kawaishi, M, Saito, K, Araya, J, Kaneko, Y, Nakayama, K, Kishi, K & Kuwano, K (2015). Pathogens in COPD exacerbations identified by comprehensive real-time PCR plus older methods. *Int J Chron Obstruct Pulmon Dis,* **10:** 2009-2016.

Simpson, JL, Baines, KJ, Horvat, JC, Essilfie, AT, Brown, AC, Tooze, M, Mcdonald, VM, Gibson, PG & Hansbro, PM (2016). COPD is characterized by increased detection of *Haemophilus influenzae*, *Streptococcus pneumoniae* and a deficiency of *Bacillus* species. *Respirology,* **21:** 697-704.

Soriano, JB, Kendrick, PJ, Paulson, KR, Gupta, V, Abrams, EM, Adedoyin, RA, Adhikari, TB, Advani, SM, Agrawal, A, Ahmadian, E, Alahdab, F, Aljunid, SM, Altirkawi, KA, Alvis-Guzman, N, Anber, NH, Andrei, CL, Anjomshoa, M, Ansari, F, Antó, JM, Arabloo, J, Athari, SM, Athari, SS, Awoke, N, Badawi, A, Banoub, JAM, Bennett, DA, Bensenor, IM, Berfield, KSS, Bernstein, RS, Bhattacharyya, K, Bijani, A, Brauer, M, Bukhman, G, Butt, ZA, Cámera, LA, Car, J, Carrero, JJ, Carvalho, F, Castañeda-Orjuela, CA, Choi, J-YJ, Christopher, DJ, Cohen, AJ, Dandona, L, Dandona, R, Dang, AK, Daryani, A, De Courten, B, Demeke, FM, Demoz, GT, De Neve, J-W, Desai, R, Dharmaratne, SD, Diaz, D, Douiri, A, Driscoll, TR, Duken, EE, Eftekhari, A, Elkout, H, Endries, AY, Fadhil, I, Faro, A, Farzadfar, F, Fernandes, E, Filip, I, Fischer, F, Foroutan, M, Garcia-Gordillo, MA, Gebre, AK, Gebremedhin, KB, Gebremeskel, GG, Gezae, KE, Ghoshal, AG, Gill, PS, Gillum, RF, Goudarzi, H, Guo, Y, Gupta, R, Hailu, GB, Hasanzadeh, A, Hassen, HY, Hay, SI, Hoang, CL, Hole, MK, Horita, N, Hosgood, HD, Hostiuc, M, Househ, M, Ilesanmi, OS, Ilic, MD, Irvani, SSN, Islam, SMS, Jakovljevic, M, Jamal, AA, Jha, RP, Jonas, JB, Kabir, Z, Kasaeian, A, Kasahun, GG, Kassa, GM, Kefale, AT, Kengne, AP, Khader, YS, Khafaie, MA, Khan, EA, Khan, J, Khubchandani, J, Kim, Y-E, Kim, YJ, Kisa, S, Kisa, A, Knibbs, LD, Komaki, H, Koul, PA, Koyanagi, A, Kumar, GA, Lan, Q, Lasrado, S, Lauriola, P, La Vecchia, C, Le, TT, Leigh, J, Levi, M, Li, S, Lopez, AD, Lotufo, PA, Madotto, F, Mahotra, NB, Majdan, M, Majeed, A, Malekzadeh, R, Mamun, AA, Manafi, N, Manafi, F, Mantovani, LG, Meharie, BG, Meles, HG, Meles, GG, Menezes, RG, Mestrovic, T, Miller, TR, Mini, GK, Mirrakhimov, EM, Moazen, B, Mohammad, KA, Mohammed, S, Mohebi, F, Mokdad, AH, Molokhia, M, Monasta, L, Moradi, M, Moradi, G, Morawska, L, Mousavi, SM, Musa, KI, Mustafa, G, Naderi, M, Naghavi, M, Naik, G, Nair, S, Nangia, V, Nansseu, JR, Nazari, J, Ndwandwe, DE, Negoi, RI, Nguyen, TH, Nguyen, CT, Nguyen, HLT, Nixon, MR, Ofori-Asenso, R, Ogbo, FA, Olagunju, AT, Olagunju, TO, Oren, E, Ortiz, JR, Owolabi, MO, P A, M, Pakhale, S, Pana, A, Panda-Jonas, S, Park, E-

K, Pham, HQ, Postma, MJ, Pourjafar, H, Poustchi, H, Radfar, A, Rafiei, A, Rahim, F, Rahman, MHU, Rahman, MA, Rawaf, S, Rawaf, DL, Rawal, L, Reiner Jr, RC, Reitsma, MB, Roever, L, Ronfani, L, Roro, EM, Roshandel, G, Rudd, KE, Sabde, YD, Sabour, S, Saddik, B, Safari, S, Saleem, K, Samy, AM, Santric-Milicevic, MM, Sao Jose, BP, Sartorius, B, Satpathy, M, Savic, M, Sawhney, M, Sepanlou, SG, Shaikh, MA, Sheikh, A, Shigematsu, M, Shirkoohi, R, Si, S, Siabani, S, Singh, V, Singh, JA, Soljak, M, Somayaji, R, Soofi, M, Soyiri, IN, Tefera, YM, Temsah, M-H, Tesfay, BE, Thakur, JS, Toma, AT, Tortajada-Girbés, M, Tran, KB, Tran, BX, Tudor Car, L, Ullah, I, Vacante, M, Valdez, PR, Van Boven, JFM, Vasankari, TJ, Veisani, Y, Violante, FS, Wagner, GR, Westerman, R, Wolfe, CDA, Wondafrash, DZ, Wondmieneh, AB, Yonemoto, N, Yoon, S-J, Zaidi, Z, Zamani, M, Zar, HJ, Zhang, Y & Vos, T (2020). Prevalence and attributable health burden of chronic respiratory diseases, 1990–2017: A systematic analysis for the global burden of disease study 2017. *The Lancet Respiratory Medicine,* **8:** 585-596.

Sze, MA, Dimitriu, PA, Suzuki, M, Mcdonough, JE, Campbell, JD, Brothers, JF, Erb-Downward, JR, Huffnagle, GB, Hayashi, S, Elliott, WM, Cooper, J, Sin, DD, Lenburg, ME, Spira, A, Mohn, WW & Hogg, JC (2015). Host response to the lung microbiome in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med,* **192:** 438-445.

Sze, MA, Hogg, JC & Sin, DD (2014). Bacterial microbiome of lungs in COPD. *Int J Chron Obstruct Pulmon Dis,* **9:** 229-238.

Twigg, HL, 3rd, Weinstock, GM & Knox, KS (2017). Lung microbiome in human immunodeficiency virus infection. *Transl Res,* **179:** 97-107.

UNAIDS. (2020). *South Africa* [Online]. Available: https://www.unaids.org/en/regionscountries/countries/southafrica [Accessed 03 October 2020].

van Gemert, F, Van Der Molen, T, Jones, R & Chavannes, N (2011). The impact of asthma and COPD in sub-Saharan Africa. *Prim Care Respir J,* **20:** 240-248.

Viviers, PJ & Van Zyl-Smit, RN 2015. Chronic obstructive pulmonary disease – diagnosis and classification of severity. South African Medical Journal, 105.

Vogelmeier, CF, Criner, GJ, Martinez, FJ, Anzueto, A, Barnes, PJ, Bourbeau, J, Celli, BR, Chen, R, Decramer, M, Fabbri, LM, Frith, P, Halpin, DM, Lopez Varela, MV, Nishimura, M,

Roche, N, Rodriguez-Roisin, R, Sin, DD, Singh, D, Stockley, R, Vestbo, J, Wedzicha, JA & Agusti, A (2017). Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease 2017 report: Gold executive summary. *Am J Respir Crit Care Med,* **195:** 557-582.

Wang, D (2020). 5 challenges in understanding the role of the virome in health and disease. *PLoS Pathog,* **16:** e1008318.

Williams, B, Landay, A & Presti, RM (2016). Microbiome alterations in HIV infection a review. *Cell Microbiol,* **18:** 645-651.

Williams, SC (2013). The other microbiome. *Proc Natl Acad Sci U S A,* **110:** 2682-2684.

Willner, D, Furlan, M, Haynes, M, Schmieder, R, Angly, FE, Silva, J, Tammadoni, S, Nosrat, B, Conrad, D & Rohwer, F (2009). Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS One,* **4:** e7370.

Wylie, KM (2017). The virome of the human respiratory tract. *Clin Chest Med,* **38:** 11-19.

Zakharkina, T, Heinzel, E, Koczulla, RA, Greulich, T, Rentz, K, Pauling, JK, Baumbach, J, Herrmann, M, Grunewald, C, Dienemann, H, Von Muller, L & Bals, R (2013). Analysis of the airway microbiota of healthy individuals and patients with chronic obstructive pulmonary disease by T-RFLP and clone sequencing. *PLoS One,* **8:** e68302.

CHAPTER 2

LITERATURE REVIEW

(Excerpts from Chapter 3 that was submitted as a review article can be found in Chapter 2)

2.1 Introduction

Microorganisms are ubiquitous and can be found everywhere (Barton and Northup, 2011). The study of these microorganisms and their environments has been termed microbial ecology (Barton and Northup, 2011). There are several different approaches (e.g. culture) to studying the microbial community structure in different environments, with molecular-based approaches being the most popular (Barton and Northup, 2011). According to Bikel *et al.* (2015), a microbiome can be considered as an ecological community (or an ecosystem) with multiple microorganisms interacting with each other and their environment. The composition of the (microbial) ecosystem in the human body is highly adaptive (changing as needed in response to outside influences), dependent on host genetics (as well as anatomical and physiological characteristics) and can be influenced by lifestyle choices (e.g. diet) and the environment (McDonald *et al.*, 2015; Lloyd-Price *et al.*, 2016; Marimón, 2018). Any change that disrupts the physiological microbial community, a term referred to as dysbiosis, can potentially influence the health of an individual and may cause a disease phenotype as a result (Lloyd-Price *et al.*, 2016).

Chronic obstructive pulmonary disease (COPD) is a disease that is influenced by microbiome alterations (Shukla *et al.*, 2017). The disease is characterised by persistent airway obstruction and inflammation of the lungs (Celli *et al.*, 2004; Global Initiative for Obstructive Lung Disease, 2019). Chronic obstructive pulmonary disease ranks as the fourth of the leading causes of deaths worldwide (Lopez-Campos *et al.*, 2016). Even though the African continent has a lower life expectancy (than Europe and the USA), due to risk factors including smoking, inhalation of biomass fuels fumes and HIV infection, which are prevalent in Africa, COPD can be considered a public health concern on the African continent (Adetunji and Bos, 2006; Sze *et al.*, 2012; Lalloo *et al.*, 2016; Macnee *et al.*, 2016; Pefura-Yone *et al.*, 2016).

This review aimed to increase the understanding of the human microbiome and COPD. Key areas of research that will be discussed including the different methods used to study the microbiome.

2.2 Overview of the human microbiome

The human microbiome can be defined as all the microorganisms present in and around the human body (including archaea, bacteria, fungi, protozoans and viruses) along with their genetic material (i.e. genomes) (Human Microbiome Project, 2012; Martin *et al.*, 2014; Marchesi and Ravel, 2015; Mammen and Sethi, 2016). However, this definition of the microbiome has been disputed, with the argument being that the environmental conditions surrounding a habitat (e.g. the human body) form part of the microbiome (Marchesi and Ravel, 2015). The argument is that the definition of a "biome" includes both biotic (living) and abiotic (non-living) factors (Marchesi and Ravel, 2015). Marchesi and Ravel (2015) argue that the microorganisms along with their genetic material should be referred to as the metagenome and only when this metagenome is combined with the environment should the term microbiome be used (Marchesi and Ravel, 2015). In this literature review, the term microbiome will be used to describe the microorganisms found in the human body.

The human body has been estimated to house over 10 trillion microbial cells, with bacterial cells predominating (Savage, 1977; Ursell *et al.*, 2012; Martin *et al.*, 2014). It has been postulated that the human microbiome has co-evolved with the human body, with these microorganisms performing essential functions for the human host, including the development of the immune system (Hooper and Gordon, 2001; Bäckhed *et al.*, 2005; Gill *et al.*, 2006; Martin *et al.*, 2014; Mammen and Sethi, 2016). It had been previously estimated that for every human cell in the human body, there are a million bacterial cells (Sze *et al.*, 2014). Recent data shows that it is a vast overestimation and the ratio is closer to 1:1, however, the number of bacteria in the human body still outnumber other microorganisms, such as fungi and archaea (Sender *et al.*, 2016). As a result, the term microbiome is often used to describe the bacterial cells and their genomes (the bacteriome) instead of the full microbiome (Zaura *et al.*, 2014).

The earliest microbiome studies have been attributed to Antonie van Leeuwenhoek in the 1680s (Porter, 1976; Ursell *et al.*, 2012). In these studies, van Leeuwenhoek observed and compared what he termed "animalcules" (Porter, 1976; Ursell *et al.*, 2012). He collected several different types of specimens including saliva, teeth scrapings and stool samples not only from himself

but from other people as well (van Leeuwenhoeck, 1677; van Leeuwenhoek, 1682; Leevvenhoeck, 1684; Dobell, 1920; Porter, 1976; Ursell *et al.*, 2012). Research of the human microbiome continued (focusing on the gastrointestinal tract), with several articles published in the 1970s (summarised in a review by Savage in 1977) (Savage, 1977; Goodrich *et al.*, 2014). While van Leeuwenhoek used microscopy to study the microbiome, these studies used culture techniques instead (Savage, 1977). In 1977, Woese and Fox were able to use 16S rRNA sequencing to differentiate bacteria phylogenetically, paving the way for future microbiome studies (Woese and Fox, 1977; Biteen *et al.*, 2016). Recent advances in DNA-based technologies resulted in the scope and scale of these projects increasing (Goodrich *et al.*, 2014; Mammen and Sethi, 2016). However, human microbiome projects focus primarily on the gastrointestinal tract (the gut) and there is no known consortium on the lung microbiome. Research into the lung microbiome has lagged behind other body sites (especially the gastrointestinal tract), in part due to a past hypothesis that stated that the healthy lung is a sterile body site and colonisation of the lung only occurs in disease (Charlson *et al.*, 2011; Mammen and Sethi, 2016). The first article published regarding the microbiome of the lung was in 2003 and focused on the lung disease, cystic fibrosis (CF) (Rogers *et al.*, 2003). In comparison, studies were done on the gut microbiome using molecular techniques as early as in 1996 (Wilson and Blitchington, 1996; Vaughan *et al.*, 2000; Zoetendal *et al.*, 2004).

2.3 Methods used to study the microbiome

Microbiomes, including the human microbiome, were previously studied using culture-based techniques (Mitchell and Glanville, 2018). These techniques have been proven to be unreliable as less than 1% of all bacteria can be cultured (Mammen and Sethi, 2016). However, Lau *et al.* (2016) were of the opinion that due to molecular methods being unable to distinguish between viable and non-viable cells, culturing of microorganism is still the best option (Lagier *et al.*, 2012; Lau *et al.*, 2016). These scientists proposed the use of enrichment steps and diverse culture conditions to elucidate the microbiome, using what has been termed “culturomics” (Lagier *et al.*, 2012). However, the use of culturomics, (especially in a diagnostic setting) can often be time-consuming and expensive, particularly for polymicrobial infections (Boase *et al.*, 2013). Additionally, culture-based methods are unable to detect viable but non-culturable bacteria (VNBC), fastidious microorganisms and viruses (that require tissue culture) using conventional culture media (Hodinka, 2013; Zhao *et al.*, 2017; Mobed *et al.*, 2019). Culture-independent methods, such as sequencing, on the other hand, are quicker (than culture-dependent methods; that take hours as opposed to days) and are becoming increasingly cheaper

and therefore may be more suitable in a diagnostic setting (Pallen *et al.*, 2010; Wang and Salazar, 2016). Culture-independent methods, used to study the microbiome can follow one of two approaches: i) a targeted approach (i.e. targeted metagenomics), where a specific region of the microbial genome is targeted, e.g. 16S rRNA or intergenic spacer (IS) region, or ii) a shotgun metagenomic approach, where all microbial genetic material is sequenced (Thurber *et al.*, 2009; Wommack *et al.*, 2012).

2.3.1 Targeted approach to study the microbiome

Using a targeted approach to study the microbiome is not a new concept and has been used since the late 20th century (in the 1970s and 1980s) (Hiergeist *et al.*, 2015). Fluorescence *in situ* hybridisation (FISH) was one of the first molecular techniques used to study the gut microbiome, by using specific probes to target a region of DNA such as the 16S rRNA region of bacteria in faecal samples (Franks *et al.*, 1998; Morgan and Huttenhower, 2012; Hiergeist *et al.*, 2015; Lloyd-Price *et al.*, 2016). Other approaches that have been used to study the microbiome include i) denaturing gradient gel electrophoresis (DGGE), ii) microarrays and iii) terminal restriction length polymorphisms (T-RFLP) (Table 2.1) (Hiergeist *et al.*, 2015; Huang *et al.*, 2017). However, the most popular approach to studying the microbiome is sequencing. (Hermann-Bank *et al.*, 2013; Hiergeist *et al.*, 2015; Hill *et al.*, 2016). Previously the Sanger sequencing method was used but this has now been replaced with next-generation sequencing technologies (NGS) (Hermann-Bank *et al.*, 2013; Hiergeist *et al.*, 2015; Hill *et al.*, 2016).

Table 2.1: Alternative methods to sequencing that have been used to study the microbiome

Method	Description	References
DGGE	PCR is used to target a specific region e.g. 16S rRNA with primers that have GC-rich tails and are run on a gel with a denaturing (chemical) gradient. The fragments will separate based on the %GC content and sequence; each band on the gel should correspond to a species	(Strathdee and Free, 2013; Hill *et al.*, 2016)
T-RFLP	Targets the 16S rRNA sequence as well, however, it utilises fluorescently labelled primers to target the sequence and is digested with restriction enzymes, followed by capillary electrophoresis	(Huang *et al.*, 2017)
Microarrays	Utilise fluorescent probes to target known sequences	(Hill *et al.*, 2016)
Quantitative PCR (qPCR)	Real-time PCR utilises probes to detect a fluorescence signal. The intensity of the signal is dependent on the amount of amplicon i.e. specific region of DNA that is targeted	(Hermann-Bank *et al.*, 2013; Hill *et al.*, 2016; Kralik and Ricchi, 2017)

DGGE: Denaturating gradient gel electrophoresis
DNA: Deoxyribonucleic acid
PCR: Polymerase chain reaction
RNA: Ribonucleic acid
T-RFLP: Terminal restriction length polymorphisms

Most of these culture-independent approaches (particularly NGS approaches) have focused on using the 16S ribosomal RNA (rRNA) gene region as a target (also known as the small subunit (SSU) rRNA) (Kembel *et al.*, 2012; Martin *et al.*, 2015). Ribosomal RNA is useful for determining phylogenetics as this protein is present in all forms of prokaryotic (as 16S rRNA) and eukaryotic organisms (as 18S rRNA) i.e. it is universal, is easily isolated and is highly conserved (i.e. the sequences and the length of the genes change very little with time) (Woese and Fox, 1977; Gürtler *et al.*, 2014; Hiergeist *et al.*, 2015). The 16S rRNA gene can be found as part of the ribosomal RNA (*rrn*) operon, together with 23S rRNA and 5S rRNA genes and intergenic spacer (ITS) regions (Gürtler *et al.*, 2014).

The 16S rRNA gene region is ideal for quantifying the microbiome as it has both conserved and hypervariable regions (Kembel *et al.*, 2012; Marsland *et al.*, 2013; Hiergeist *et al.*, 2015; Amato, 2017). To date, nine hypervariable regions within the 16S rRNA gene have been identified and are commonly referred to as V1-V9 (Mammen and Sethi, 2016; Nguyen *et al.*, 2016; Amato, 2017). None of these hypervariable regions can distinguish all bacteria (from each other), however, some show more promise than others (Tremblay *et al.*, 2015; Mammen and Sethi, 2016; Amato, 2017). Two sets of regions are popular and have been used in microbiome studies: i) The V1-V3 region and ii) V3-V4 region (region of choice for the Illumina platforms, as per the manufacturer's advice) (Tremblay *et al.*, 2015; Mammen and Sethi, 2016; Amato, 2017).

Primers are designed to bind to the conserved regions of DNA, however the amplicons produced need to span across the hypervariable regions to be discriminatory (Hiergeist *et al.*, 2015; Amato, 2017). Selecting which the primer pair should be used for a study is dependent on not only the coverage that the primer pair offers but also on the sequence length that is required (Parada *et al.*, 2016). The sequence length required is platform-dependent; e.g. PacBio (Pacific Biosciences, USA) can sequence the entire 16S rRNA gene (PacBio generates long reads up to 20 kb and the 16S rRNA gene is 1 550 bp) and may therefore use different primers than MiSeq (Illumina, USA), which is only able to run short reads (150 bp to 350 bp) (Clarridge, 2004; Rhoads and Au, 2015; Amato, 2017; Faner *et al.*, 2017; Pollock *et al.*, 2018). The incorrect selection of primers could lead to bias against a species or even an entire phylum (Klindworth *et al.*, 2013; Tremblay *et al.*, 2015). The choice of primers has a significant impact on a dataset and if different primers are used to study the same microbiome, different datasets (for each primer pair) may occur which can significantly impact the results (relative abundances

may be different or may cluster differently) to the point, where studies using different primer pairs cannot be compared to each other (Tremblay *et al.*, 2015; Hiergeist *et al.*, 2016).

Regardless of the advances, NGS has allowed a better understanding of the microbiome, however, its use in diagnostic settings are currently not feasible (Hamady and Knight, 2009; Budding *et al.*, 2016). While the cost of NGS has decreased significantly (due to newer technologies that can read more base-pair sequences in a single run and are more accurate), it is still relatively expensive to be used in a clinical diagnostic setting (as part of routine diagnostics), especially in resource-limited settings (Hamady and Knight, 2009; Budding *et al.*, 2016; Goodwin *et al.*, 2016; Boers *et al.*, 2019; Avila-Rios *et al.*, 2020). Additionally, NGS generates large amounts of data, which requires bioinformatics analysis by trained personnel and is time-consuming and can cause a delay in the time to results (Hamady and Knight, 2009; Budding *et al.*, 2016). Fingerprinting (or profiling) techniques provide an alternative solution to this problem by reducing cost and (sometimes) saving time (Daniels *et al.*, 2014). While there are several fingerprinting/profiling techniques available to study the microbiome including T-RFLP, none of these methods have been standardised and are often not reproducible between different researchers and different laboratories (Hamady and Knight, 2009; Eck *et al.*, 2017; Huang *et al.*, 2017).

Budding *et al.* (2010) developed a method termed the "IS-Pro" (intergenic spacer profiling) method to resolve the shortcomings of the available methods. The advantage of the IS-Pro method over other currently available methods is that it is standardised, reproducible, easy to use, doesn't require expensive equipment (to be purchased) and it is fast (results are available one day after uploading to the IS-Pro software) (Eck *et al.*, 2017). The IS-Pro method targets the intergenic spacer region that occurs between the 16S and 23S rRNA genes in the *rrn* operon (Gürtler *et al.*, 2014; Eck *et al.*, 2017). This region of the DNA is highly polymorphic (and yet species-specific) and the IS-Pro method uses the variation in length and sequence polymorphism to identify and differentiate the bacteria within a sample (Budding *et al.*, 2016). The IS-Pro method has been validated and successfully used to characterise the gastrointestinal (gut) microbiome in several disease states, where it is highly reproducible (Budding *et al.*, 2010; Budding *et al.*, 2014; Grasman *et al.*, 2014; Rutten *et al.*, 2015; Aguirre *et al.*, 2016; de Meij *et al.*, 2016a; de Meij *et al.*, 2016b; Janssens *et al.*, 2016; Eck *et al.*, 2017; Lankelma *et al.*, 2017; Muller *et al.*, 2017). Although the IS-Pro method has been used to characterise other

microbiomes, such as the vaginal microbiome, it has not been used before on sputum specimens to study the lung microbiome (Budding *et al.*, 2016; Koedooder *et al.*, 2018).

2.3.2 Metagenomics approach to study the microbiome

The term metagenome was first used by Handelsman *et al.* (1998) to describe the collective genomes of soil microorganisms. In this initial metagenomics study, DNA was isolated from bacteria, digested by restriction enzymes and cloned into vectors and screened for products of interest, such as antibiotics. Metagenomics has come a long way since these initial studies as the development of newer sequencing platforms (NGS) has resulted in higher throughput, cheaper cost per base sequencing and has resulted in the exclusion of the cloning step, thereby reducing time and money (Bragg and Tyson, 2014).

This metagenomic approach to sequence microbial communities has since become known as shotgun metagenomics and can be loosely defined as random sequencing of the total DNA from a microbial community (Bragg and Tyson, 2014; Amato, 2017). The first step to shotgun metagenomics is the same as with the targeted approach i.e. the extraction of DNA, which is followed by shearing of the DNA (Zhou *et al.*, 2015). The DNA can be sheared or fragmented using several different methods including restriction enzyme digestion and sonication (Zhou *et al.*, 2015). The DNA is ligated to adapters that act as priming sites for sequencing (van Dijk *et al.*, 2014; Zhou *et al.*, 2015). Next-generation sequencing will yield multiple short reads that are assembled and annotated using bioinformatic approaches (Zhou *et al.*, 2015).

Unlike bacteria, viruses lack a consensus sequence, making metagenomics an ideal approach to study viral diversity within an environment (i.e. virome) (Wylie *et al.*, 2012; Amato, 2017). Viruses are extremely diverse, differing in size and can be double-strand DNA (dsDNA) viruses, reverse transcriptase DNA viruses, single-stranded DNA (ssDNA) viruses, double-strand RNA (dsRNA) viruses, single-stranded RNA (ssRNA) viruses (both positive and negative sense) and reverse transcriptase viruses (The International Committee on Taxonomy of Viruses (ICTV), 2012; Cadwell, 2015). However, viral DNA obtained from total DNA of a sample represents less than 0.1% of the total DNA (due to the small size of viral genomes), even though viruses outnumber other microorganisms such as bacteria (for every microbial cell there are approximately 10 viruses) (Qin *et al.*, 2010; Bikel *et al.*, 2015; Amato, 2017). The best way to improve viral DNA and RNA isolation and to obtain adequate sequencing depth is to purify viral particles (VP) before extraction (Goodrich *et al.*, 2014; Bikel *et al.*, 2015).

Enriching the sample (for viral particles) can be done using physical means (filtration and/or density gradient centrifugation), enzymatic means (usually DNase) or non-specific amplification (Datta *et al.*, 2015; Kleiner *et al.*, 2015). Thereafter, ssDNA and ssRNA viruses need to be converted to dsDNA by using reverse transcriptase PCR (RT-PCR) to create cDNA (Lysholm *et al.*, 2012; Waugh *et al.*, 2015). The non-specific amplification procedure uses a single primer that is sequence-independent and was developed by Reyes and Kim (1991), this method is known as sequence-independent single primer amplification (SISPA) (Reyes and Kim, 1991; Datta *et al.*, 2015). The procedure has undergone several modifications, including the addition of DNase I treatment and the use of random primers for PCR amplification of DNA and RNA (Froussard, 1992; Allander *et al.*, 2001; Allander *et al.*, 2005; Lysholm *et al.*, 2012; Kallies *et al.*, 2019).

2.3.3 Analysis of microbiome data generated

Regardless of the method used to study the microbiome (or the virome), the data generated directly from NGS often requires additional analysis (Kuczynski *et al.*, 2011a). Next-generation sequencing platforms will generate an output file which is either a fastq file or a fasta file along with a qual file (Ju and Zhang, 2015). The fastq file contains a combination of the sequencing (i.e. nucleotide) data, like the fasta file and the quality score data associated with the sequencing data (which can be stored separately as a qual file) (Cock *et al.*, 2010). Since most NGS platforms can generate large amounts of sequences per run, it is often quicker and cheaper to run samples together in a single run (multiplex) (Di Bella *et al.*, 2013).

There are several pipelines which are available to study the microbiome (i.e. the bacterial microbiome) including metagenomics-rapid annotation using subsystems technology (MG-RAST), mothur, quantitative insights into microbial ecology (QIIME), QIIME2, the ribosomal database project (RDP) pyrosequencing tools, workflow for the alignment, taxonomy and ecology of ribosomal sequences (W.A.T.E.R.S) and visualization and analysis of microbial population structures (VAMPS) (Meyer *et al.*, 2008; Cole *et al.*, 2009; Schloss *et al.*, 2009; Caporaso *et al.*, 2010; Hartman *et al.*, 2010; Kuczynski *et al.*, 2011a; Ursell *et al.*, 2012; Huse *et al.*, 2014; Amato, 2017). The most frequently used of these pipelines are mothur and QIIME; due to their high accuracy, ability to identify operational taxonomic units (OTUs) to a genus level and their ability to use any reference database (Plummer and Twin, 2015; Bik, 2016; Amato, 2017; Almeida *et al.*, 2018).

Quantitative insights into microbial ecology (QIIME) is a python-based software that uses command-line prompts (Kuczynski *et al.*, 2011b; Ashton *et al.*, 2016; Lakhujani and Badapanda, 2017). The bioinformatics workflow for 16S rRNA gene analysis, using a program such as QIIME usually involves the following steps (or a variation thereof): i) creating a mapping file, ii) de-multiplexing, iii) quality filtering (including the removal of chimeras), iv) OTU picking, v) taxonomic assignment of OTUs, vi) construction of OTU table, vii) OTU filtering, viii) rarefaction and ix) diversity analysis (Caporaso *et al.*, 2010; Kuczynski *et al.*, 2011a; Kuczynski *et al.*, 2011b; McDonald *et al.*, 2012a; Morgan and Huttenhower, 2012; Navas-Molina *et al.*, 2013; Jervis-Bardy *et al.*, 2015; Ju and Zhang, 2015; Ashton *et al.*, 2016; Bik, 2016; Lakhujani and Badapanda, 2017).

The mapping file is a text file (.txt) that contains the sample name, a description of the sample, the barcodes and primers used and any metadata associated with the sample (Kuczynski *et al.*, 2011b; Navas-Molina *et al.*, 2013). This information is required for the processing of samples (for de-multiplexing and the removal of primers and barcodes) and subsequent analysis (e.g. β-diversity) (Kuczynski *et al.*, 2011b; Navas-Molina *et al.*, 2013). During de-multiplexing (a crucial step when multiple samples have been included in a single run), the sequences within the fastq file are "separated" and linked back to the relevant samples and the primers and barcode sequences are removed (Kuczynski *et al.*, 2011b; Navas-Molina *et al.*, 2013; Ju and Zhang, 2015). Quality filtering is applied to the sequencing reads to ensure that the downstream analysis is not affected e.g. diversity estimates may be inflated due to poor quality reads (Bokulich *et al.*, 2013; Kumar *et al.*, 2014; Ju and Zhang, 2015; Amato, 2017). Several criteria can be applied to the sequencing reads to improve their quality including i) removal of all sequences that are too short or too long i.e. sequence length, ii) removal of sequences with a certain length of homopolymers (a section of sequence that has the same (single) base repeated consecutively), iii) removal of ambiguous bases, iv) removal of chimeric sequences (sequences that have formed from the sequences of two or more microorganisms) and v) removal of bases with low quality (Phred) scores (Ju and Zhang, 2015; Amato, 2017). These Phred scores can be found in either the .qual file (which is associated with a particular fasta file) or form part of the fastq file (Cock *et al.*, 2010). The Phred score is the probability that the given base is incorrect and is usually denoted with a Q (Ewing and Green, 1998; Ewing *et al.*, 1998; Cock *et al.*, 2010; Bokulich *et al.*, 2013; Navas-Molina *et al.*, 2013; Lee *et al.*, 2016). The Phred score is calculated with the formula $Q = -10 \log_{10} P$; where Q is the quality value for the base and P is the probability that the base is incorrect (Ewing and Green, 1998; Ewing *et al.*, 1998; Cock *et al.*,

2010; Bokulich *et al.*, 2013; Navas-Molina *et al.*, 2013; Lee *et al.*, 2016). A Phred score of 10 can be interpreted that there is a 1 in 10 chance of an incorrect base and the accuracy of the base is 90% (Ewing and Green, 1998; Ewing *et al.*, 1998; Cock *et al.*, 2010; Bokulich *et al.*, 2013; Navas-Molina *et al.*, 2013; Lee *et al.*, 2016).

After the sequencing reads have been quality filtered, the next step is to cluster the sequences into OTUs (Navas-Molina *et al.*, 2013; Goodrich *et al.*, 2014; Ju and Zhang, 2015; Amato, 2017). Each OTU is equivalent to a microbial taxon and the level of sequence similarity denotes the taxonomic rank (Goodrich *et al.*, 2014; Franzen *et al.*, 2015). Even though there is no unified species definition/concept for bacterial species, a 97% sequence similarity of the 16S rRNA gene is typically used (Stackebrandt and Goebel, 1994; Konstantinidis *et al.*, 2006; Goodrich *et al.*, 2014; Kim *et al.*, 2014; Franzen *et al.*, 2015). There are three different approaches which can be used to cluster OTUs (also known as OTU picking): i) closed reference, ii) *de novo* reference or iii) open reference (Navas-Molina *et al.*, 2013; Goodrich *et al.*, 2014; Ju and Zhang, 2015; Amato, 2017).

The closed reference method clusters each sequence from the dataset against sequences in existing reference databases such as Greengenes, RDP or SILVA and an OTU is assigned if there is ≥97% identity (DeSantis *et al.*, 2006; Pruesse *et al.*, 2007; Cole *et al.*, 2009; McDonald *et al.*, 2012b; Navas-Molina *et al.*, 2013; Goodrich *et al.*, 2014; Amato, 2017). A disadvantage of this method is that it discards any sequence that fails to match against the chosen database, however, this method is faster than the others (Goodrich *et al.*, 2014; Amato, 2017). According to the *de novo* approach, sequences are grouped/clustered against each other based on sequence identity (97% identity) without using an external database (Navas-Molina *et al.*, 2013; Goodrich *et al.*, 2014; Amato, 2017). The open reference approach combines both the closed reference and *de novo* methods (Navas-Molina *et al.*, 2013; Goodrich *et al.*, 2014; Amato, 2017). Using the open reference approach, each sequence is matched against the reference database and if it matches, an OTU is assigned (Navas-Molina *et al.*, 2013; Goodrich *et al.*, 2014; Amato, 2017). However, if the sequence does not match it is clustered using the *de novo* approach (Navas-Molina *et al.*, 2013; Goodrich *et al.*, 2014; Amato, 2017). The open reference approach is the recommended approach, as it ensures that all sequences are kept (potentially new microorganism) and it is quicker than the *de novo* method (Navas-Molina *et al.*, 2013; Goodrich *et al.*, 2014; Amato, 2017).

Once OTUs have been picked, these clusters need to be assigned a taxon (Navas-Molina *et al.*, 2013; Goodrich *et al.*, 2014; Ju and Zhang, 2015; Amato, 2017). Using QIIME, a reference sequence (the default setting is to choose the most abundant sequence) is chosen for each OTU (Navas-Molina *et al.*, 2013). The way the taxa is assigned depends on the method used; the closed reference method assigns the taxa directly to each sequence during the OTU picking process from the database whereas with the *de novo* method the OTUs have to be assigned to a taxon using a reference dataset (such as the Greengenes database) after clustering (Navas-Molina *et al.*, 2013).

After the taxonomic assignment has occurred, an OTU table is constructed (Kuczynski *et al.*, 2011b; Navas-Molina *et al.*, 2013). This OTU table shows the abundance of each OTU within each sample in the dataset and is generated in a Biological Observation Matrix (BIOM) format (Kuczynski *et al.*, 2011b; McDonald *et al.*, 2012a; Navas-Molina *et al.*, 2013). A second quality filtering step, referred to as OTU filtering, is performed after the OTU table has been generated and it involves the removal of OTUs that are present in low numbers and any unwanted taxa, such as archaeal or host DNA (Navas-Molina *et al.*, 2013).

In a sequencing run, the number of sequences obtained (sequencing depth) can differ between samples for technical reasons and not biological reasons, which can affect diversity estimates (Goodrich *et al.*, 2014). To account for this variable sequence depth, a process termed rarefaction is applied (Goodrich *et al.*, 2014). In the rarefaction (also known as random sampling) approach, the dataset is normalised by randomly selecting the same amount of sequences from each sample (Goodrich *et al.*, 2014; Ju and Zhang, 2015). The final step in microbiome studies is to perform alpha (within sample) and beta (between sample) diversity analysis (Kuczynski *et al.*, 2011b; Morgan and Huttenhower, 2012; Navas-Molina *et al.*, 2013; Goodrich *et al.*, 2014; Ju and Zhang, 2015).

2.3.4 Statistics used in microbiome studies

There are two diversity measures of importance in microbiome studies: alpha diversity and beta diversity (Lozupone and Knight, 2008; Kuczynski *et al.*, 2011b; Morgan and Huttenhower, 2012; Navas-Molina *et al.*, 2013; Goodrich *et al.*, 2014; Ju and Zhang, 2015). Alpha diversity refers to the bacterial diversity within a single sample while beta diversity describes the diversity between samples (Knight *et al.*, 2018). The alpha diversity provides information on how complex a sample is, i.e. the more bacteria there is in a sample (higher alpha diversity),

the more interactions occur within the sample, whereas beta diversity shows how similar the different samples are to each other in terms of their bacterial composition (Mammen and Sethi, 2016; Stubbendieck *et al.*, 2016; Finotello *et al.*, 2018).

The research question determines which diversity measure(s) is appropriate for data analysis (Navas-Molina *et al.*, 2013). Selection of the appropriate measure(s) for analysis is based on the following study characteristics: i) is the aim of the study to test for alpha diversity or beta diversity? ii) is the presence/absence of particular taxa the only information required or is the abundance important? (qualitative measures vs quantitative measures) and iii) are all taxa regarded as equally related to each other or are the taxa considered divergently related, i.e. not all species are equally related to each other [species (taxon)-based measures vs divergent (phylogenetic)-based measures] (Lozupone and Knight, 2008; Hamady and Knight, 2009).

Alpha diversity measures provide information on how diverse a single sample is and this can be compared to other samples; it is useful when comparing a diseased individual to a healthy individual to determine if the diseased individual's microbiome is less or more diverse (Lozupone and Knight, 2008). However, even if two communities have similar alpha diversity measures, it does not mean that the two communities share the same taxa (Wagner *et al.*, 2018). Beta diversity measures show the number of shared species between communities (Lozupone and Knight, 2008). When deciding whether to use qualitative (presence/absence) or quantitative measures, the following points should be taken into consideration: i) quantitative measures are most useful when the data has a strong environmental filter (if subtle changes occur, qualitative measures are unable to take note of the difference) and ii) qualitative measures are most useful when rare species are present; with presence/absence data rare species are given the same weight as common species and as a result rare species are emphasised (Podani *et al.*, 2013; Jovel *et al.*, 2016). A phylogenetic approach would provide more evolutionary information; however, when studying a new environment, there may be a new taxon whose lineage has not been defined (Zaura, 2012; Chao *et al.*, 2016). In this instance, it would be more appropriate to use a taxon-based approach (Zaura, 2012; Chao *et al.*, 2016)

The most used statistical measures used for alpha diversity are Chao1, the Shannon index and the Simpson index (Morris *et al.*, 2014). According to Morris *et al.* (2014), an ideal alpha diversity measure does not exist and each alpha diversity measure interprets results differently, however, by using more than one alpha diversity measure, a more complete understanding of

the interactions within the community may be possible. Table 2.2 summarises the advantages and disadvantages of each statistical method to measure alpha diversity.

Table 2.2: Summary of characteristics of alpha diversity measures that can be used in microbiome studies

Statistical tool	Taxon/ Phylogenetic	Equations	Advantages	Disadvantages	References
Qualitative					
Chao1	Taxon	$S_{Chao1} = S_{obs} + \frac{n_1^2}{2n_2}$ where S_{obs} is the number of observed species, n_1 is the number of singletons (single reads) and n_2 is the number of doubletons	Precise	All species are regarded as equally related. Requires abundance data (e.g. OTU table)	(Chao, 1984; Hughes *et al.*, 2001; Lozupone and Knight, 2008; Magurran and McGill, 2010; Lemos *et al.*, 2011; Magurran, 2013; Ashton *et al.*, 2016)
Abundance-base coverage (ACE)	Taxon	$S_{ACE} = S_{abund} + \frac{S_{rare}}{C_{ACE}} + \frac{F_1}{C_{ACE}} \gamma_{ACE}^2$ where S_{abund} is the number of abundant species, S_{rare} is the number of rare species, $C_{ACE}=1-F_1/N_{rare}$ (F_1 is the number of species with i individuals) and $N_{rare} = \sum_{i=1}^{10} iF_i$	Considers both rare and abundant species	All species are regarded as equally related. Only provides information on the species observed	(Chazdon *et al.*, 1998; Hughes *et al.*, 2001; Lozupone and Knight, 2008; Lemos *et al.*, 2011; Magurran, 2013; Ashton *et al.*, 2016)
Phylogenetic Diversity (PD)	Phylogenetic	PD = (N-1) + no. of internal nodes of the minimum spanning path, where N is the size of the taxa	Provides both branch length and topographical information	Requires a phylogenetic tree; More weight is given to richness (over evenness); analysis is difficult with populations of different sample sizes	(Faith, 1992; Lozupone and Knight, 2008; Magurran, 2013; Lean and Maclaurin, 2016)

OTU: Operational taxonomic unit

Table 2.2: Summary of characteristics of alpha diversity measures that can be used in microbiome studies (continued)

Statistical tool	Taxon/ Phylogenetic	Equations	Advantages	Disadvantages	References
Quantitative					
Shannon's Index	Taxon	$H = -\sum_i p_i \ln p_i$; where p_i is the number of individuals in species s_i	Confounds species richness and evenness; sensitive to rarer species	All species are regarded as equally related; Sensitive to sample size; Values have no absolute meaning	(Shannon, 1984; Lozupone and Knight, 2008; Allen *et al.*, 2009; Lemos *et al.*, 2011; Daly *et al.*, 2018; Willis, 2019)
Simpson's Index	Taxon	$D = -\sum_i p_i^2$; where p_i is the number of individuals in species s_i	Suitable for smaller sample sizes; robust	All species are regarded as equally related; Requires abundance data; not intuitive; Values have no absolute meaning; does not account for unobserved species	(Simpson, 1949; Lozupone and Knight, 2008; Allen *et al.*, 2009; Lemos *et al.*, 2011; Magurran, 2013; Daly *et al.*, 2018; Willis, 2019)
Theta (θ)	Phylogenetic	$\theta(\pi) = \sum_{i=1}^{k}\sum_{j<i} p_i p_j d_{ij}$ where k is the number of distinct sequences, p_i is the frequency of the first (*i*th) sequence, p_j is the frequency of the second sequence (*j*th)and d_{ij} is the number of (nucleotide) differences between the two sequences	Provides a phylogenetic measurement	Richness is not considered	(Martin, 2002; Lozupone and Knight, 2008)
Jackknife	Unknown	$JACK1 = SO + \frac{r1(n-1)}{n}$; where SO is the number of species observed in *n* quadrants and *r*1 is the number of species present in one quadrant	Precise; useful in populations where there is resampling	Sensitive to sample size	(Heltshe and Forrester, 1983; Palmer, 1990; Morgan and Huttenhower, 2012; Magurran, 2013)

The Bray-Curtis (also known as Sorenson quantitative index), unweighted UniFrac and weighted UniFrac are the preferred statistical tools for measurement of beta diversity (Zhao *et al.*, 2015). Table 2.3 shows the various beta-diversity measures that can be used to study the microbiome and Figure 2.1 provides information on how to choose a beta diversity measure in the context of different study designs.

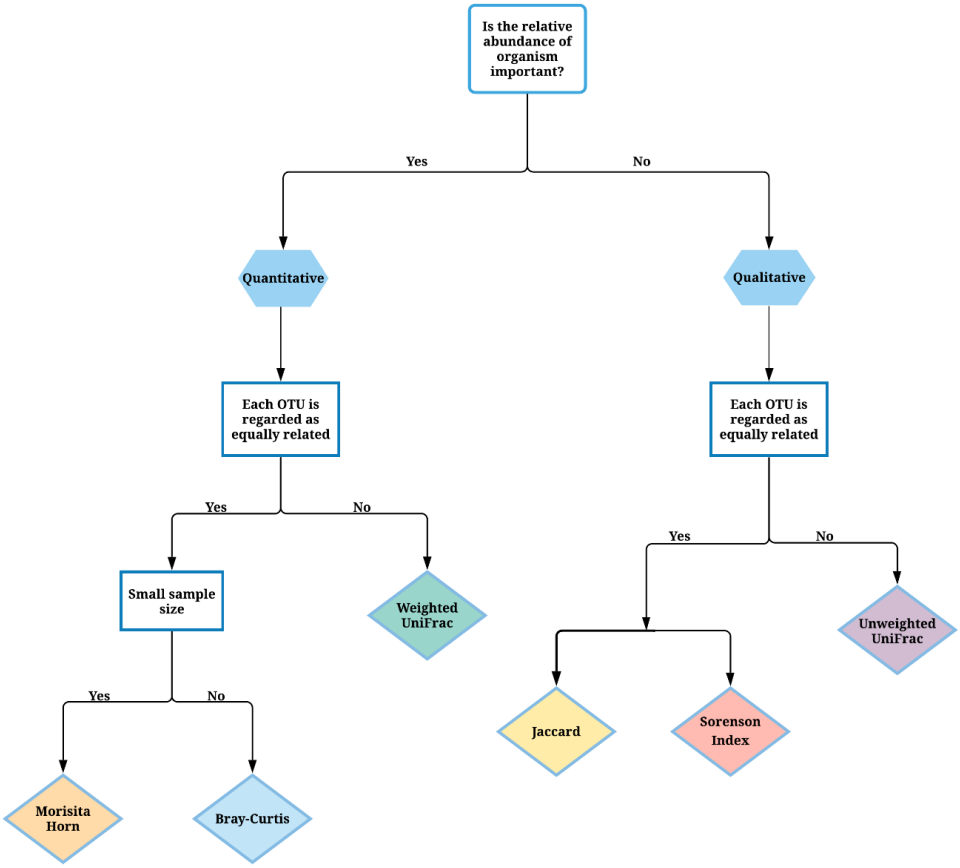

Figure 2.1: Algorithm to guide the choice of statistical measures to determine beta diversity in microbiome studies. Step 1 is choosing between a quantitative or a qualitative measure. Step 2 is deciding whether to consider the phylogenetic relationship between operational taxonomic units (OTUs). Other considerations, such as sample size, help inform the final decision on which measure to use (Koleff *et al.*, 2003; Chao *et al.*, 2006; Lozupone *et al.*, 2007; Lozupone and Knight, 2008; Magurran and McGill, 2010; Chang *et al.*, 2011; Lemos *et al.*, 2011; Evans and Matsen, 2012; Morgan and Huttenhower, 2012; Li *et al.*, 2013; Magurran, 2013; Rempala and Seweryn, 2013; Wong *et al.*, 2016; Xia and Sun, 2017; Wagner *et al.*, 2018).

Table 2.3: Summary of characteristics of beta diversity measures that are used in microbiome studies

Statistical tool	Taxon/ Phylogenetic	Equations	Input	Output (results)	Interpretation of results	Pros and Cons	References
Qualitative							
Sorenson Index/ Dice's coefficient	Taxon	$\boldsymbol{\beta_{sor}} = \frac{2a}{\alpha_1+\alpha_2}$; where a is the total number of species that occur in both populations, α_1 is the total number of species in population 1 and α_2 is the total number of species in population 2	OTU table	A value between 0 and 1	The closer the number is to one, the more similar the samples are	Pros: Simple and intuitive Cons: All species are regarded as equally related	(Sørensen, 1948; Koleff *et al.*, 2003; Chao *et al.*, 2006; Lozupone and Knight, 2008; Lemos *et al.*, 2011; Li *et al.*, 2013)
Jaccard	Taxon	$\boldsymbol{\beta_j} = \frac{a}{\alpha_1+\alpha_2-a}$; where a is the total number of species that occur in both populations, α_1 is the total number of species in population 1 and α_2 is the total number of species in population 2	OTU table	A value between 0 and 1	The closer the number is to one, the more similar the samples are	Pros: Simple and intuitive Cons: All species are regarded as equally related	(Jaccard, 1912; Koleff *et al.*, 2003; Chao *et al.*, 2006; Lozupone and Knight, 2008; Lemos *et al.*, 2011)
Unweighted UniFrac	Phylogenetic	$\boldsymbol{U} = \frac{\sum_i^n bi\lvert A_i-B_i\rvert}{\sum_i^n bi}$; where b_i is the branch length from branch *i*, A_i is the number of sequences/reads from branch *i* in population A and B_i is the number of sequences/reads from branch *i* in population B	Phylogenetic tree	A phylogenetic tree which indicates from which sample the sequences are from at the end of the node (from one sample, both samples, etc.)	If a node is shared between samples; the branch length will be shared indicating a similarity.	Pros: can compare samples from different conditions Cons: Gives to much weight to rare OTUs	(Lozupone and Knight, 2005; Lozupone *et al.*, 2007; Lozupone and Knight, 2008; Chang *et al.*, 2011; Xia and Sun, 2017)

OTU: Operational taxonomic unit

Table 2.3: Summary of characteristics of beta diversity measures that are used in microbiome studies (continued)

Statistical tool	Taxon/ Phylogenetic	Equations	Input	Output (results)	Interpretation of results	Pros and Cons	References
Quantitative							
Sorenson quantitative index/ Bray-Curtis Index	Taxon	$\boxed{BC_{ij} = \frac{S_i+S_j-C_{ij}}{S_i+S_j}}$; where S_i is the number of species in population *i*, S_j is the number of species in population *j* and C_{ij} is the total number of species (at the location with the fewest species)	OTU table	A value between 0 and 1	The closer the number is to one, the more similar the samples are	Pros: Robust Cons: sensitive to sample size; samples populations must be the same size	(Chao *et al.*, 2006; Lozupone and Knight, 2008; Magurran and McGill, 2010; Morgan and Huttenhower, 2012; Li *et al.*, 2013; Schroeder and Jenkins, 2018)
Morisita-Horn measures	Taxon	$\boxed{C_{MH} = \frac{2\sum_{i=1}^{s} pi_1 pi_2}{\sum_{i=1}^{s} pi_1^2 + \sum_{i=1}^{s} pi_2^2}}$; where pi_1 is the proportional abundance (percentage) of species in *i* in population 1 and pi_2 and pi_1 is the proportional abundance (percentage) of species in *i* in population 2	OTU table	A value between 0 and 1	The closer the number is to one, the more similar the samples are	Pros: Not sensitive to sample size Cons: can overlook rarer OTUs	(Morisita, 1959; Horn, 1966; Chao *et al.*, 2006; Lozupone and Knight, 2008; Magurran and McGill, 2010; Magurran, 2013; Rempala and Seweryn, 2013; Wagner *et al.*, 2018)
Weighted UniFrac	Phylogenetic	$\boxed{U = \sum_i^n bi \left\lvert \frac{A_i}{A_T} - \frac{B_i}{B_T} \right\rvert}$; where b_i is the branch length from branch *i*, A_i is the number of sequences/reads from branch *i* in population A, A_T is the total number of sequences/reads in population A, B_i is the number of sequences/reads from branch *i* in population B and B_T is the total number of sequences/reads in population B	Phylogenetic tree	A phylogenetic tree	A weight is given to the sequences based on their relative abundance. The width of the branch indicates the weight	Pros: can compare samples from different conditions Cons: Gives too much weight to more abundant OTUs	(Lozupone *et al.*, 2007; Lozupone and Knight, 2008; Evans and Matsen, 2012; Wong *et al.*, 2016; Xia and Sun, 2017)

OTU: Operational taxonomic unit

Beta diversity measures provide information on whether there are variations in the microbial composition between different populations or groups, but this measure is unable to identify the factors that are responsible for such variation (Tuomisto and Ruokolainen, 2006; Legendre, 2007). Variations between populations, if present, may be caused by i) biological interactions within the community, ii) environmental conditions (another variable) or iii) random variation (no known cause for the variation) (Legendre *et al.*, 2005). The best approach to understanding the variation in beta diversity is to perform multivariate analysis (Tuomisto and Ruokolainen, 2006).

Multivariate analysis of microbiome data can be performed in two ways: i) the distance-based approach that uses distance/dissimilarity matrices (beta diversity measures) such as the Bray-Curtis measure, or ii) the canonical approach that uses raw data i.e. OTU table (Legendre and Legendre, 2012; *GUSTA ME*, 2014; Buttigieg and Ramette, 2014). The distance-based approach is discussed in more detail below. The canonical approach uses the OTU table and requires that some assumptions be made on the relationship between the groups (linear, unimodal, etc.), i.e. how the data will be distributed (Ramette, 2007; Buttigieg and Ramette, 2014). Table 2.4 summarises the various distance-based and canonical multivariate tests that are available.

Table 2.4: Examples of multivariate tests to analyse microbiome data (Paliy and Shankar, 2016)

Test	Abbreviations	Raw data/Distance-based	Type of assumed relationship	Exploratory/ Interpretive/ Discriminatory	Ordination/ clustering
Principal coordinate analysis	PCoA	Distance-based	N/A	Exploratory	Ordination
Hierarchical clustering	HCA	Distance-based	N/A	Exploratory	Clustering
k-means clustering	N/A	Distance-based	N/A	Exploratory	Clustering
Nonmetric multidimensional scaling	NMDS	Distance-based	N/A	Exploratory	Ordination
Orthogonal projections to latent structure discriminant analysis	OPLS-DA	Raw data	Linear	Discriminatory	Ordination
redundancy analysis	RDA	Raw data	Linear	Interpretive	Ordination
Discriminatory function analysis	DFA/ LDA	Raw data	Linear	Discriminatory	Ordination
Canonical correlation analysis	CCorA	Raw data	Linear	Interpretive	Ordination
Canonical correspondence analysis	CCA	Raw data	Unimodal	Interpretive	Ordination
Principal component analysis	PCA	Raw data	Linear	Exploratory	Ordination
Correspondence analysis	CA	Raw data	Unimodal	Exploratory	Ordination
Detrended correspondence analysis	DCA	Raw data	Unimodal	Exploratory	Ordination
Procrustes analysis	PA	Any data	N/A	Interpretive	Ordination
*Hypothesis Tests**					
Multivariate analysis of variance with permutation	PERMANOVA	Distance-based	N/A	Interpretive	N/A
Analysis of group similarities	ANOSIM	Distance-based	N/A	Interpretive	N/A
Mantel test	N/A	Distance-based	N/A	Interpretive	N/A

N/A-Not applicable
*Hypothesis tests: used to test for significant differences between groups. Used after canonical (raw data) or distance-based approach.

In the distance-based approach, the first step is to ensure that all the data is in the same scale and format (Anderson, 2001; Ramette, 2007). This is achieved by standardising and normalising the data (Anderson, 2001; Ramette, 2007). The second step is to choose a distance measure to be used, e.g. Bray-Curtis (Anderson, 2001; Ramette, 2007). The third step is to visualise the similarity and dissimilarity between objects using cluster analysis or ordination (Anderson, 2001; Ramette, 2007). Patterns in a dataset may be observed using either cluster analysis or ordination (Anderson, 2001; Ramette, 2007). The more similar the samples are, the closer the samples will cluster (Frades and Matthiesen, 2010).

There are two types of multivariate clustering: hierarchical and *k*-means clustering (user-defined clustering; the user decides how many groups the data should be clustered into) (Anderson, 2001; Ramette, 2007; Buttigieg and Ramette, 2014). Hierarchical clustering is more appropriate for small datasets whereas *k*-means clustering is the most suitable tool for large datasets (Buttigieg and Ramette, 2014; Rodriguez *et al.*, 2019). There are several different

hierarchical clustering methods, including i) single-linkage clustering (also known as nearest neighbour clustering) e.g. minimum spanning tree (MST), ii) complete-linkage clustering and iii) average-linkage e.g. unweighted pair-group method with arithmetic mean (UPGMA) clustering (Legendre and Legendre, 2012; Buttigieg and Ramette, 2014). The user-defined method, *k*-means clustering uses an algorithm which requires three parameters from the user: i) the number of clusters, which is defined as *K*, ii) cluster initialisation (choosing initial clusters) and iii) a distance matrix (Khan and Ahmad, 2004; Ramette, 2007; Jain, 2010; Bai *et al.*, 2012; Buttigieg and Ramette, 2014).

The term ordination can be defined as "the arrangement of units in some order" (Legendre and Legendre, 2012). In ecology, ordination is used to visualise objects on reference axes (Ramette, 2007; Legendre and Legendre, 2012). Ideally, each descriptor in the study should be plotted as an axis; however, if there are more than three descriptors, it is not possible to be visualised on paper (Legendre and Legendre, 2012). As a result, the axes are chosen based on descriptors that the researchers are interested in (Legendre and Legendre, 2012). As the graph(s) represent the variability in a reduced space (dimensionally), these methods are referred to as ordination in reduced space (Legendre and Legendre, 2012). An example of an ordination method is principal coordinate analysis (PCoA) (Ramette, 2007; Legendre and Legendre, 2012; Buttigieg and Ramette, 2014; Paliy and Shankar, 2016). Clustering can be combined with ordination in a method called non-metric dimensional scaling (NMDS) (Ramette, 2007; Legendre and Legendre, 2012; Buttigieg and Ramette, 2014; Paliy and Shankar, 2016).

The last step in the distance approach (for multivariate analysis) is to test for the significant differences between the groups (Anderson, 2001; Ramette, 2007). Several test statistics can be used including analysis of similarities (ANOSIM), the Mantel test and permutational multivariate analysis of variance (PERMANOVA) (Ramette, 2007; Paliy and Shankar, 2016). The most popular test statistics is the PERMANOVA method, in part due to the fact it can be used in studies which have a small sample size (Tang *et al.*, 2016). Each of these methods test a different null hypothesis (Anderson and Walsh, 2013).

Choosing the appropriate approach (and tests) for multivariate analysis can be complicated for researchers who do not have a thorough understanding of statistical analytical methods and as such the risk of making the incorrect conclusions is higher (Buttigieg and Ramette, 2014). To help researchers understand multivariate analysis and to choose the right tools, Buttigieg and

Ramette (2014) developed an interactive website called GUSTA ME (https://sites.google.com/site/mb3gustame/home), that acts as a resource tool for microbial ecologist and other researchers studying the microbiome.

These statistical analyses (including α-diversity, β-diversity and multivariate analysis) can be performed using software tools (Hodkinson and Grice, 2015). There are several software tools currently available, including MATLAB (Hodkinson and Grice, 2015). One of the more popular tools is "R", an open-source software tool; which has several packages specific for microbiome data including "phyloseq", "picante" and "micropower" (Kembel *et al.*, 2010; McMurdie and Holmes, 2013; Navas-Molina *et al.*, 2013; Kelly *et al.*, 2015). In addition to the statistical analysis, these tools can also be used to visualise data (Navas-Molina *et al.*, 2013)

2.3.5 Visualisation of microbiome data

The data generated from microbiome studies is complex and multi-dimensional (Foster *et al.*, 2012). Most microbiome studies aim to understand a specific biological question or test a specific hypothesis; however, it is difficult to sort through all the different layers of information to answer these questions (Foster *et al.*, 2012). By using visualisation techniques, researchers can find patterns in the data and critically analyse and interpret the data (Foster *et al.*, 2012). However, due to the sheer number of data visualisation techniques, data visualisation in microbiome studies can be challenging (Foster *et al.*, 2012; Vazquez-Baeza *et al.*, 2013).

One of the first ways in which microbiome data can be visualised is in an OTU table (McDonald *et al.*, 2012a; Sedlar *et al.*, 2016). Most bioinformatics pipelines, including QIIME, create an OTU table (in a BIOM file) during the workflow process (Sedlar *et al.*, 2016; Dhariwal *et al.*, 2017). However, it is difficult to answer research questions based on the OTU table alone, especially in large datasets, as the information is presented in an unsorted tabular format (Dhariwal *et al.*, 2017). As a result, the OTU table is not used for the analysis itself but rather it is often a starting point for other visualisation techniques such as heat maps and Venn diagrams (Sedlar *et al.*, 2016). In QIIME, the OTU table can be used to summarise the relative abundance of each taxon (each taxon is shown as a percentage of the total taxa within the sample) in plots such as bar and pie graphs (Navas-Molina *et al.*, 2013; Huse *et al.*, 2014).

Alpha diversity measures are often depicted as box plots (box and whisker diagram) or as rarefaction curves (used more with Sanger sequencing) (Navas-Molina *et al.*, 2013; Dhariwal

et al., 2017). The beta diversity analysis is depicted using hierarchical clustering (as a dendrogram), non-metric multidimensional scaling (NMDS) or principal coordinate analysis (PCoA) (Legendre and Legendre, 2012; Navas-Molina *et al.*, 2013; Dhariwal *et al.*, 2017).

2.4 Factors that influence the microbial composition

The local environment within an anatomical site can affect the microbiome, with each site in the body having a unique microbiome (Weinstock, 2012; Zhou *et al.*, 2013; Taylor *et al.*, 2016). Other factors that influence the microbiome include: i) the growth (reproductive) rate of microorganism within the anatomical site, ii) the addition (immigration) of new microorganisms to the anatomical site and iii) the removal (elimination or extinction) of microorganisms from the anatomical site (Dickson *et al.*, 2015a).

The growth of the microorganisms within a local environment (within the human body) is dependent on physiochemical factors, such as temperature, oxygen tension, pH and nutrient supply (Dickson *et al.*, 2015a; Lopes *et al.*, 2015; Taylor *et al.*, 2016). Interactions between host epithelial cells, the concentration of inflammatory cells (e.g. alveolar macrophages) and interspecies interactions, such as competition with other microorganisms, in the local environment can affect the growth rate as well (Venkataraman *et al.*, 2015; Dickson *et al.*, 2016). Surfactant, which is produced by alveolar type II cells (AECII/ATII cells) can inhibit growth, due to its antibacterial properties (Palange and Simonds, 2013; Haghi *et al.*, 2014; Standring, 2015; Adar *et al.*, 2016). The pulmonary surfactant consists of phospholipids and proteins (mostly surfactant proteins SP-A, SP-B and SP-C) and is recycled by AECII cells or removed by alveolar macrophages (Palange and Simonds, 2013; Haghi *et al.*, 2014; Standring, 2015).

Microorganisms can enter (immigrate to) the lung from either inhalation or microaspiration (O'Dwyer *et al.*, 2016). Inhalation is the process of air intake into the respiratory system (Palange and Simonds, 2013). Air that is breathed in from the environment contains, along with microorganisms, gases (both oxygen and toxic gases) and particles (Nicod, 2005; Simkhovich *et al.*, 2008; Palange and Simonds, 2013). Factors such as geography and climate can affect the microbiome (Beck *et al.*, 2012; Kim *et al.*, 2017; Twigg *et al.*, 2017). The type of air particles, gases and microorganisms found is dependent on the local environment, for example, the air from an urban environment may contain higher levels of metal particles (from exhaust fumes, etc.) than the air from a rural environment (Simkhovich *et al.*, 2008; Mateos *et al.*, 2018).

Microaspiration, on the other hand, is when a small volume of matter from the gastrointestinal tract or oropharynx is inhaled into the respiratory tract (Lee *et al.*, 2010). It is through this process (microaspiration) that microorganisms from the gastrointestinal tract and oral cavity can enter the respiratory tract and contribute to the lung microbiome (Lee *et al.*, 2010; Beck *et al.*, 2012; Budden *et al.*, 2017; Chung, 2017).

When particles, such as microorganisms enter the lung (through breathing in), these are cleared through mucociliary clearance, the primary defence mechanism of the lung (see Figure 2.2) (Dickson *et al.*, 2016; O'Riordan and Smaldone, 2016; Bustamante-Marin and Ostrowski, 2017). The airways in a healthy individual have two layers, ciliated epithelial cells and an airway surface layer, which is sub-divided into two layers, a mucus layer and low viscosity periciliary layer (facilitates ciliary beating) (Bustamante-Marin and Ostrowski, 2017). The particles are trapped in the mucus layer and transported up the trachea by the beating cilia (Dickey *et al.*, 2015; Bustamante-Marin and Ostrowski, 2017). The particles i.e. microorganisms are either coughed up (exiting the human body) or are swallowed (enter the gastrointestinal tract) (Dickey *et al.*, 2015; Bustamante-Marin and Ostrowski, 2017). The other way in which microorganisms are eliminated from the lung is through innate and adaptive immune defences (Dickson *et al.*, 2016).

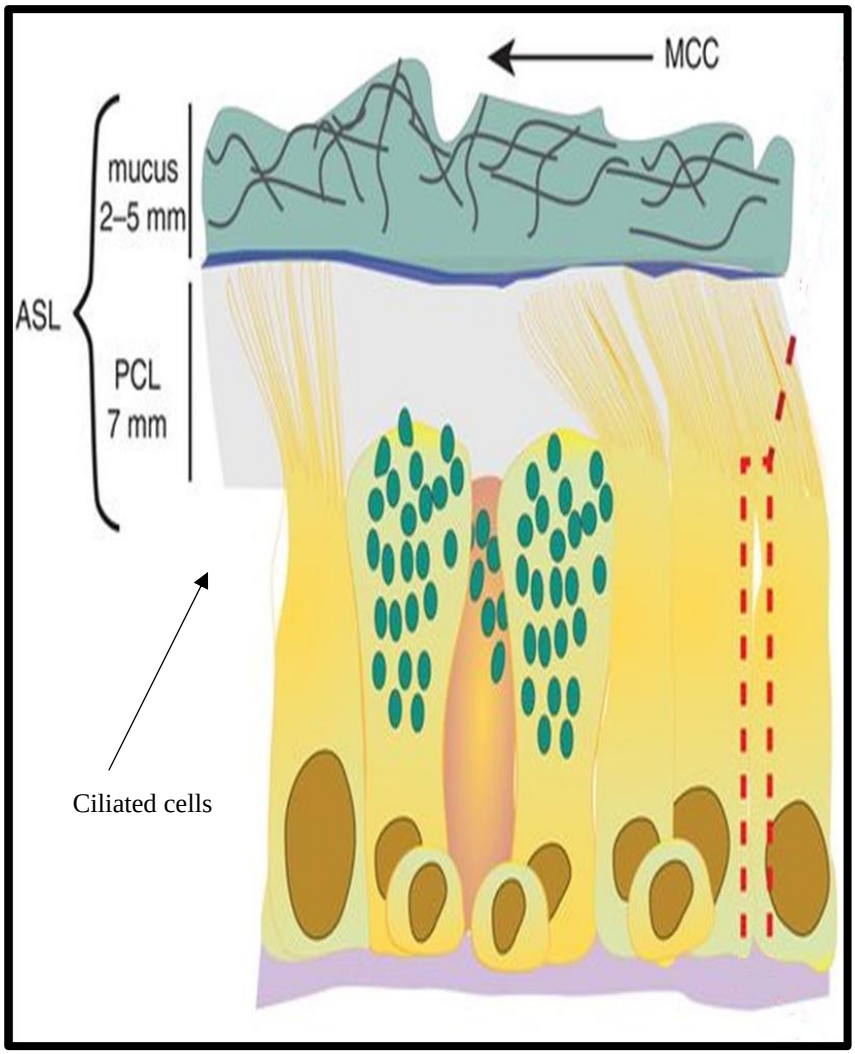

Figure 2.2: A diagrammatical representation of mucociliary clearance components (MCC). The airway surface liquid (ASL) layer is divided into a mucus layer (mobile) in the top and periciliary layer (stationary) on the bottom. The ciliated cells are present as part of the periciliary layer (PCL) as well as below it. In some instances, a surfactant layer (shown in blue below the mucus layer) is present (Bustamante-Marin and Ostrowski, 2017).

Additionally, the microbiome of the gastrointestinal tract (gut) can affect the lung microbiome and *vice versa*, via the gut-lung axis (O'Dwyer *et al.*, 2016; Taylor *et al.*, 2016; Budden *et al.*, 2017). The idea is that there is cross-talk between the gastrointestinal tract and lungs and that immunological changes in one may affect the other (Marsland *et al.*, 2015; Budden *et al.*, 2017). Additionally, the gastrointestinal tract may act as a source of metabolites for the lung (Marsland *et al.*, 2015). There have been studies that suggest that dietary changes, such as the addition of fibre (and its metabolites), can change the lung microbiome by affecting the immunological

component in the lung i.e. immune responses in the gastrointestinal tract e.g. cytokine response may result in an immune response elsewhere e.g. in the lings (Marsland *et al.*, 2015; Budden *et al.*, 2017; Kim *et al.*, 2017). The microbiome of a healthy lung is primarily affected by immigration and elimination of microorganisms from the lungs, however, in a diseased lung, the growth rate of the microorganism present in the lung primarily affects the microbiome (Dickson *et al.*, 2016).

2.5. Microbial composition of the healthy lung

The respiratory system can be subdivided into two sections: the upper respiratory tract (URT) and lower respiratory tract (LRT) (Man *et al.*, 2017). The URT consists of the anterior nares, nasal passages, paranasal sinuses, nasopharynx, oropharynx and a portion of the larynx (above vocal cords) (Man *et al.*, 2017). The LRT portion of the respiratory system starts at the trachea, branches off into bronchi (and subsequently bronchioles) and ends in millions of alveoli, where gas exchange occurs (Hogan *et al.*, 2014; Man *et al.*, 2017). The entire LRT is lined with epithelium (Hogan *et al.*, 2014). However, in a mature LRT, the type of epithelium differs in structure and composition throughout the LRT (Hogan *et al.*, 2014). There are over 40 different types of epithelial cells in a mature LRT (Li *et al.*, 2015). The LRT can be separated into three regions based on the epithelial structure and is as follows: i) the trachea and bronchi, ii) the bronchioles and iii) the alveoli (Li *et al.*, 2015). The epithelial cells in the trachea and bronchi are pseudostratified columnar epithelial cells and include basal, club (Clara), ciliated and goblet cells (Palange and Simonds, 2013; Li *et al.*, 2015). Dispersed amongst these cells are submucosal glands (Fahy and Dickey, 2010; Li *et al.*, 2015). These glands consist of mucous cells and serous cells (Fahy and Dickey, 2010). The bronchioles are made up of ciliated cells, goblet cells, neuroendocrine (Kulchitsky) cells and secretory club cells (Palange and Simonds, 2013; Li *et al.*, 2015). The alveoli are composed of two types of epithelial cells (pneumocytes) and connective tissue (Palange and Simonds, 2013; Li *et al.*, 2015; Standring, 2015). These alveoli cells are referred to as alveolar type I (AECI/ATI cells) and AECII/ATII cells (Guillot *et al.*, 2013; Palange and Simonds, 2013). These cells produce components of the extracellular matrix (ECM) and growth factors (Palange and Simonds, 2013). The AECI cells are responsible for gas exchange (Palange and Simonds, 2013; Hogan *et al.*, 2014).

In the healthy lung, nutrient supply is low and this may contribute to the low biomass in the lung (Makino *et al.*, 2003; Dickson *et al.*, 2016; Scheiermann and Klinman, 2017; Vecchio-Pagan *et al.*, 2017). It has been suggested that the lung microbiome in healthy individuals is

transient, with the constant movement of bacteria (Charlson *et al.*, 2011; Budden *et al.*, 2017). However, several phyla are predominant in the healthy lung and include *Firmicutes*, *Proteobacteria*, *Actinobacteria* and *Bacteroidetes* (Beck *et al.*, 2012; O'Dwyer *et al.*, 2016; Marimón, 2018). At a genus level, *Prevotella*, *Veillonella* and *Streptococcus* were found to be present in all the studies done using healthy volunteers (as of 2018) (Chung, 2017; Marimón, 2018). In the elderly, *Rothia* and *Lactobacillus* species were more prevalent (Marimón, 2018).

In the healthy lung, viruses can be found as well (Jankauskaite *et al.*, 2018). The healthy lung virome has limited diversity and is comprised of mostly DNA viruses and bacteriophages (Jankauskaite *et al.*, 2018). Additionally, retroviruses can be incorporated into the human genome and have been found in the lung; however, the effect that these viruses may have on diseases in the lung is unknown (Flight *et al.*, 2019). One of the most common virus families found is the *Anelloviridae* (Flight *et al.*, 2019). The *Anelloviridae* are non-enveloped ssDNA viruses (negative sense), that can mutate at a high rate (Spandole *et al.*, 2015). This family of viruses (i.e. *Anelloviridae*) has not been associated with disease in humans (Abbas *et al.*, 2019).

2.6 Changes in the lung microbiome during disease

During lung disease the respiratory ecosystem changes (Dickson *et al.*, 2016). Factors such as cell biology and innate defences may be altered (Huffnagle and Dickson, 2015). Changes in nutrient supply e.g. accumulation of inflammatory by-products due to reactive oxygen species (ROS) and reactive nitrogen species (RNS), results in some phyla increasing and outgrowing other bacteria in the lungs (Winter and Baumler, 2014; Scales *et al.*, 2016). In lung diseases, mucus is often hyper secreted (Williams *et al.*, 2006; Dickson *et al.*, 2015b). Mucus has a gel-like structure and its main component is mucin, a glycoprotein that is primarily consisting of 50% to 90% carbon (Rabiu and Gibson, 2002; Fahy and Dickey, 2010; Alrahman and Yoon, 2017). Bacteria, such as *Pseudomonas aeruginosa* (an opportunistic pathogen) can utilise carbon as a growth medium to outgrow competitors (Rabiu and Gibson, 2002; Alrahman and Yoon, 2017). Table 2.5. shows the different changes to the lung microbiome based on different diseases.

Table 2.5: Overview of the changes to the lung microbiome in different lung diseases and HIV

Disease	Change to the microbiome	References
COPD	Increase in *Proteobacteria* (increases with disease severity and with exacerbations)	(Erb-Downward *et al.*, 2011; Dickey *et al.*, 2015; Adar *et al.*, 2016; Mammen and Sethi, 2016)
HIV	Lower alpha diversity and increased prevalence of *Tropheryma whipplei*	(Twigg *et al.*, 2017).
CF	*Burkholderia*, *Pseudomonas* and *Staphylococcus* are present in high abundances	(Hery-Arnaud *et al.*, 2019)
TB	Higher alpha diversity	(Hong *et al.*, 2016)
Asthma	Increase in *Proteobacteria* and a decrease in *Bacteroidetes*	(O'Dwyer *et al.*, 2016)

CF: Cystic fibrosis
COPD: Chronic obstructive pulmonary disease
HIV: Human immunodeficiency virus
TB: Tuberculosis

2.7 An overview of chronic obstructive pulmonary disease

Chronic obstructive pulmonary disease (COPD) is a complex respiratory disease which is characterised by persistent respiratory symptoms due to exposure to noxious particles or gases (Sarioglu *et al.*, 2016; Vogelmeier *et al.*, 2017). Chronic obstructive pulmonary disease accounts for 5.1% of the global mortality and impaired quality of life (Lee *et al.*, 2016; Oliveira *et al.*, 2018). The disease affects over 300 million people worldwide and the reported prevalence of COPD in sub-Saharan Africa ranges from 4.1% to 24.8% (Salvi, 2015). The highest reported prevalence (23.8%) was from a study conducted in Cape Town, South Africa that formed part of the Burden of Obstructive Lung Disease (BOLD) study conducted in 2006 (Buist *et al.*, 2007).

2.7.1 Pathogenesis and clinical manifestations of chronic obstructive pulmonary disease

At an anatomical level, COPD can affect the small airways, large airways and lung parenchyma that results in bronchiolitis, chronic bronchitis or emphysema, respectively (MacNee, 2006; Macnee *et al.*, 2016; Global Initiative for Obstructive Lung Disease, 2019). These changes occur as a result of chronic inflammation in the lungs (Hogg *et al.*, 2017; Caramori *et al.*, 2018; Global Initiative for Obstructive Lung Disease, 2019). This inflammation is a result of both innate and adaptive immune responses (Brusselle *et al.*, 2011). Figure 2.3 depicts the innate and adaptive responses in the COPD lung.

The inflammatory response in COPD is not a normal response and appears to occur as a result of chronic irritants such as smoke (Global Initiative for Obstructive Lung Disease, 2019). Cigarette smoke generates both a particulate fraction and a gas fraction, each of which contains

over 10^{15} free radicals (Fischer *et al.*, 2015). These free radicals, i.e. ROS, cause a shift in the normal balance of oxidants and antioxidants, causing oxidative stress (Fischer *et al.*, 2015; McGuinness and Sapey, 2017). This oxidative stress causes damage to DNA, proteins and lipids (McGuinness and Sapey, 2017; Ng Kee Kwong *et al.*, 2017). The damage caused to DNA includes shortening of telomere length (accelerates ageing) and histone acetylation/deacetylation (changes gene expression) (Milic *et al.*, 2015; Eapen *et al.*, 2017).

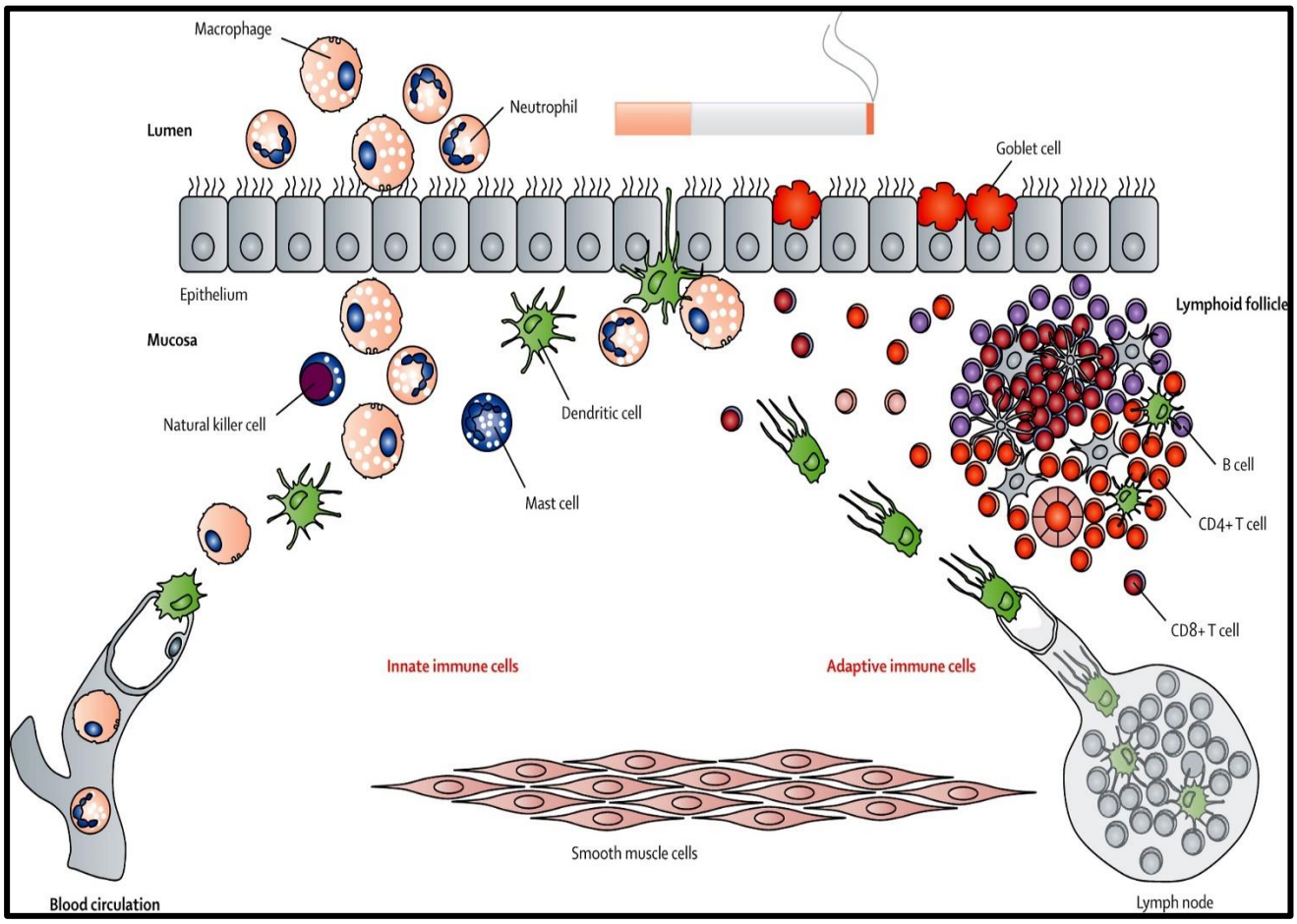

Figure 2.3: Diagram showing the innate and adaptive immune components in chronic obstructive pulmonary disease. Smoke activates innate immune responses by activating the epithelial cells, macrophages and natural killer (NK) cells. Dendritic cells activate the adaptive immune response including B cells and T cells (Brusselle *et al.*, 2011).

In addition to the direct damage to the cells and tissues, the ROS can initiate inflammation by inducing proinflammatory cytokines, chemokines and proteases (Fischer *et al.*, 2015; Footitt *et al.*, 2016; Eapen *et al.*, 2017). The ROS can also suppress the phagocytotic and efferocytotic

(removal of dead cells and debris) abilities of neutrophils and alveolar macrophages; increasing apoptotic cells and bacteria (Fischer *et al.*, 2015; Eapen *et al.*, 2017; Yamasaki and Eeden, 2018). Additionally, the neutrophils and macrophages (which are increased in the COPD lung) contribute to inflammation through the release of chemokines, cytokines, ROS, and proteases, such as neutrophil elastase and matrix metallopeptidases (MMP12, MMP-9 and MMP-1) (Chung and Adcock, 2008; King, 2015; Eapen *et al.*, 2017). These proteases can degrade collagen, lung parenchyma and other cells, causing tissue damage, resulting in emphysema (Dey *et al.*, 2018).

Chronic obstructive pulmonary disease covers several different clinical phenotypes (Papaioannou *et al.*, 2009). Individuals with COPD often present with breathlessness (dyspnoea), chronic coughing and sputum production (Lee *et al.*, 2016; Vogelmeier *et al.*, 2017). Breathlessness/dyspnoea is defined as "a subjective experience of breathing discomfort that consists of qualitatively distinct sensations that vary in intensity" by the American Thoracic Society (ATS) (American Thoracic Society, 1999; Robson, 2017). Breathlessness is not constant and can vary according to activities (Mullerova *et al.*, 2014). A chronic cough in this context is defined as a cough that persists for more than eight weeks (Irwin and Madison, 2000; Martin and Harrison, 2015). Data suggests that females are more susceptible to chronic coughing (as a symptom); which may explain why females with COPD report respiratory symptoms more often (Dicpinigaitis and Rauf, 1998; Kastelik *et al.*, 2002; Martinez *et al.*, 2007; Kavalcikova-Bogdanova *et al.*, 2016; Plevkova *et al.*, 2017).

2.7.2 Clinical diagnosis and assessment of chronic obstructive pulmonary disease

Diagnosing COPD is not an easy endeavour, primarily since there is no clear-cut definition for the disease; however, all definitions agree that the disease is pulmonary and is heterogeneous (Andreeva *et al.*, 2017). In this study, the Global Initiative of Chronic Obstructive Lung Disease (GOLD) guidelines definition is used, which states that COPD is "characterised by persistent respiratory symptoms and airflow limitation that is due to airway and/or alveolar abnormalities usually caused by significant exposure to noxious particles or gases" (Global Initiative for Obstructive Lung Disease, 2019).

If a person over the age of 40 years old has shortness of breath, chronic cough and sputum production combined with a history of smoking or exposure to other risk factors such as pollution, biofuels and occupational hazards e.g. dust from mines, fumes and gases (the

symptoms are discussed in detail in section 2.2.1), a diagnosis of COPD should be considered according to the GOLD guidelines (Vogelmeier *et al.*, 2017). While these symptoms and risk factors are indicative of COPD, other tests need to be performed to confirm the diagnosis and to assess the severity of the disease (Vogelmeier *et al.*, 2017). The South African guidelines are the same as the GOLD guidelines except for additional risk factors, such as HIV and previous *Mycobacterium tuberculosis* infection which should be taken into consideration as well (Abdool-Gaffar *et al.*, 2019).

Spirometry, (a test used to determine the lung function of an individual) is performed as follows: first, the patient takes a deep breath (inhalation) and then the patient exhales (blows out) as fast as they possibly can into the instrument (Bailey, 2012; Koegelenberg *et al.*, 2012; Vogelmeier *et al.*, 2017). This test is repeated two more times, to get an accurate reading and can be repeated up to seven times (Koegelenberg *et al.*, 2012). The instrument measures the volumes of exhaled air and plots it on a volume (y-axis) vs time (x-axis) graph or a flow (y-axis) vs volume (x-axis) graph, which is referred to as a spirogram (Koegelenberg *et al.*, 2012). Several measures are obtained from the instrument and include: i) forced vital capacity (FVC); the maximum volume of air exhaled after maximum inhalation and ii) forced expiratory volume in one second (FEV_1); the volume of air exhaled in the first second (Koegelenberg *et al.*, 2012; Global Initiative for Obstructive Lung Disease, 2019). A ratio of these two values (often expressed as a percentage), FEV_1/FVC is used as well (Koegelenberg *et al.*, 2012; Global Initiative for Obstructive Lung Disease, 2019). A person suffering from COPD typically shows decreased FEV_1 and FVC values (Koegelenberg *et al.*, 2012; Global Initiative for Obstructive Lung Disease, 2019). However, to determine if the airflow limitation is reversible or not, a bronchodilator (dilates the airways) is used (Koegelenberg *et al.*, 2012; Global Initiative for Obstructive Lung Disease, 2019). A post-bronchodilator FEV_1/FVC ratio of <0.70 is typically used for the diagnosis of COPD, however, this cut-off value has been brought into question as it may lead to overdiagnosis in the elderly and underdiagnosis in younger adults (Pellegrino *et al.*, 2005; Culver *et al.*, 2017). The European Respiratory Society (ERS) and the American Thoracic Society (ATS) recommend the use of lower limits of normal (LLN) values instead (Pellegrino *et al.*, 2005; Culver *et al.*, 2017; Global Initiative for Obstructive Lung Disease, 2019). These values (LLN) consider the lower five percent of a healthy population as abnormal (Brazzale *et al.*, 2016; Global Initiative for Obstructive Lung Disease, 2019). An FEV_1/FVC ratio that is below the LLN is suggestive of obstruction (Brazzale *et al.*, 2016; Global Initiative for Obstructive Lung Disease, 2019). Even though the FEV_1/FVC ratio may lead to under-

/overdiagnosis, GOLD still recommends the use of this ratio over LLN; stating that it is the only parameter used in diagnosis (Global Initiative for Obstructive Lung Disease, 2019). Table 2.6 shows the tests used for the diagnosis and assessment of COPD and their advantages and disadvantages.

Table 2.6: Tests for the diagnosis and assessment of chronic obstructive pulmonary disease and their advantages and disadvantages

Test	Description of the test	Advantages	Disadvantages	Reference
		Lung Physiology		
Spirometry	Measurement of the volume of air that is exhaled and inhaled over a period of time	Non-invasive and sensitive	Equipment requires training to use	(Miller *et al.*, 2005; Make and Martinez, 2008; Moore, 2012; Gold and Koth, 2016)
Lung volume test	Detects the volume of air in the lungs after inhalation, exhalation and after a tidal breath (volume of air displaced during normal breathing)	Able to detect airflow limitations as the disease progress an increase in lung capacity occurs	N/A	(Make and Martinez, 2008; Papaioannou *et al.*, 2009; Bailey, 2012; Global Initiative for Obstructive Lung Disease, 2019)
Arterial blood gas	A blood test that measures the pH, oxygen and carbon dioxide levels in the blood. It is the lungs ability to move oxygen and remove carbon dioxide	Useful in determining if oxygen therapy will help the patient	Only recommended for patients with possible respiratory failure	(Make and Martinez, 2008; Papaioannou *et al.*, 2009; McKeever *et al.*, 2016; Global Initiative for Obstructive Lung Disease, 2019)
Pulse oximetry	Measures the oxygen saturation	Non-invasive	Not as accurate as arterial blood gas	(Make and Martinez, 2008; Amalakanti and Pentakota, 2016; Global Initiative for Obstructive Lung Disease, 2019)
Diffusing capacity test (DLCO)	Measurement of carbon monoxide (CO) transfer from alveoli to red blood cells i.e. the diffusion of CO (due to its high affinity for haemoglobin)	Is indicative for emphysema; as the disease progress there are fewer alveoli and as such less diffusion. Non-invasive	Requires trained individuals	(Matheson *et al.*, 2007; Make and Martinez, 2008; Papaioannou *et al.*, 2009; Bailey, 2012; Lumb, 2016b; Global Initiative for Obstructive Lung Disease, 2019)
		Lung structure		
Radiology e.g. X-ray	Provides images of the lung to visualise any changes	Rules out other possible causes of symptoms (differential diagnosis) Can detect emphysema	Not diagnostic for COPD	(Washko, 2010; Global Initiative for Obstructive Lung Disease, 2019)

N/A: Not available

Table 2.6: Tests for the diagnosis and assessment of chronic obstructive pulmonary disease and their advantages and disadvantages

Test	Description of the test	Advantages	Disadvantages	Reference
Genetics				
Alpha-1 antitrypsin deficiency screening	Screens for a genetic marker on the *SERPINA1* gene that encodes for protease inhibitor (protects the lung tissue from destruction)	N/A	The frequency of this gene is much lower in the Asian and African populations than in Caucasian populations and therefore cannot be used for diagnosis in these populations	(de Serres, 2002)
Patient-reported				
Questionnaires e.g. Modified British Medical Research Council (mMRC) and COPD assessment test (CAT)	Patients are asked to answer questions based on their symptoms	Gives information on the severity of symptoms	Cannot be used in the diagnosis of COPD	(Global Initiative for Obstructive Lung Disease, 2019)
Physical capacity				
Exercise/Physical activity testing e.g. Six-minute walk test	Patients are monitored for physical signs of fatigue and breathlessness, the distance walked in six minutes and pulse oximetry during the exercise	Able to predict patients who are at higher risk for morbidity/mortality	N/A	(A. T. S. Committee on Proficiency Standards for Clinical Pulmonary Function Laboratories, 2002; Enright, 2003; Holland *et al.*, 2014; Enright, 2016; Waatevik *et al.*, 2016; Global Initiative for Obstructive Lung Disease, 2019)

N/A: Not available

Some of the above methods (Table 2.6) e.g. X-ray cannot be used on their own to diagnose COPD, however, these tests help rule out other diseases which may have similar symptoms and may help in disease management by providing additional information, such as lung structure (Global Initiative for Obstructive Lung Disease, 2019). While breathlessness and coughing are suggestive of COPD, these symptoms may also be suggestive of other lung diseases, such as asthma (Lumb, 2016a; Anzueto and Miravitlles, 2018; Global Initiative for Obstructive Lung Disease, 2019). Table 2.7 summarises these diseases and how to differentiate them from COPD. Asthma shares many overlapping features with COPD including airway narrowing (airflow limitation) and inflammation (Price *et al.*, 2010; Lange *et al.*, 2016). However, the type of

inflammation differs between the two diseases; asthma is primarily eosinophilic inflammation and COPD is primarily neutrophilic inflammation (however, some COPD patients do present with an eosinophilic phenotype) (Postma and Rabe, 2015; Loureiro, 2016). However, in older patients, these two diseases may overlap, in a condition known as asthma-COPD overlap syndrome (ACOS/ACO) (Postma and Rabe, 2015).

Table 2.7: Differential diagnosis of chronic obstructive pulmonary disease

Disease	How to differentiate the disease from COPD	Reference
Asthma	Post-bronchodilator spirometry (COPD shows limited reversibility whereas asthma often shows reversibility after therapy) Age of onset (asthma usually has an early onset) Asthmatic often have allergies	(Celli *et al.*, 2004; Postma and Rabe, 2015; Global Initiative for Obstructive Lung Disease, 2019)
Bronchiectasis	Chest CT (bronchial dilation and wall thickening is present in bronchiectasis)	(Celli *et al.*, 2004; Price *et al.*, 2010; Global Initiative for Obstructive Lung Disease, 2019)
Obliterative bronchiolitis	Chest CT shows areas of decreased lung density	(Celli *et al.*, 2004; Price *et al.*, 2010; Burgel *et al.*, 2013; Global Initiative for Obstructive Lung Disease, 2019)
Diffuse panbronchiolitis	High-resolution chest CT shows hyperinflated areas Mostly seen in Asians, rare in Caucasians	(Celli *et al.*, 2004; Price *et al.*, 2010; Burgel *et al.*, 2013; Global Initiative for Obstructive Lung Disease, 2019)

COPD: Chronic obstructive pulmonary disease
CT: chest tomography

The management of COPD requires the assessment of four factors: i) degree of airflow limitation (spirometry), ii) symptoms [through tests such as questionnaires, such as the COPD assessment test (CAT) or modified British Medical Research Council test(mMRC), six-minute walk test, etc.], iii) risk of exacerbation (number of exacerbations per year) and iv) comorbidities (Papaioannou *et al.*, 2009; Lange *et al.*, 2016; Global Initiative for Obstructive Lung Disease, 2019). There are several assessment tools, including the "ABCD" tool (featured in the 2011 GOLD update and has replaced the previously used GOLD stages 1-4) and the newly refined ABCD tool (Global Initiative for Obstructive Lung Disease, 2019). The refined ABCD tool (recommended by GOLD) is a combination of both the GOLD stages and the ABCD classification previously used, using both spirometric and symptomatic data (Global Initiative for Obstructive Lung Disease, 2019). The FEV_1 values for the GOLD stages (now known as Grades) are as follows: i) GOLD 1 is ≥80, ii) GOLD 2 is 50 to 79, iii) GOLD 3 is 30 to 49 and iv) GOLD 4 is <30. The ABCD classification has four groups that are divided as follows: i) group A; patients with either no exacerbations or a mild exacerbation (doesn't require hospitalisation) that have a low mMRC and CAT, ii) group B; patients with either no

exacerbations or a mild exacerbation that have high mMRC and CAT scores, iii) group C; patients with either two or more exacerbations or a single exacerbation requiring hospitalisation have low mMRC and CAT scores and iv)group D; patients with either two or more exacerbations or a single exacerbation requiring hospitalisation have high mMRC and CAT scores (Global Initiative for Obstructive Lung Disease, 2019).

The ABCD classification can be summarised as follows; group A consists of patients with barely any symptoms and no exacerbation risk, group B consists of patients that are symptomatic but have no exacerbation risk, group C consists of patients with little to no symptoms but with a high exacerbation risk and group D consists of patients that are symptomatic and have a high exacerbation risk (Lange *et al.*, 2016; Global Initiative for Obstructive Lung Disease, 2019). Assessment of spirometry and symptoms helps with the management of COPD itself, however, comorbidities also need to be taken into account in disease management as they may influence several factors including increased hospitalisation (Global Initiative for Obstructive Lung Disease, 2019).

2.7.3 Management and treatment of chronic obstructive pulmonary disease

The first step in the management of COPD is to stop the patient from smoking i.e. smoking cessation (Global Initiative for Obstructive Lung Disease, 2019). By quitting smoking patients show an initial increase in lung function (even though lung inflammation persists after smoking cessation) and have a better response to other therapies (Jimenez-Ruiz *et al.*, 2015; Global Initiative for Obstructive Lung Disease, 2019). Smoking cessation is done using nicotine replacement products, such as a transdermal patch or by prescribing antidepressants or varenicline (Global Initiative for Obstructive Lung Disease, 2019). The use of vaccines, such as the influenza vaccine and the pneumococcal vaccine, is recommended by GOLD, as these vaccines can reduce the rate of infections (Ruso *et al.*, 2015; Ambrosino and Bertella, 2018; Global Initiative for Obstructive Lung Disease, 2019).

Stable COPD is managed by reducing disease symptoms and exacerbation (Global Initiative for Obstructive Lung Disease, 2019). This management is done through the use of bronchodilators, antimuscarinic drugs, methylxanthines, inhaled corticosteroids, phosphodiesterase-4 (PDE4) inhibitors and antibiotics [such as azithromycin, to reduce the risk of exacerbations in patients who are critically ill, require mechanical ventilation or present with the three cardinal symptoms (increased dyspnoea, increased sputum volume and increased

sputum purulence)] (Global Initiative for Obstructive Lung Disease, 2019) (Table 2.8). In addition to pharmacological (drug) therapy, COPD is managed through pulmonary rehabilitation (increase in physical activity) and oxygen therapy (where needed) (Palange and Simonds, 2013; Global Initiative for Obstructive Lung Disease, 2019). Patients are also encouraged to self-manage i.e. monitor the signs and symptoms of the disease, address risk factors (such as diet), adhere to medications and follow-up with doctors/nurses (Palange and Simonds, 2013; Global Initiative for Obstructive Lung Disease, 2019).

Table 2.8: List of the different drugs used to treat chronic obstructive pulmonary disease and their modes of action and recommended usage (Abdool-Gaffar *et al.*, 2019; Global Initiative for Chronic Obstructive Lung Disease, 2020)

Drug	Mode of action	Examples	Role in COPD therapy	Recommendations
Short-acting beta-antagonist (SABA)	Alters smooth muscle tone; allows the widening of the airways	Salbutamol	Short-term relief	For use in mild COPD (symptomatic management)
Long-acting beta-antagonist (LABA)	Alters smooth muscle tone; allows the widening of the airways	Formoterol	Decreases exacerbations and symptoms improve	For use in moderate COPD
Short-acting anticholinergic (SAMA)	Block the effects of acetylcholine	Ipratropium bromide	Short-term relief	For use in mild COPD (symptomatic management)
Long-acting anticholinergic (LAMA)	Block the effects of acetylcholine	Tiotropium	Decreases exacerbations and hospital visits	For use in moderate COPD
Methylxanthine	Has a bronchodilator effect; however, this drug is highly toxic	Theophylline	Improved quality of life	No recommendations
Corticosteroids	Anti-inflammatory	Fluticasone	Improves lung function and decreases exacerbations	Use in combination therapy (only inhaled therapy recommended)
Phosphodiesterase inhibitors	Inhibit the breakdown of cyclic AMP	Roflumilast	Decreases exacerbations	For use in severe COPD with a history of exacerbations
Mucolytics	Break down mucus	Erdostreine	May decrease exacerbations	Not recommended by the South African Thoracic Society (SATS)

COPD: Chronic obstructive pulmonary disease

An exacerbation in COPD is defined as an "acute worsening in respiratory symptoms that results in additional therapy" (Palange and Simonds, 2013; Global Initiative for Obstructive Lung Disease, 2019). These exacerbations have a high impact on morbidity and mortality as well as quality, however, the majority of these exacerbations go unreported (to healthcare

providers) (Wilkinson *et al.*, 2004; Palange and Simonds, 2013; Global Initiative for Obstructive Lung Disease, 2019). There are three categories of exacerbations: i) mild, ii) moderate and iii) severe (requires hospitalisation) (Global Initiative for Obstructive Lung Disease, 2019). Exacerbations can be triggered by either an infection (bacterial or viral) or by the environment e.g. pollution (Palange and Simonds, 2013; Global Initiative for Obstructive Lung Disease, 2019). Chronic obstructive pulmonary disease exacerbations are treated with two goals in mind: i) reduce the impact of the exacerbation and ii) prevent further exacerbations (Global Initiative for Obstructive Lung Disease, 2019). Pharmacological options are the mainstay for treatment of exacerbations and include the use of short-acting bronchodilators (act quickly), systemic corticosteroids and antibiotics (Palange and Simonds, 2013; Global Initiative for Obstructive Lung Disease, 2019).

2.7.4 Chronic obstructive pulmonary disease and human immunodeficiency virus

The human immunodeficiency virus (HIV) is one of the leading causes of death in South Africa (Statistics South Africa, 2018). As of 2018, 13.1% of the South African population has been infected with HIV i.e. are HIV-positive (Statistics South Africa, 2018). Amongst the provinces in South Africa, KwaZulu-Natal has the highest prevalence and Western Cape has the lowest prevalence with Gauteng having an intermediate prevalence (Shisana *et al.*, 2014; Human Sciences Research Council (HSRC), 2018). Table 2.9 shows the HIV prevalence in South Africa in 2012 and 2017.

Table 2.9: The HIV prevalence in the 15 to 49 age group from 2012 to 2017, per province in South Africa (Shisana *et al.*, 2014; Human Sciences Research Council (HSRC), 2018).

Province	2012 (%)	2017 (%)
Western Cape	7.8	12.6
Northern Cape	11.9	13.9
Limpopo	13.9	17.2
Gauteng	17.8	17.6
North West	20.3	22.7
Mpumalanga	21.8	22.8
Eastern Cape	19.9	25.2
Free State	20.4	25.5
KwaZulu Natal	27.9	27.0

In South Africa, there are about 3.4 million people on antiretroviral therapy (ART) and the ART programme is one of the largest in the world (Moorhouse *et al.*, 2019). With the use of ART, HIV-positive individuals live longer (than in the pre-ART era) (Lalloo *et al.*, 2016). However,

HIV-positive individuals on ARTs show signs of chronic inflammation and accelerated/premature ageing that causes complications (comorbidities) such as liver disease, heart disease, diabetes and pulmonary disease (Deeks *et al.*, 2013; Butler *et al.*, 2018).

Human immunodeficiency virus is considered an independent risk factor for COPD, as individuals with HIV show a higher prevalence of emphysema and a decreased FEV_1 (independent of smoking status) (Drummond *et al.*, 2016; Lalloo *et al.*, 2016; Presti *et al.*, 2017; Bigna *et al.*, 2018). Besides smoking, previous lung infections by bacteria (such as *Pneumocystis jirovecii* and *Tropheryma whipplei*), biomass exposure (e.g. burning of wood or coal), pulmonary tuberculosis, inadequate inflammatory response and oxidative stress have been implicated as the cause of COPD development in HIV-positive individuals (Drummond *et al.*, 2016; Lalloo *et al.*, 2016; Presti *et al.*, 2017; Bigna *et al.*, 2018). Additionally, HIV is associated with increased frequency of exacerbations, especially in individuals with a low CD4 count (Collini and Morris, 2016; Depp *et al.*, 2016; Drummond *et al.*, 2016).

2.8 South African healthcare system

In South Africa, despite the high HIV burden, healthcare is a low economic priority (Mayosi and Benatar, 2014; Malakoane *et al.*, 2020). In most healthcare institutions, long waiting times, lack of proper medicines and inadequate safety precautions are common (Malakoane *et al.*, 2020). The South African healthcare system can be divided into private and public sectors (Malakoane *et al.*, 2020). The public sector is further divided into primary (clinics), secondary (district hospitals) and tertiary (academic hospitals) facilities (Malakoane *et al.*, 2020). Only 16% of the population has access to private healthcare and only 30% of the country's doctors work in the public sector, placing an additional burden on the healthcare system (Mayosi and Benatar, 2014). Most HIV infected individuals only have access to the public sector and as such do not have proper access to healthcare (only have access to an overburden system) (Bogart *et al.*, 2013). Additionally, for proper management of diseases like COPD, access to tertiary institutions is required, to which most of the population does not have access to (Abdool-Gaffar *et al.*, 2019).

2.9 Summary

The human microbiome constitutes all the archaea, bacteria, fungi, protozoans and viruses found in and on the human body along with their genetic material. The most common microorganisms in the human microbiome are bacteria. These bacteria i.e. the bacteriome have

been studied using targeted molecular approaches, such as 16S rDNA sequencing (also known as targeted metagenomics) and the IS-Pro method. Both methods target the 16S rRNA gene which is conserved in all bacteria. The targeted metagenomics sequencing targets hypervariable regions of the 16S rRNA gene that can be used to differentiate between different bacterial genera. The IS-Pro method targets the intergenic spacer region between the 16S rRNA and 23S rRNA genes and is variable in both length and sequence content and can be used to identify bacteria to a species level.

Additionally, the bacteriome has been studied using a metagenomics approach. This approach also allows the study of the viruses i.e. the virome and any other sequences of interest e.g. parasites. The metagenomics approach is useful as all the DNA present in a sample is sequenced and identification of the viruses is conducted using bioinformatics. The data generated from the targeted approaches, such as 16S rRNA sequencing and the IS-Pro method also require additional analysis using bioinformatics approaches. The IS-Pro method uses its propriety software for analysis. However, analysis of 16S rRNA sequencing and metagenomics data can be conducted using online tools such as MG-RAST or python-based tools, such as QIIME2. Steps can be undertaken to ensure good quality sequences and taxonomic assignment of the OTUs i.e. the sequences have occurred. After these steps have been performed and an OTU table generated; the data is compared within the group (alpha diversity) and between groups (beta diversity). Beta diversity measures are used in conjunction with multivariate analysis to compare groups or populations to determine if there is any variation in the microbial composition and which of the factors may be responsible.

Factors that have been shown to affect microbial composition includes pH, temperature and the introduction/removal of bacteria to/from the environment e.g. the human lung. In the human healthy lung, immigration and elimination of microorganisms primarily affect the microbiome whereas as in the diseased lung, the growth rate of microorganism is the primary affecter. The predominating phyla in a healthy lung are *Firmicutes*, *Proteobacteria*, *Actinobacteria* and *Bacteroidetes*. The most common genera include *Prevotella*, *Veillonella* and *Streptococcus*. In disease states, this microbiome is altered e.g. in the COPD lung, *Proteobacteria* increases with disease severity.

Chronic obstructive pulmonary disease is a progressive respiratory disease that is characterised by irreversible airflow limitations. Clinical phenotypes include breathlessness, chronic cough

and sputum production. This disease is diagnosed using clinical features and spirometry. Risk factors for the disease include dust from mines, HIV and previous *Mycobacterium tuberculosis* infection. There is no cure for this disease, however, COPD can be managed through interventions, such as smoking cessation, vaccinations (to prevent exacerbations), drugs e.g. short-acting beta antagonist and oxygen therapy. However, in South Africa, the diagnosis and management of COPD is complicated by the poor healthcare system that has resulted in long waiting times and lack of proper medications at healthcare institutions.

This literature review highlights that while there have been major advances in the field of human microbiome studies, the COPD lung microbiome still requires further investigation. In this study, the sputum microbiome of the COPD lung was investigated. It was expected that even though the targeted metagenomics and the IS-Pro methods will yield different outputs; both these methods would generate an OTU table that can be used to compare the two different technologies. Both these methods targeted the 16S rRNA gene and as such similar microbial composition and diversity were expected. Studies have shown that the stable and exacerbation states of the disease have similar microbial profiles; however, the abundances of these phyla may change during the different disease states; which was what was expected in this study. It was expected that with the virome data, mostly DNA viruses would have been identified and that most of these viruses (including RNA viruses) would be known respiratory viruses. This study was expected to improve the current understanding of the COPD lung microbiome and the IS-Pro method was selected as a possible alternative tool to study the microbiome.

References

A. T. S. Committee on Proficiency Standards for Clinical Pulmonary Function Laboratories (2002). ATS statement: Guidelines for the six-minute walk test. *Am J Respir Crit Care Med,* **166:** 111-117.

Abbas, AA, Young, JC, Clarke, EL, Diamond, JM, Imai, I, Haas, AR, Cantu, E, Lederer, DJ, Meyer, K, Milewski, RK, Olthoff, KM, Shaked, A, Christie, JD, Bushman, FD & Collman, RG (2019). Bidirectional transfer of *Anelloviridae* lineages between graft and host during lung transplantation. *Am J Transplant,* **19:** 1086-1097.

Abdool-Gaffar, MS, Calligaro, G, Wong, ML, Smith, C, Lalloo, UG, Koegelenberg, CFN, Dheda, K, Allwood, BW, Goolam-Mahomed, A & Van Zyl-Smit, RN (2019). Management of chronic obstructive pulmonary disease-a position statement of the South African thoracic society: 2019 update. *J Thorac Dis,* **11:** 4408-4427.

Adar, SD, Huffnagle, GB & Curtis, JL (2016). The respiratory microbiome: An underappreciated player in the human response to inhaled pollutants? *Ann Epidemiol,* **26:** 355-359.

Adetunji, J & Bos, E (2006). Levels and trends in mortality in sub-Saharan Africa: An overview. *In:* Jamison, D. T., Feachem, R. G., Makgoba, M. W., Bos, E. R., Baingana, F. K., Hofman, K. J. & Rogo, K. O. (eds.) Disease and mortality in sub-Saharan Africa. Washington (DC): The International Bank for Reconstruction and Development/The World Bank.

Aguirre, M, Eck, A, Koenen, ME, Savelkoul, PH, Budding, AE & Venema, K (2016). Diet drives quick changes in the metabolic activity and composition of human gut microbiota in a validated *in vitro* gut model. *Res Microbiol,* **167:** 114-125.

Allander, T, Emerson, SU, Engle, RE, Purcell, RH & Bukh, J (2001). A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. *Proc Natl Acad Sci U S A,* **98:** 11609-11614.

Allander, T, Tammi, MT, Eriksson, M, Bjerkner, A, Tiveljung-Lindell, A & Andersson, B (2005). Cloning of a human parvovirus by molecular screening of respiratory tract samples. *Proc Natl Acad Sci U S A,* **102:** 12891-12896.

Allen, B, Kon, M & Bar-Yam, Y (2009). A new phylogenetic diversity measure generalizing the Shannon index and its application to phyllostomid bats. *Am Nat,* **174:** 236-243.

Almeida, A, Mitchell, AL, Tarkowska, A & Finn, RD (2018). Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. *Gigascience,* **7**.

Alrahman, MA & Yoon, SS (2017). Identification of essential genes of *Pseudomonas aeruginosa* for its growth in airway mucus. *J Microbiol,* **55:** 68-74.

Amalakanti, S & Pentakota, MR (2016). Pulse oximetry overestimates oxygen saturation in COPD. *Respir Care,* **61:** 423-427.

Amato, KR (2017). An introduction to microbiome analysis for human biology applications. *Am J Hum Biol,* **29**.

Ambrosino, N & Bertella, E (2018). Lifestyle interventions in prevention and comprehensive management of COPD. *Breathe (Sheff),* **14:** 186-194.

American Thoracic Society (1999). Dyspnea. Mechanisms, assessment, and management: A consensus statement. *Am J Respir Crit Care Med,* **159:** 321-340.

Anderson, MJ (2001). A new method for non-parametric multivariate analysis of variance. *Austral ecology,* **26:** 32-46.

Anderson, MJ & Walsh, DCI (2013). PERMANOVA, ANOSIM, and the mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing? *Ecological Monographs,* **83:** 557-574.

Andreeva, E, Pokhaznikova, M, Lebedev, A, Moiseeva, I, Kuznetsova, O & Degryse, JM (2017). Spirometry is not enough to diagnose COPD in epidemiological studies: A follow-up study. *NPJ Prim Care Respir Med,* **27:** 62.

Anzueto, A & Miravitlles, M (2018). Considerations for the correct diagnosis of COPD and its management with bronchodilators. *Chest,* **154:** 242-248.

Ashton, JJ, Beattie, RM, Ennis, S & Cleary, DW (2016). Analysis and interpretation of the human microbiome. *Inflamm Bowel Dis,* **22:** 1713-1722.

Avila-Rios, S, Parkin, N, Swanstrom, R, Paredes, R, Shafer, R, Ji, H & Kantor, R (2020). Next-generation sequencing for HIV drug resistance testing: Laboratory, clinical, and implementation considerations. *Viruses,* **12**.

Bäckhed, F, Ley, RE, Sonnenburg, JL, Peterson, DA & Gordon, JI (2005). Host-bacterial mutualism in the human intestine. *Science,* **307:** 1915-1920.

Bai, L, Liang, J, Dang, C & Cao, F (2012). A cluster centers initialization method for clustering categorical data. *Expert Syst Appl,* **39:** 8022-8029.

Bailey, KL (2012). The importance of the assessment of pulmonary function in COPD. *Med Clin North Am*, **96:** 745-752.

Barton, LL & Northup, DE (2011). Microbial ecology, Hoboken, New Jersey, John Wiley & Sons.

Beck, JM, Young, VB & Huffnagle, GB (2012). The microbiome of the lung. *Transl Res*, **160:** 258-266.

Bigna, JJ, Kenne, AM, Asangbeh, SL & Sibetcheu, AT (2018). Prevalence of chronic obstructive pulmonary disease in the global population with HIV: A systematic review and meta-analysis. *Lancet Glob Health*, **6:** e193-e202.

Bik, EM (2016). The hoops, hopes, and hypes of human microbiome research. *Yale J Biol Med*, **89:** 363-373.

Bikel, S, Valdez-Lara, A, Cornejo-Granados, F, Rico, K, Canizales-Quinteros, S, Soberon, X, Del Pozo-Yauner, L & Ochoa-Leyva, A (2015). Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: Towards a systems-level understanding of human microbiome. *Comput Struct Biotechnol J*, **13:** 390-401.

Biteen, JS, Blainey, PC, Cardon, ZG, Chun, M, Church, GM, Dorrestein, PC, Fraser, SE, Gilbert, JA, Jansson, JK, Knight, R, Miller, JF, Ozcan, A, Prather, KA, Quake, SR, Ruby, EG, Silver, PA, Taha, S, Van Den Engh, G, Weiss, PS, Wong, GC, Wright, AT & Young, TD (2016). Tools for the microbiome: Nano and beyond. *ACS Nano*, **10:** 6-37.

Boase, S, Foreman, A, Cleland, E, Tan, L, Melton-Kreft, R, Pant, H, Hu, FZ, Ehrlich, GD & Wormald, PJ (2013). The microbiome of chronic rhinosinusitis: Culture, molecular diagnostics and biofilm detection. *BMC Infect Dis*, **13:** 210.

Boers, SA, Jansen, R & Hays, JP (2019). Understanding and overcoming the pitfalls and biases of next-generation sequencing (NGS) methods for use in the routine clinical microbiological diagnostic laboratory. *Eur J Clin Microbiol Infect Dis*, **38:** 1059-1070.

Bogart, LM, Chetty, S, Giddy, J, Sypek, A, Sticklor, L, Walensky, RP, Losina, E, Katz, JN & Bassett, IV (2013). Barriers to care among people living with HIV in South Africa: Contrasts between patient and healthcare provider perspectives. *AIDS Care*, **25:** 843-853.

Bokulich, NA, Subramanian, S, Faith, JJ, Gevers, D, Gordon, JI, Knight, R, Mills, DA & Caporaso, JG (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat Methods*, **10:** 57-59.

Bragg, L & Tyson, GW (2014). Metagenomics using next-generation sequencing. *Methods Mol Biol*, **1096:** 183-201.

Brazzale, D, Hall, G & Swanney, MP (2016). Reference values for spirometry and their use in test interpretation: A position statement from the Australian and New Zealand society of respiratory science. *Respirology*, **21:** 1201-1209.

Brusselle, GG, Joos, GF & Bracke, KR (2011). New insights into the immunology of chronic obstructive pulmonary disease. *The Lancet*, **378:** 1015-1026.

Budden, KF, Gellatly, SL, Wood, DL, Cooper, MA, Morrison, M, Hugenholtz, P & Hansbro, PM (2017). Emerging pathogenic links between microbiota and the gut-lung axis. *Nat Rev Microbiol*, **15:** 55-63.

Budding, AE, Grasman, ME, Eck, A, Bogaards, JA, Vandenbroucke-Grauls, CM, Van Bodegraven, AA & Savelkoul, PH (2014). Rectal swabs for analysis of the intestinal microbiota. *PLoS One*, **9:** e101344.

Budding, AE, Grasman, ME, Lin, F, Bogaards, JA, Soeltan-Kaersenhout, DJ, Vandenbroucke-Grauls, CM, Van Bodegraven, AA & Savelkoul, PH (2010). IS-Pro: High-throughput molecular fingerprinting of the intestinal microbiota. *FASEB J*, **24:** 4556-4564.

Budding, AE, Hoogewerf, M, Vandenbroucke-Grauls, CM & Savelkoul, PH (2016). Automated broad-range molecular detection of bacteria in clinical samples. *J Clin Microbiol*, **54:** 934-943.

Buist, AS, Mcburnie, MA, Vollmer, WM, Gillespie, S, Burney, P, Mannino, DM, Menezes, AMB, Sullivan, SD, Lee, TA, Weiss, KB, Jensen, RL, Marks, GB, Gulsvik, A & Nizankowska-Mogilnicka, E (2007). International variation in the prevalence of COPD (the BOLD study): A population-based prevalence study. *The Lancet*, **370:** 741-750.

Burgel, PR, Bergeron, A, De Blic, J, Bonniaud, P, Bourdin, A, Chanez, P, Chinet, T, Dalphin, JC, Devillier, P, Deschildre, A, Didier, A, Kambouchner, M, Knoop, C, Laurent, F, Nunes, H,

Perez, T, Roche, N, Tillie-Leblond, I & Dusser, D (2013). Small airways diseases, excluding asthma and COPD: An overview. *Eur Respir Rev*, **22:** 131-147.

Bustamante-Marin, XM & Ostrowski, LE (2017). Cilia and mucociliary clearance. *Cold Spring Harb Perspect Biol*, **9**.

Butler, I, Macleod, W, Majuba, PP & Tipping, B (2018). Human immunodeficiency virus infection and older adults: A retrospective single-site cohort study from Johannesburg, South Africa. *South Afr J HIV Med*, **19:** 838.

Buttigieg, PL & Ramette, A (2014). A guide to statistical analysis in microbial ecology: A community-focused, living review of multivariate data analyses. *FEMS Microbiol Ecol*, **90:** 543-550.

Cadwell, K (2015). The virome in host health and disease. *Immunity*, **42:** 805-813.

Caporaso, JG, Kuczynski, J, Stombaugh, J, Bittinger, K, Bushman, FD, Costello, EK, Fierer, N, Pena, AG, Goodrich, JK, Gordon, JI, Huttley, GA, Kelley, ST, Knights, D, Koenig, JE, Ley, RE, Lozupone, CA, Mcdonald, D, Muegge, BD, Pirrung, M, Reeder, J, Sevinsky, JR, Turnbaugh, PJ, Walters, WA, Widmann, J, Yatsunenko, T, Zaneveld, J & Knight, R (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*, **7:** 335-336.

Caramori, G, Ruggeri, P, Di Stefano, A, Mumby, S, Girbino, G, Adcock, IM & Kirkham, P (2018). Autoimmunity and COPD: Clinical implications. *Chest*, **153:** 1424-1431.

Celli, BR, Macnee, W, Agusti, A, Anzueto, A, Berg, B, Buist, AS, Calverley, PMA, Chavannes, N, Dillard, T, Fahy, B, Fein, A, Heffner, J, Lareau, S, Meek, P, Martinez, F, Mcnicholas, W, Muris, J, Austegard, E, Pauwels, R, Rennard, S, Rossi, A, Siafakas, N, Tiep, B, Vestbo, J, Wouters, E & Zuwallack, R (2004). Standards for the diagnosis and treatment of patients with COPD: A summary of the ATS/ERS position paper. *Eur Respir J*, **23:** 932-946.

Chang, Q, Luan, Y & Sun, F (2011). Variance adjusted weighted UniFrac: A powerful beta diversity measure for comparing communities based on phylogeny. *BMC Bioinformatics*, **12:** 118.

Chao, A (1984). Nonparametric estimation of the number of classes in a population. *Scand J Stat,* **11:** 265-270.

Chao, A, Chazdon, RL, Colwell, RK & Shen, TJ (2006). Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics,* **62:** 361-371.

Chao, A, Chiu, C-H & Jost, L (2016). Phylogenetic diversity measures and their decomposition: A framework based on hill numbers. *In:* Pellens, R. & Grandcolas, P. (eds.) Biodiversity conservation and phylogenetic systematics: Preserving our evolutionary heritage in an extinction crisis. Cham: Springer International Publishing.

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

Chazdon, RL, Colwell, RK, Denslow, JS & Guariguata, MR (1998). Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of northeastern Costa Rica. *In:* Dallmeier, F. & Comiskey, J. A. (eds.) Forest biodiversity research, monitoring and modeling: Conceptual background and old world case studies. Paris: Parthenon Publishing.

Chung, KF (2017). Airway microbial dysbiosis in asthmatic patients: A target for prevention and treatment? *J Allergy Clin Immunol,* **139:** 1071-1081.

Chung, KF & Adcock, IM (2008). Multifaceted mechanisms in COPD: Inflammation, immunity, and tissue repair and destruction. *Eur Respir J,* **31:** 1334-1356.

Clarridge, JE, 3rd (2004). Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin Microbiol Rev,* **17:** 840-862, table of contents.

Cock, PJ, Fields, CJ, Goto, N, Heuer, ML & Rice, PM (2010). The Sanger fastq file format for sequences with quality scores, and the Solexa/Illumina fastq variants. *Nucleic Acids Res,* **38:** 1767-1771.

Cole, JR, Wang, Q, Cardenas, E, Fish, J, Chai, B, Farris, RJ, Kulam-Syed-Mohideen, AS, Mcgarrell, DM, Marsh, T, Garrity, GM & Tiedje, JM (2009). The ribosomal database project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Res*, **37:** D141-145.

Collini, P & Morris, A (2016). Maintaining lung health with longstanding HIV. *Curr Opin Infect Dis*, **29:** 31-38.

Culver, BH, Graham, BL, Coates, AL, Wanger, J, Berry, CE, Clarke, PK, Hallstrand, TS, Hankinson, JL, Kaminsky, DA, Macintyre, NR, Mccormack, MC, Rosenfeld, M, Stanojevic, S, Weiner, DJ & ATS Committee on Proficiency Standards for Pulmonary Function Laboratories (2017). Recommendations for a standardized pulmonary function report. An official American thoracic society technical statement. *Am J Respir Crit Care Med*, **196:** 1463-1472.

Daly, A, Baetens, J & De Baets, B (2018). Ecological diversity: Measuring the unmeasurable. *Mathematics*, **6**.

Daniels, L, Budding, AE, De Korte, N, Eck, A, Bogaards, JA, Stockmann, HB, Consten, EC, Savelkoul, PH & Boermeester, MA (2014). Fecal microbiome analysis as a diagnostic test for diverticulitis. *Eur J Clin Microbiol Infect Dis*, **33:** 1927-1936.

Datta, S, Budhauliya, R, Das, B, Chatterjee, S, Vanlalhmuaka & Veer, V (2015). Next-generation sequencing in clinical virology: Discovery of new viruses. *World J Virol*, **4:** 265-276.

de Meij, TG, Budding, AE, De Groot, EF, Jansen, FM, Frank Kneepkens, CM, Benninga, MA, Penders, J, Van Bodegraven, AA & Savelkoul, PH (2016a). Composition and stability of intestinal microbiota of healthy children within a Dutch population. *FASEB J*, **30:** 1512-1522.

de Meij, TG, De Groot, EF, Eck, A, Budding, AE, Kneepkens, CM, Benninga, MA, Van Bodegraven, AA & Savelkoul, PH (2016b). Characterization of microbiota in children with chronic functional constipation. *PLoS One*, **11:** e0164731.

de Serres, FJ (2002). Worldwide racial and ethnic distribution of alpha1-antitrypsin deficiency. *Chest*, **122:** 1818-1829.

Deeks, SG, Lewin, SR & Havlir, DV (2013). The end of AIDS: HIV infection as a chronic disease. *The Lancet*, **382:** 1525-1533.

Depp, TB, Mcginnis, KA, Kraemer, K, Akgun, KM, Edelman, EJ, Fiellin, DA, Butt, AA, Crystal, S, Gordon, AJ, Freiberg, M, Gibert, CL, Rimland, D, Bryant, KJ & Crothers, K (2016). Risk factors associated with acute exacerbation of chronic obstructive pulmonary disease in HIV-infected and uninfected patients. *AIDS*, **30:** 455-463.

DeSantis, TZ, Hugenholtz, P, Larsen, N, Rojas, M, Brodie, EL, Keller, K, Huber, T, Dalevi, D, Hu, P & Andersen, GL (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol*, **72:** 5069-5072.

Dey, T, Kalita, J, Weldon, S & Taggart, CC (2018). Proteases and their inhibitors in chronic obstructive pulmonary disease. *J Clin Med*, **7:** 244-264.

Dhariwal, A, Chong, J, Habib, S, King, IL, Agellon, LB & Xia, J (2017). MicrobiomeAnalyst: A web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Res*, **45:** W180-W188.

Di Bella, JM, Bao, Y, Gloor, GB, Burton, JP & Reid, G (2013). High throughput sequencing methods and analysis for microbiome research. *J Microbiol Methods*, **95:** 401-414.

Dickey, BF, Knowles, MR & Boucher, RC (2015). Mucociliary clearance. *In:* Grippi, M. A., Elias, J. A., Fishman, J. A., Kotloff, R. M., Pack, A. I., Senior, R. M. & Siegel, M. D. (eds.) Fishman's pulmonary diseases and disorders, 5e. New York, NY: McGraw-Hill Education.

Dickson, RP, Erb-Downward, JR, Freeman, CM, Mccloskey, L, Beck, JM, Huffnagle, GB & Curtis, JL (2015a). Spatial variation in the healthy human lung microbiome and the adapted island model of lung biogeography. *Ann Am Thorac Soc*, **12:** 821-830.

Dickson, RP, Erb-Downward, JR & Huffnagle, GB (2015b). Homeostasis and its disruption in the lung microbiome. *Am J Physiol Lung Cell Mol Physiol*, **309:** L1047-1055.

Dickson, RP, Erb-Downward, JR, Martinez, FJ & Huffnagle, GB (2016). The microbiome and the respiratory tract. *Annu Rev Physiol*, **78:** 481-504.

Dicpinigaitis, PV & Rauf, K (1998). The influence of gender on cough reflex sensitivity. *Chest*, **113:** 1319-1321.

Dobell, C (1920). The discovery of the intestinal protozoa of man. *Proc R Soc Med,* **13:** 1-15.

Drummond, MB, Kunisaki, KM & Huang, L (2016). Obstructive lung diseases in HIV: A clinical review and identification of key future research needs. *Semin Respir Crit Care Med,* **37:** 277-288.

Eapen, MS, Myers, S, Walters, EH & Sohal, SS (2017). Airway inflammation in chronic obstructive pulmonary disease (COPD): A true paradox. *Expert Rev Respir Med,* **11:** 827-839.

Eck, A, De Groot, EFJ, De Meij, TGJ, Welling, M, Savelkoul, PHM & Budding, AE (2017). Robust microbiota-based diagnostics for inflammatory bowel disease. *J Clin Microbiol,* **55:** 1720-1732.

Enright, PL (2003). The six-minute walk test. *Respir Care,* **48:** 783-785.

Enright, PL (2016). Oxygen desaturation during a 6-min walk identifies a COPD phenotype with an increased risk of morbidity and mortality. *Eur Respir J,* **48:** 1-2.

Erb-Downward, JR, Thompson, DL, Han, MK, Freeman, CM, Mccloskey, L, Schmidt, LA, Young, VB, Toews, GB, Curtis, JL, Sundaram, B, Martinez, FJ & Huffnagle, GB (2011). Analysis of the lung microbiome in the "healthy" smoker and in COPD. *PLoS One,* **6:** e16384.

Evans, SN & Matsen, FA (2012). The phylogenetic kantorovich-rubinstein metric for environmental sequence samples. *J R Stat Soc Series B Stat Methodol,* **74:** 569-592.

Ewing, B & Green, P (1998). Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res,* **8:** 186-194.

Ewing, B, Hillier, L, Wendl, MC & Green, P (1998). Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res,* **8:** 175-185.

Fahy, JV & Dickey, BF (2010). Airway mucus function and dysfunction. *N Engl J Med,* **363:** 2233-2247.

Faith, DP (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation,* **61:** 1-10.

Faner, R, Sibila, O, Agusti, A, Bernasconi, E, Chalmers, JD, Huffnagle, GB, Manichanh, C, Molyneaux, PL, Paredes, R, Perez Brocal, V, Ponomarenko, J, Sethi, S, Dorca, J & Monso, E

(2017). The microbiome in respiratory medicine: Current challenges and future perspectives. *Eur Respir J,* **49**.

Finney, LJ, Feary, JR, Leonardi-Bee, J, Gordon, SB & Mortimer, K (2013). Chronic obstructive pulmonary disease in sub-Saharan Africa: A systematic review. *Int J Tuberc Lung Dis,* **17:** 583-589.

Finotello, F, Mastrorilli, E & Di Camillo, B (2018). Measuring the diversity of the human microbiota with targeted next-generation sequencing. *Brief Bioinform,* **19:** 679-692.

Fischer, BM, Voynow, JA & Ghio, AJ (2015). COPD: Balancing oxidants and antioxidants. *Int J Chron Obstruct Pulmon Dis,* **10:** 261-276.

Flight, WG, Turkington, CJR & Clokie, MRJ (2019). Viruses and the lung microbiome. *In:* Cox, M., Ege, M. & Von Mutius, E. (eds.) The lung microbiome. Sheffield: European Respiratory Society.

Footitt, J, Mallia, P, Durham, AL, Ho, WE, Trujillo-Torralbo, MB, Telcian, AG, Del Rosario, A, Chang, C, Peh, HY, Kebadze, T, Aniscenko, J, Stanciu, L, Essilfie-Quaye, S, Ito, K, Barnes, PJ, Elkin, SL, Kon, OM, Wong, WS, Adcock, IM & Johnston, SL (2016). Oxidative and nitrosative stress and histone deacetylase-2 activity in exacerbations of COPD. *Chest,* **149:** 62-73.

Foster, JA, Bunge, J, Gilbert, JA & Moore, JH (2012). Measuring the microbiome: Perspectives on advances in DNA-based techniques for exploring microbial life. *Brief Bioinform,* **13:** 420-429.

Frades, I & Matthiesen, R (2010). Overview on techniques in cluster analysis. Bioinformatics methods in clinical research. Humana Press.

Franks, AH, Harmsen, HJ, Raangs, GC, Jansen, GJ, Schut, F & Welling, GW (1998). Variations of bacterial populations in human feces measured by fluorescent *in situ* hybridization with group-specific 16S rRNA-targeted oligonucleotide probes. *Appl Environ Microbiol,* **64:** 3336-3345.

Franzen, O, Hu, J, Bao, X, Itzkowitz, SH, Peter, I & Bashir, A (2015). Improved OTU-picking using long-read 16S rRNA gene amplicon sequencing and generic hierarchical clustering. *Microbiome,* **3:** 43.

Froussard, P (1992). A random-PCR method (rPCR) to construct whole cDNA library from low amounts of RNA. *Nucleic Acids Res,* **20:** 2900-2900.

Gill, SR, Pop, M, Deboy, RT, Eckburg, PB, Turnbaugh, PJ, Samuel, BS, Gordon, JI, Relman, DA, Fraser-Liggett, CM & Nelson, KE (2006). Metagenomic analysis of the human distal gut microbiome. *Science,* **312:** 1355-1359.

Global Initiative for Chronic Obstructive Lung Disease (2020). Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease (2020 report).

Global Initiative for Obstructive Lung Disease. (2019). *2019 global strategy for the prevention, diagnosis and management of COPD* [Online]. [Accessed 07 January 2019].

Gold, WM & Koth, LL (2016). Pulmonary function testing. *In:* V. Courtney Broaddus, Robert J. Mason, Joel D. Ernst, Talmadge E. King, Stephen C. Lazarus, John F. Murray, Jay A. Nadel, Arthur S. Slutsky & Michael B. Gotway (eds.) Murray and Nadel's textbook of respiratory medicine. Sixth ed. Philadelphia: W.B. Saunders.

Goodrich, JK, Di Rienzi, SC, Poole, AC, Koren, O, Walters, WA, Caporaso, JG, Knight, R & Ley, RE (2014). Conducting a microbiome study. *Cell,* **158:** 250-262.

Goodwin, S, Mcpherson, JD & Mccombie, WR (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nat Rev Genet,* **17:** 333-351.

Grasman, M, Van Der Borden, R, Budding, A, Eck Hauer, A, Savelkoul, P & Van Bodegraven, A (2014). P668 interspace microbiome profiling (IS-Pro) enables to differentiate IBD subclasses and disease activity by specific loss of bacterial diversity. *J Crohns Colitis,* **8**.

Guillot, L, Nathan, N, Tabary, O, Thouvenin, G, Le Rouzic, P, Corvol, H, Amselem, S & Clement, A (2013). Alveolar epithelial cells: Master regulators of lung homeostasis. *Int J Biochem Cell Biol,* **45:** 2568-2573.

Gürtler, V, Subrahmanyam, G, Shekar, M, Maiti, B & Karunasagar, I (2014). Chapter 12-bacterial typing and identification by genomic analysis of 16S–23S rRNA intergenic

transcribed spacer (ITS) sequences. *In:* Michael Goodfellow, Iain Sutcliffe & Chun, J. (eds.) Methods in microbiology. Academic Press.

GUSTA ME [Online]. (2014). Available: https://mb3is.megx.net/gustame/home [Accessed 11 April 2018].

Haghi, M, Ong, HX, Traini, D & Young, P (2014). Across the pulmonary epithelial barrier: Integration of physicochemical properties and human cell models to study pulmonary drug formulations. *Pharmacol Ther,* **144:** 235-252.

Hamady, M & Knight, R (2009). Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res,* **19:** 1141-1152.

Handelsman, J, Rondon, MR, Brady, SF, Clardy, J & Goodman, RM (1998). Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chem Biol,* **5:** R245-249.

Hartman, AL, Riddle, S, Mcphillips, T, Ludascher, B & Eisen, JA (2010). Introducing W.A.T.E.R.S.: A workflow for the alignment, taxonomy, and ecology of ribosomal sequences. *BMC Bioinformatics,* **11:** 317.

Heltshe, JF & Forrester, NE (1983). Estimating species richness using the Jackknife procedure. *Biometrics,* **39:** 1-11.

Hermann-Bank, ML, Skovgaard, K, Stockmarr, A, Larsen, N & Molbak, L (2013). The gut microbiotassay: A high-throughput qPCR approach combinable with next generation sequencing to study gut microbial diversity. *BMC Genomics,* **14:** 788.

Hery-Arnaud, G, Boutin, S, Cuthbertson, L, Elborn, SJ & Tunney, MM (2019). The lung and gut microbiome: What has to be taken into consideration for cystic fibrosis? *J Cyst Fibros,* **18:** 13-21.

Hiergeist, A, Glasner, J, Reischl, U & Gessner, A (2015). Analyses of intestinal microbiota: Culture versus sequencing. *ILAR J,* **56:** 228-240.

Hiergeist, A, Reischl, U, Priority Program Intestinal Microbiota Consortium/ Quality Assessment, p & Gessner, A (2016). Multicenter quality assessment of 16S ribosomal DNA-

sequencing for microbiome analyses reveals high inter-center variability. *Int J Med Microbiol*, **306:** 334-342.

Hill, C, Ross, RP, Stanton, C & O'toole, PW (2016). The human microbiome in health and disease. *In:* Unden, G., Thines, E. & Schüffler, A. (eds.) Host - pathogen interaction. John Wiley & Sons.

Hodinka, RL (2013). Point: Is the era of viral culture over in the clinical microbiology laboratory? *J Clin Microbiol*, **51:** 2-4.

Hodkinson, BP & Grice, EA (2015). Next-generation sequencing: A review of technologies and tools for wound microbiome research. *Adv Wound Care (New Rochelle)*, **4:** 50-58.

Hogan, BL, Barkauskas, CE, Chapman, HA, Epstein, JA, Jain, R, Hsia, CC, Niklason, L, Calle, E, Le, A, Randell, SH, Rock, J, Snitow, M, Krummel, M, Stripp, BR, Vu, T, White, ES, Whitsett, JA & Morrisey, EE (2014). Repair and regeneration of the respiratory system: Complexity, plasticity, and mechanisms of lung stem cell function. *Cell Stem Cell*, **15:** 123-138.

Hogg, JC, Pare, PD & Hackett, TL (2017). The contribution of small airway obstruction to the pathogenesis of chronic obstructive pulmonary disease. *Physiol Rev*, **97:** 529-552.

Holland, AE, Spruit, MA, Troosters, T, Puhan, MA, Pepin, V, Saey, D, Mccormack, MC, Carlin, BW, Sciurba, FC, Pitta, F, Wanger, J, Macintyre, N, Kaminsky, DA, Culver, BH, Revill, SM, Hernandes, NA, Andrianopoulos, V, Camillo, CA, Mitchell, KE, Lee, AL, Hill, CJ & Singh, SJ (2014). An official European respiratory society/American thoracic society technical standard: Field walking tests in chronic respiratory disease. *Eur Respir J*, **44:** 1428-1446.

Hong, BY, Maulen, NP, Adami, AJ, Granados, H, Balcells, ME & Cervantes, J (2016). Microbiome changes during tuberculosis and antituberculous therapy. *Clin Microbiol Rev*, **29:** 915-926.

Hooper, LV & Gordon, JL (2001). Commensal host-bacterial relationships in the gut. *Science*, **292:** 1115-1118.

Horn, HS (1966). Measurement of "overlap" in comparative ecological studies. *Am Nat*, **100:** 419-424.

Huang, YJ, Erb-Downward, JR, Dickson, RP, Curtis, JL, Huffnagle, GB & Han, MK (2017). Understanding the role of the microbiome in chronic obstructive pulmonary disease: Principles, challenges, and future directions. *Transl Res,* **179:** 71-83.

Huffnagle, GB & Dickson, RP (2015). The bacterial microbiota in inflammatory lung diseases. *Clin Immunol,* **159:** 177-182.

Hughes, JB, Hellmann, JJ, Ricketts, TH & Bohannan, BJ (2001). Counting the uncountable: Statistical approaches to estimating microbial diversity. *Appl Environ Microbiol,* **67:** 4399-4406.

Human Microbiome Project, C (2012). A framework for human microbiome research. *Nature,* **486:** 215-221.

Human Sciences Research Council (HSRC). (2018). *The fifth south South African national HIV prevalence, incidence and behaviour survey, 2017 (sabssm v1)* [Online]. Available: www.hsrc.ac.za/uploads/pageContent/9234/SABSSMV_Impact_Assessment_Summary_ZA_ADS_cleared_PDFA4.pdf [Accessed January 2019].

Huse, SM, Mark Welch, DB, Voorhis, A, Shipunova, A, Morrison, HG, Eren, AM & Sogin, ML (2014). VAMPS: A website for visualization and analysis of microbial population structures. *BMC Bioinformatics,* **15:** 41.

Irwin, RS & Madison, JM (2000). The diagnosis and treatment of cough. *N Engl J Med,* **343:** 1715-1721.

Jaccard, P (1912). The distribution of the flora in the alpine zone. *New Phytol,* **11:** 37-50.

Jain, AK (2010). Data clustering: 50 years beyond *k*-means. *Pattern Recognit Lett,* **31:** 651-666.

Jankauskaite, L, Miseviciene, V, Vaideliene, L & Kevalas, R (2018). Lower airway virology in health and disease-from invaders to symbionts. *Medicina (Kaunas),* **54:** 72-87.

Janssens, PL, Penders, J, Hursel, R, Budding, AE, Savelkoul, PH & Westerterp-Plantenga, MS (2016). Long-term green tea supplementation does not change the human gut microbiota. *PLoS One,* **11:** e0153134.

Jervis-Bardy, J, Leong, LE, Marri, S, Smith, RJ, Choo, JM, Smith-Vaughan, HC, Nosworthy, E, Morris, PS, O'leary, S, Rogers, GB & Marsh, RL (2015). Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of Illumina MiSeq data. *Microbiome,* **3:** 19.

Jimenez-Ruiz, CA, Andreas, S, Lewis, KE, Tonnesen, P, Van Schayck, CP, Hajek, P, Tonstad, S, Dautzenberg, B, Fletcher, M, Masefield, S, Powell, P, Hering, T, Nardini, S, Tonia, T & Gratziou, C (2015). Statement on smoking cessation in COPD and other pulmonary diseases and in smokers with comorbidities who find it difficult to quit. *Eur Respir J,* **46:** 61-79.

Jovel, J, Patterson, J, Wang, W, Hotte, N, O'keefe, S, Mitchel, T, Perry, T, Kao, D, Mason, AL, Madsen, KL & Wong, GK (2016). Characterization of the gut microbiome using 16S or shotgun metagenomics. *Front Microbiol,* **7:** 459.

Ju, F & Zhang, T (2015). 16S rRNA gene high-throughput sequencing data mining of microbial diversity and interactions. *Appl Microbiol Biotechnol,* **99:** 4119-4129.

Kallies, R, Holzer, M, Brizola Toscan, R, Nunes Da Rocha, U, Anders, J, Marz, M & Chatzinotas, A (2019). Evaluation of sequencing library preparation protocols for viral metagenomic analysis from pristine aquifer groundwaters. *Viruses,* **11**.

Kastelik, JA, Thompson, RH, Aziz, I, Ojoo, JC, Redington, AE & Morice, AH (2002). Sex-related differences in cough reflex sensitivity in patients with chronic cough. *Am J Respir Crit Care Med,* **166:** 961-964.

Kavalcikova-Bogdanova, N, Buday, T, Plevkova, J & Song, WJ (2016). Chronic cough as a female gender issue. *Adv Exp Med Biol,* **905:** 69-78.

Kelly, BJ, Gross, R, Bittinger, K, Sherrill-Mix, S, Lewis, JD, Collman, RG, Bushman, FD & Li, H (2015). Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA. *Bioinformatics,* **31:** 2461-2468.

Kembel, SW, Cowan, PD, Helmus, MR, Cornwell, WK, Morlon, H, Ackerly, DD, Blomberg, SP & Webb, CO (2010). Picante: R tools for integrating phylogenies and ecology. *Bioinformatics,* **26:** 1463-1464.

Kembel, SW, Wu, M, Eisen, JA & Green, JL (2012). Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Comput Biol*, **8:** e1002743.

Khan, SS & Ahmad, A (2004). Cluster center initialization algorithm for *k*-means clustering. *Pattern Recognit Lett*, **25:** 1293-1302.

Kim, D, Hofstaedter, CE, Zhao, C, Mattei, L, Tanes, C, Clarke, E, Lauder, A, Sherrill-Mix, S, Chehoud, C, Kelsen, J, Conrad, M, Collman, RG, Baldassano, R, Bushman, FD & Bittinger, K (2017). Optimizing methods and dodging pitfalls in microbiome research. *Microbiome*, **5:** 52.

Kim, M, Oh, HS, Park, SC & Chun, J (2014). Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int J Syst Evol Microbiol*, **64:** 346-351.

King, PT (2015). Inflammation in chronic obstructive pulmonary disease and its role in cardiovascular disease and lung cancer. *Clin Transl Med*, **4:** 68.

Kleiner, M, Hooper, LV & Duerkop, BA (2015). Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. *BMC Genomics*, **16:** 7.

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

Knight, R, Vrbanac, A, Taylor, BC, Aksenov, A, Callewaert, C, Debelius, J, Gonzalez, A, Kosciolek, T, Mccall, LI, Mcdonald, D, Melnik, AV, Morton, JT, Navas, J, Quinn, RA, Sanders, JG, Swafford, AD, Thompson, LR, Tripathi, A, Xu, ZZ, Zaneveld, JR, Zhu, Q, Caporaso, JG & Dorrestein, PC (2018). Best practices for analysing microbiomes. *Nat Rev Microbiol*, **16:** 410-422.

Koedooder, R, Singer, M, Schoenmakers, S, Savelkoul, PHM, Morre, SA, De Jonge, JD, Poort, L, Cuypers, WSS, Budding, AE, Laven, JSE & ReceptIVFity Study Group (2018). The ReceptIVFity cohort study protocol to validate the urogenital microbiome as predictor for IVF or IVF/ICSI outcome. *Reprod Health*, **15:** 202.

Koegelenberg, CF, Swart, F & Irusen, EM (2012). Guideline for office spirometry in adults, 2012. *S Afr Med J,* **103:** 52-62.

Koleff, P, Gaston, KJ & Lennon, JJ (2003). Measuring beta diversity for presence–absence data. *J Anim Ecol,* **72:** 367-382.

Konstantinidis, KT, Ramette, A & Tiedje, JM (2006). The bacterial species definition in the genomic era. *Philos Trans R Soc Lond B Biol Sci,* **361:** 1929-1940.

Kralik, P & Ricchi, M (2017). A basic guide to real time PCR in microbial diagnostics: Definitions, parameters, and everything. *Front Microbiol,* **8:** 108.

Kuczynski, J, Lauber, CL, Walters, WA, Parfrey, LW, Clemente, JC, Gevers, D & Knight, R (2011a). Experimental and analytical tools for studying the human microbiome. *Nat Rev Genet,* **13:** 47-58.

Kuczynski, J, Stombaugh, J, Walters, WA, Gonzalez, A, Caporaso, JG & Knight, R (2011b). Using QIIME to analyze 16S rRNA gene sequences from microbial communities. *Curr Protoc Bioinformatics,* **36:** 10.17.11-10.17.20.

Kumar, R, Eipers, P, Little, RB, Crowley, M, Crossman, DK, Lefkowitz, EJ & Morrow, CD (2014). Getting started with microbiome analysis: Sample acquisition to bioinformatics. *Curr Protoc Hum Genet,* **82:** 18.18.11–18.18.29.

Lagier, JC, Armougom, F, Million, M, Hugon, P, Pagnier, I, Robert, C, Bittar, F, Fournous, G, Gimenez, G, Maraninchi, M, Trape, JF, Koonin, EV, La Scola, B & Raoult, D (2012). Microbial culturomics: Paradigm shift in the human gut microbiome study. *Clin Microbiol Infect,* **18:** 1185-1193.

Lakhujani, V & Badapanda, C (2017). Prepare_taxa_charts.py: A Python program to automate generation of publication ready taxonomic pie chart images from QIIME. *Genom Data,* **12:** 97-101.

Lalloo, UG, Pillay, S, Mngqibisa, R, Abdool-Gaffar, S & Ambaram, A (2016). HIV and COPD: A conspiracy of risk factors. *Respirology,* **21:** 1166-1172.

Lange, P, Halpin, DM, O'donnell, DE & Macnee, W (2016). Diagnosis, assessment, and phenotyping of COPD: Beyond FEV_1. *Int J Chron Obstruct Pulmon Dis*, **11 Special Issue:** 3-12.

Lankelma, JM, Birnie, E, Weehuizen, TAF, Scicluna, BP, Belzer, C, Houtkooper, RH, Roelofs, J, De Vos, AF, Van Der Poll, T, Budding, AE & Wiersinga, WJ (2017). The gut microbiota as a modulator of innate immunity during melioidosis. *PLoS Negl Trop Dis*, **11:** e0005548.

Lau, JT, Whelan, FJ, Herath, I, Lee, CH, Collins, SM, Bercik, P & Surette, MG (2016). Capturing the diversity of the human gut microbiota through culture-enriched molecular profiling. *Genome Med*, **8:** 72.

Lean, C & Maclaurin, J (2016). The value of phylogenetic diversity. *In:* Pellens, R. & Grandcolas, P. (eds.) Biodiversity conservation and phylogenetic systematics: Preserving our evolutionary heritage in an extinction crisis. Cham: Springer International Publishing.

Lee, JS, Collard, HR, Raghu, G, Sweet, MP, Hays, SR, Campos, GM, Golden, JA & King, TE, Jr. (2010). Does chronic microaspiration cause idiopathic pulmonary fibrosis? *Am J Med*, **123:** 304-311.

Lee, SW, Kuan, CS, Wu, LS & Weng, JT (2016). Metagenome and metatranscriptome profiling of moderate and severe COPD sputum in Taiwanese Han males. *PLoS One*, **11:** e0159066.

Leevvenhoeck, A (1684). An abstract of a letter from Mr. Anthony Leevvenhoeck at Delft, dated sep. 17. 1683 containing some microscopical observations, about animals in the scurf of the teeth, the substance call'd worms in the nose, the cuticula consisting of scales. *Phil Trans*, **14:** 568-574.

Legendre, P (2007). Studying beta diversity: Ecological variation partitioning by multiple regression and canonical analysis. *J Plant Ecol*, **1:** 3-8.

Legendre, P, Borcard, D & Peres-Neto, PR (2005). Analyzing beta diversity: Partitioning the spatial variation of community composition data. *Ecological Monographs*, **75:** 435-450.

Legendre, P & Legendre, LF (2012). Numerical ecology, Oxford, UK, Elsevier.

Lemos, LN, Fulthorpe, RR, Triplett, EW & Roesch, LF (2011). Rethinking microbial diversity analysis in the high throughput sequencing era. *J Microbiol Methods*, **86:** 42-51.

Li, F, He, J, Wei, J, Cho, WC & Liu, X (2015). Diversity of epithelial stem cell types in adult lung. *Stem Cells International*, **2015:** 728307.

Li, K, Bihan, M & Methe, BA (2013). Analyses of the stability and core taxonomic memberships of the human microbiome. *PLoS One*, **8:** e63139.

Lloyd-Price, J, Abu-Ali, G & Huttenhower, C (2016). The healthy human microbiome. *Genome Med*, **8:** 51.

Lopes, SP, Azevedo, NF & Pereira, MO (2015). Microbiome in cystic fibrosis: Shaping polymicrobial interactions for advances in antibiotic therapy. *Crit Rev Microbiol*, **41:** 353-365.

Lopez-Campos, JL, Tan, W & Soriano, JB (2016). Global burden of COPD. *Respirology*, **21:** 14-23.

Loureiro, CC (2016). Blurred lines. Eosinophilic COPD: ACOS or COPD phenotype? *Rev Port Pneumol (2006)*, **22:** 279-282.

Lozupone, C & Knight, R (2005). UniFrac: A new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol*, **71:** 8228-8235.

Lozupone, CA, Hamady, M, Kelley, ST & Knight, R (2007). Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl Environ Microbiol*, **73:** 1576-1585.

Lozupone, CA & Knight, R (2008). Species divergence and the measurement of microbial diversity. *FEMS Microbiol Rev*, **32:** 557-578.

Lumb, AB (2016a). Airways disease. Nunn's applied respiratory physiology. Eight ed. Italy: Elsevier Health Sciences.

Lumb, AB (2016b). Diffusion of respiratory gases. Nunn's applied respiratory physiology. Eight ed. Italy: Elsevier Health Sciences.

Lysholm, F, Wetterbom, A, Lindau, C, Darban, H, Bjerkner, A, Fahlander, K, Lindberg, AM, Persson, B, Allander, T & Andersson, B (2012). Characterization of the viral microbiome in patients with severe lower respiratory tract infections, using metagenomic sequencing. *PLoS One*, **7:** e30875.

Macnee, W (2006). Pathology, pathogenesis, and pathophysiology. *BMJ,* **332:** 1202-1204.

Macnee, W, Vestbo, J & Agusti, A (2016). COPD: Pathogenisis and natural history. *In:* V. Courtney Broaddus, Robert J. Mason, Joel D. Ernst, Talmadge E. King, Stephen C. Lazarus, John F. Murray, Jay A. Nadel, Arthur S. Slutsky & Michael B. Gotway (eds.) Murray and Nadel's textbook of respiratory medicine. Sixth ed. Philadelphia: W.B. Saunders.

Magurran, AE (2013). Measuring biological diversity, Hoboken, United Kingdom, John Wiley & Sons, Incorporated.

Magurran, AE & Mcgill, BJ (2010). Biological diversity: Frontiers in measurement and assessment, Oxford, United Kingdom, OUP Oxford.

Make, BJ & Martinez, FJ (2008). Assessment of patients with chronic obstructive pulmonary disease. *Proc Am Thorac Soc,* **5:** 884-890.

Makino, W, Cotner, JB, Sterner, RW & Elser, JJ (2003). Are bacteria more like plants or animals? Growth rate and resource dependence of bacterial C : N : P stoichiometry. *Funct Ecol,* **17:** 121-130.

Malakoane, B, Heunis, JC, Chikobvu, P, Kigozi, NG & Kruger, WH (2020). Public health system challenges in the Free State, South Africa: A situation appraisal to inform health system strengthening. *BMC Health Serv Res,* **20:** 58.

Mammen, MJ & Sethi, S (2016). COPD and the microbiome. *Respirology,* **21:** 590-599.

Man, WH, De Steenhuijsen Piters, WA & Bogaert, D (2017). The microbiota of the respiratory tract: Gatekeeper to respiratory health. *Nat Rev Microbiol,* **15:** 259-270.

Marchesi, JR & Ravel, J (2015). The vocabulary of microbiome research: A proposal. *Microbiome,* **3:** 31.

Marimón, JM (2018). The lung microbiome in health and respiratory diseases. *Clin Pulm Med,* **25:** 131-137.

Marsland, BJ, Trompette, A & Gollwitzer, ES (2015). The gut-lung axis in respiratory disease. *Ann Am Thorac Soc,* **12 Suppl 2:** S150-156.

Marsland, BJ, Yadava, K & Nicod, LP (2013). The airway microbiome and disease. *Chest*, **144:** 632-637.

Martin, AP (2002). Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Appl Environ Microbiol*, **68:** 3673-3682.

Martin, C, Burgel, PR, Lepage, P, Andrejak, C, De Blic, J, Bourdin, A, Brouard, J, Chanez, P, Dalphin, JC, Deslee, G, Deschildre, A, Gosset, P, Touqui, L & Dusser, D (2015). Host-microbe interactions in distal airways: Relevance to chronic airway diseases. *Eur Respir Rev*, **24:** 78-91.

Martin, MJ & Harrison, TW (2015). Causes of chronic productive cough: An approach to management. *Respir Med*, **109:** 1105-1113.

Martin, R, Miquel, S, Langella, P & Bermudez-Humaran, LG (2014). The role of metagenomics in understanding the human microbiome in health and disease. *Virulence*, **5:** 413-423.

Martinez, FJ, Curtis, JL, Sciurba, F, Mumford, J, Giardino, ND, Weinmann, G, Kazerooni, E, Murray, S, Criner, GJ, Sin, DD, Hogg, J, Ries, AL, Han, M, Fishman, AP, Make, B, Hoffman, EA, Mohsenifar, Z, Wise, R & National Emphysema Treatment Trial Research, G (2007). Sex differences in severe pulmonary emphysema. *Am J Respir Crit Care Med*, **176:** 243-252.

Mateos, AC, Amarillo, AC, Carreras, HA & Gonzalez, CM (2018). Land use and air quality in urban environments: Human health risk assessment due to inhalation of airborne particles. *Environ Res*, **161:** 370-380.

Matheson, MC, Raven, J, Johns, DP, Abramson, MJ & Walters, EH (2007). Associations between reduced diffusing capacity and airflow obstruction in community-based subjects. *Respir Med*, **101:** 1730-1737.

Mayosi, BM & Benatar, SR (2014). Health and health care in South Africa--20 years after Mandela. *N Engl J Med*, **371:** 1344-1353.

McDonald, D, Birmingham, A & Knight, R (2015). Context and the human microbiome. *Microbiome*, **3:** 52.

McDonald, D, Clemente, JC, Kuczynski, J, Rideout, JR, Stombaugh, J, Wendel, D, Wilke, A, Huse, S, Hufnagle, J, Meyer, F, Knight, R & Caporaso, JG (2012a). The biological observation

matrix (biom) format or: How I learned to stop worrying and love the ome-ome. *Gigascience*, **1:** 7.

McDonald, D, Price, MN, Goodrich, J, Nawrocki, EP, Desantis, TZ, Probst, A, Andersen, GL, Knight, R & Hugenholtz, P (2012b). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J*, **6:** 610-618.

McGuinness, AJ & Sapey, E (2017). Oxidative stress in COPD: Sources, markers, and potential mechanisms. *J Clin Med*, **6:** 21-39.

McKeever, TM, Hearson, G, Housley, G, Reynolds, C, Kinnear, W, Harrison, TW, Kelly, AM & Shaw, DE (2016). Using venous blood gas analysis in the assessment of COPD exacerbations: A prospective cohort study. *Thorax*, **71:** 210-215.

McMurdie, PJ & Holmes, S (2013). Phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*, **8:** e61217.

Meyer, F, Paarmann, D, D'souza, M, Olson, R, Glass, EM, Kubal, M, Paczian, T, Rodriguez, A, Stevens, R, Wilke, A, Wilkening, J & Edwards, RA (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9:** 386-394.

Milic, M, Frustaci, A, Del Bufalo, A, Sánchez-Alarcón, J, Valencia-Quintana, R, Russo, P & Bonassi, S (2015). DNA damage in non-communicable diseases: A clinical and epidemiological perspective. *Mutat Res*, **776:** 118-127.

Miller, MR, Hankinson, J, Brusasco, V, Burgos, F, Casaburi, R, Coates, A, Crapo, R, Enright, P, van Der Grinten, CP, Gustafsson, P, Jensen, R, Johnson, DC, Macintyre, N, Mckay, R, Navajas, D, Pedersen, OF, Pellegrino, R, Viegi, G & Wanger, J (2005). Standardisation of spirometry. *Eur Respir J*, **26:** 319-338.

Mitchell, AB & Glanville, AR (2018). The human respiratory microbiome: Implications and impact. *Semin Respir Crit Care Med*, **39:** 199-212.

Mobed, A, Baradaran, B, Guardia, Mdl, Agazadeh, M, Hasanzadeh, M, Rezaee, MA, Mosafer, J, Mokhtarzadeh, A & Hamblin, MR (2019). Advances in detection of fastidious bacteria: From

microscopic observation to molecular biosensors. *TrAC Trends in Analytical Chemistry*, **113:** 157-171.

Moore, VC (2012). Spirometry: Step by step. *Breathe*, **8:** 232-240.

Moorhouse, M, Maartens, G, Venter, WDF, Moosa, MY, Steegen, K, Jamaloodien, K, Fox, MP & Conradie, F (2019). Third-line antiretroviral therapy program in the South African public sector: Cohort description and virological outcomes. *J Acquir Immune Defic Syndr*, **80:** 73-78.

Morgan, XC & Huttenhower, C (2012). Chapter 12: Human microbiome analysis. *PLoS Comput Biol*, **8:** e1002808.

Morisita, M (1959). Measuring of interspecific association and similarity between communities. *Mem. Fac. Sci. Kyushu Univ. Series E*, **3:** 65-80.

Morris, EK, Caruso, T, Buscot, F, Fischer, M, Hancock, C, Maier, TS, Meiners, T, Muller, C, Obermaier, E, Prati, D, Socher, SA, Sonnemann, I, Waschke, N, Wubet, T, Wurst, S & Rillig, MC (2014). Choosing and using diversity indices: Insights for ecological applications from the German biodiversity exploratories. *Ecol Evol*, **4:** 3514-3524.

Muller, PH, De Meij, TGJ, Westedt, M, De Groot, EFJ, Allaart, CF, Brinkman, DMC, Schonenberg-Meinema, D, Van Den Berg, M, Van Suijlekom-Smit, LWA, Van Rossum, M, Budding, AE & Ten Cate, R (2017). Disturbance of microbial core species in new-onset juvenile idiopathic arthritis. *J Pediat Infect Dis*, **12:** 131-135.

Mullerova, H, Lu, C, Li, H & Tabberer, M (2014). Prevalence and burden of breathlessness in patients with chronic obstructive pulmonary disease managed in primary care. *PLoS One*, **9:** e85540.

Navas-Molina, JA, Peralta-Sanchez, JM, Gonzalez, A, Mcmurdie, PJ, Vazquez-Baeza, Y, Xu, Z, Ursell, LK, Lauber, C, Zhou, H, Song, SJ, Huntley, J, Ackermann, GL, Berg-Lyons, D, Holmes, S, Caporaso, JG & Knight, R (2013). Advancing our understanding of the human microbiome using QIIME. *Methods Enzymol*, **531:** 371-444.

Ng Kee Kwong, F, Nicholson, AG, Harrison, CL, Hansbro, PM, Adcock, IM & Chung, KF (2017). Is mitochondrial dysfunction a driving mechanism linking COPD to nonsmall cell lung carcinoma? *Eur Respir Rev*, **26:** 170040.

Nguyen, NP, Warnow, T, Pop, M & White, B (2016). A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *NPJ Biofilms Microbiomes,* **2:** 16004.

Nicod, LP (2005). Lung defences: An overview. *Eur Respir Rev,* **14:** 45-50.

O'Dwyer, DN, Dickson, RP & Moore, BB (2016). The lung microbiome, immunity, and the pathogenesis of chronic lung disease. *J Immunol,* **196:** 4839-4847.

O'Riordan, TG & Smaldone, GC (2016). Aerosol deposition and clearance. *In:* V. Courtney Broaddus, Robert J. Mason, Joel D. Ernst, Talmadge E. King, Stephen C. Lazarus, John F. Murray, Jay A. Nadel, Arthur S. Slutsky & Michael B. Gotway (eds.) Murray and Nadel's textbook of respiratory medicine. Sixth ed. Philadelphia: W.B. Saunders.

Oliveira, AS, Munha, J, Bugalho, A, Guimaraes, M, Reis, G, Marques, A & GI DPOC Grupo de Interesse na Doenca Pulmonar Obstrutiva Cronica (2018). Identification and assessment of COPD exacerbations. *Pulmonology,* **24:** 42-47.

Palange, P & Simonds, AK (2013). ERS handbook of respiratory medicine, European Respiratory Society.

Paliy, O & Shankar, V (2016). Application of multivariate statistical techniques in microbial ecology. *Mol Ecol,* **25:** 1032-1057.

Pallen, MJ, Loman, NJ & Penn, CW (2010). High-throughput sequencing and clinical microbiology: Progress, opportunities and challenges. *Curr Opin Microbiol,* **13:** 625-631.

Palmer, MW (1990). The estimation of species richness by extrapolation. *Ecology,* **71:** 1195-1198.

Papaioannou, AI, Loukides, S, Gourgoulianis, KI & Kostikas, K (2009). Global assessment of the COPD patient: Time to look beyond fev1? *Respir Med,* **103:** 650-660.

Parada, AE, Needham, DM & Fuhrman, JA (2016). Every base matters: Assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ Microbiol,* **18:** 1403-1414.

Pefura-Yone, EW, Kengne, AP, Balkissou, AD, Magne-Fotso, CG, Ngo-Yonga, M, Boulleys-Nana, JR, Efe-De-Melingui, NR, Ndjeutcheu-Moualeu, PI, Mbele-Onana, CL, Kenmegne-

Noumsi, EC, Kolontchang-Yomi, BL, Theubo-Kamgang, BJ, Ebouki, ER, Djuikam-Kamga, CK, Amougou, F, Mboumtou, L, Petchou-Talla, EL, Kuaban, C & Respiratory Health Survey Group in Cameroon (RHSGC) (2016). Prevalence of obstructive lung disease in an African country using definitions from different international guidelines: A community based cross-sectional survey. *BMC Res Notes*, **9:** 124.

Pellegrino, R, Viegi, G, Brusasco, V, Crapo, RO, Burgos, F, Casaburi, R, Coates, A, Van Der Grinten, CP, Gustafsson, P, Hankinson, J, Jensen, R, Johnson, DC, Macintyre, N, Mckay, R, Miller, MR, Navajas, D, Pedersen, OF & Wanger, J (2005). Interpretative strategies for lung function tests. *Eur Respir J*, **26:** 948-968.

Plevkova, J, Buday, T, Kavalcikova-Bogdanova, N, Ioan, I & Demoulin-Alexikova, S (2017). Sex differences in cough reflex. *Respir Physiol Neurobiol*, **245:** 122-129.

Plummer, E & Twin, J (2015). A comparison of three bioinformatics pipelines for the analysis of preterm gut microbiota using 16S rRNA gene sequencing data. *Journal of Proteomics & Bioinformatics*, **8**.

Podani, J, Ricotta, C & Schmera, D (2013). A general framework for analyzing beta diversity, nestedness and related community-level phenomena based on abundance data. *Ecol. Complex*, **15:** 52-61.

Pollock, J, Glendinning, L, Wisedchanwet, T & Watson, M (2018). The madness of microbiome: Attempting to find consensus "best practice" for 16S microbiome studies. *Appl Environ Microbiol*, **84:** e02627-02617.

Porter, JR (1976). Antony van leeuwenhoek: Tercentenary of his discovery of bacteria. *Bacteriol Rev*, **40:** 260-269.

Postma, DS & Rabe, KF (2015). The asthma-COPD overlap syndrome. *N Engl J Med*, **373:** 1241-1249.

Presti, RM, Flores, SC, Palmer, BE, Atkinson, JJ, Lesko, CR, Lau, B, Fontenot, AP, Roman, J, Mcdyer, JF & Twigg, HL, 3rd (2017). Mechanisms underlying HIV-associated noninfectious lung disease. *Chest*, **152:** 1053-1060.

Price, DB, Yawn, BP & Jones, RC (2010). Improving the differential diagnosis of chronic obstructive pulmonary disease in primary care. *Mayo Clin Proc*, **85:** 1122-1129.

Pruesse, E, Quast, C, Knittel, K, Fuchs, BM, Ludwig, W, Peplies, J & Glockner, FO (2007). Silva: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res*, **35:** 7188-7196.

Qin, J, Li, R, Raes, J, Arumugam, M, Burgdorf, KS, Manichanh, C, Nielsen, T, Pons, N, Levenez, F, Yamada, T, Mende, DR, Li, J, Xu, J, Li, S, Li, D, Cao, J, Wang, B, Liang, H, Zheng, H, Xie, Y, Tap, J, Lepage, P, Bertalan, M, Batto, JM, Hansen, T, Le Paslier, D, Linneberg, A, Nielsen, HB, Pelletier, E, Renault, P, Sicheritz-Ponten, T, Turner, K, Zhu, H, Yu, C, Li, S, Jian, M, Zhou, Y, Li, Y, Zhang, X, Li, S, Qin, N, Yang, H, Wang, J, Brunak, S, Dore, J, Guarner, F, Kristiansen, K, Pedersen, O, Parkhill, J, Weissenbach, J, Meta, HITC, Bork, P, Ehrlich, SD & Wang, J (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464:** 59-65.

Rabiu, BA & Gibson, GR (2002). Carbohydrates: A limit on bacterial diversity within the colon. *Biol Rev*, **77:** 443-453.

Ramette, A (2007). Multivariate analyses in microbial ecology. *FEMS Microbiol Ecol*, **62:** 142-160.

Rempala, GA & Seweryn, M (2013). Methods for diversity and overlap analysis in t-cell receptor populations. *J Math Biol*, **67:** 1339-1368.

Reyes, GR & Kim, JP (1991). Sequence-independent, single-primer amplification (SISPA) of complex DNA populations. *Molecular and Cellular Probes*, **5:** 473-481.

Rhoads, A & Au, KF (2015). PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics*, **13:** 278-289.

Robson, A (2017). Dyspnoea, hyperventilation and functional cough: A guide to which tests help sort them out. *Breathe*, **13:** 45-50.

Rodriguez, MZ, Comin, CH, Casanova, D, Bruno, OM, Amancio, DR, Costa, LDF & Rodrigues, FA (2019). Clustering algorithms: A comparative approach. *PLoS One*, **14:** e0210236.

Rogers, GB, Hart, CA, Mason, JR, Hughes, M, Walshaw, MJ & Bruce, KD (2003). Bacterial diversity in cases of lung infection in cystic fibrosis patients: 16S ribosomal DNA (rDNA) length heterogeneity PCR and 16S rDNA terminal restriction fragment length polymorphism profiling. *Journal of Clinical Microbiology,* **41:** 3548-3558.

Ruso, S, Marco, FM, Martínez-Carbonell, JA & Carratalá, JA (2015). Bacterial vaccines in chronic obstructive pulmonary disease: Effects on clinical outcomes and cytokine levels. *APMIS,* **123:** 556-561.

Rutten, NB, Gorissen, DM, Eck, A, Niers, LE, Vlieger, AM, Besseling-Van Der Vaart, I, Budding, AE, Savelkoul, PH, Van Der Ent, CK & Rijkers, GT (2015). Long term development of gut microbiota composition in atopic children: Impact of probiotics. *PLoS One,* **10:** e0137681.

Salvi, S (2015). The silent epidemic of COPD in Africa. *The Lancet Global Health,* **3:** e6-e7.

Sarioglu, N, Hismiogullari, AA, Bilen, C & Erel, F (2016). Is the COPD assessment test (CAT) effective in demonstrating the systemic inflammation and other components in COPD? *Rev Port Pneumol (2006),* **22:** 11-17.

Savage, DC (1977). Microbial ecology of the gastrointestinal tract. *Annu Rev Microbiol,* **31:** 107-133.

Scales, BS, Dickson, RP & Huffnagle, GB (2016). A tale of two sites: How inflammation can reshape the microbiomes of the gut and lungs. *J Leukoc Biol,* **100:** 943-950.

Scheiermann, J & Klinman, DM (2017). Three distinct pneumotypes characterize the microbiome of the lung in BALB/cJ mice. *PLoS One,* **12:** e0180561.

Schloss, PD, Westcott, SL, Ryabin, T, Hall, JR, Hartmann, M, Hollister, EB, Lesniewski, RA, Oakley, BB, Parks, DH, Robinson, CJ, Sahl, JW, Stres, B, Thallinger, GG, Van Horn, DJ & Weber, CF (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol,* **75:** 7537-7541.

Schroeder, PJ & Jenkins, DG (2018). How robust are popular beta diversity indices to sampling error? *Ecosphere,* **9:** e02100.

Sedlar, K, Videnska, P, Skutkova, H, Rychlik, I & Provaznik, I (2016). Bipartite graphs for visualization analysis of microbiome data. *Evol Bioinform Online*, **12:** 17-23.

Sender, R, Fuchs, S & Milo, R (2016). Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol*, **14:** e1002533.

Shannon, CE (1984). A mathematical theory of communication. *Bell Syst Tech J.*, **27:** 379-423.

Shisana, O, Rehle, T, Simbayi, LC, Zuma, K, Jooste, S, Zungu, N, Labadarios, D & Onoya, D (2014). South African national HIV prevalence, incidence and behaviour survey, 2012, Cape Town, HSRC Press.

Shukla, SD, Budden, KF, Neal, R & Hansbro, PM (2017). Microbiome effects on immunity, health and disease in the lung. *Clin Transl Immunology*, **6:** e133.

Simkhovich, BZ, Kleinman, MT & Kloner, RA (2008). Air pollution and cardiovascular injury epidemiology, toxicology, and mechanisms. *J Am Coll Cardiol*, **52:** 719-726.

Simpson, EH (1949). Measurement of diversity. *Nature*, **163:** 688.

Sørensen, T (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biol. Skr.*, **5:** 1-34.

Spandole, S, Cimponeriu, D, Berca, LM & Mihaescu, G (2015). Human anelloviruses: An update of molecular, epidemiological and clinical aspects. *Arch Virol*, **160:** 893-908.

Stackebrandt, E & Goebel, BM (1994). 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*, **44:** 846-849.

Standring, S (2015). Pleura, lungs, trachea and bronchi. *In:* Standring, S. (ed.) Gray's anatomy e-book: The anatomical basis of clinical practice. 41 ed.: Elsevier Health Sciences.

Statistics South Africa. (2018). *Mid-year population estimastes - 2017* [Online]. Available: www.statssa.gov.za/publications/P0302/P303022018.pdf [Accessed 27 January 2019].

Strathdee, F & Free, A (2013). Denaturing gradient gel electrophoresis (DGGE). DNA electrophoresis. Springer.

Stubbendieck, RM, Vargas-Bautista, C & Straight, PD (2016). Bacterial communities: Interactions to scale. *Front Microbiol*, **7:** 1234.

Sze, MA, Dimitriu, PA, Hayashi, S, Elliott, WM, Mcdonough, JE, Gosselink, JV, Cooper, J, Sin, DD, Mohn, WW & Hogg, JC (2012). The lung tissue microbiome in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*, **185:** 1073-1080.

Sze, MA, Hogg, JC & Sin, DD (2014). Bacterial microbiome of lungs in COPD. *Int J Chron Obstruct Pulmon Dis*, **9:** 229-238.

Tang, ZZ, Chen, G & Alekseyenko, AV (2016). PERMANOVA-s: Association test for microbial community composition that accommodates confounders and multiple distances. *Bioinformatics*, **32:** 2618-2625.

Taylor, SL, Wesselingh, S & Rogers, GB (2016). Host-microbiome interactions in acute and chronic respiratory infections. *Cell Microbiol*, **18:** 652-662.

The International Committee on Taxonomy of Viruses (ICTV) (2012). Part II. The viruses. *In:* King, A. M. Q., Adams, M. J., Carstens, E. B. & Lefkowitz, E. J. (eds.) Virus taxonomy. San Diego: Elsevier.

Thurber, RV, Haynes, M, Breitbart, M, Wegley, L & Rohwer, F (2009). Laboratory procedures to generate viral metagenomes. *Nat Protoc*, **4:** 470-483.

Tremblay, J, Singh, K, Fern, A, Kirton, ES, He, S, Woyke, T, Lee, J, Chen, F, Dangl, JL & Tringe, SG (2015). Primer and platform effects on 16S rRNA tag sequencing. *Front Microbiol*, **6:** 771.

Tuomisto, H & Ruokolainen, K (2006). Analyzing or explaining beta diversity? Understanding the targets of different methods of analysis. *Ecology*, **87:** 2697-2708.

Twigg, HL, 3rd, Weinstock, GM & Knox, KS (2017). Lung microbiome in human immunodeficiency virus infection. *Transl Res*, **179:** 97-107.

Ursell, LK, Metcalf, JL, Parfrey, LW & Knight, R (2012). Defining the human microbiome. *Nutr Rev*, **70 Suppl 1:** S38-44.

van Dijk, EL, Auger, H, Jaszczyszyn, Y & Thermes, C (2014). Ten years of next-generation sequencing technology. *Trends Genet*, **30:** 418-426.

van Leeuwenhoeck, A (1677). Observations, communicated to the publisher by Mr. Antony van leewenhoeck, in a Dutch letter of the 9th octob. 1676. Here english'd: Concerning little animals by him observed in rain-well-sea- and snow water; as also in water wherein pepper had lain infused. *Phil Trans*, **12:** 821-831.

van Leeuwenhoek, A (1682). A letter from m. Anthony leeuwenhoeck, in answer to some former letters sent to him by r. H. Containing some further observations made by him, about the fabrick and texture of the fibres of muscles-communicated to the royal society. *In:* Hooke, R. (ed.) Philosophical collections.

Vaughan, EE, Schut, F, Heilig, HGHJ, Zoetendal, EG, De Vos, WM & Akkermans, ADL (2000). A molecular view of the intestina ecosystem. *Current Issues in Intestinal Microbiology*, **1:** 1-12.

Vazquez-Baeza, Y, Pirrung, M, Gonzalez, A & Knight, R (2013). EMPEROR: A tool for visualizing high-throughput microbial community data. *Gigascience*, **2:** 16.

Vecchio-Pagan, B, Bewick, S, Mainali, K, Karig, DK & Fagan, WF (2017). A stoichioproteomic analysis of samples from the human microbiome project. *Front Microbiol*, **8:** 1119.

Venkataraman, A, Bassis, CM, Beck, JM, Young, VB, Curtis, JL, Huffnagle, GB & Schmidt, TM (2015). Application of a neutral community model to assess structuring of the human lung microbiome. *MBio*, **6:** e02284-02214.

Vogelmeier, CF, Criner, GJ, Martinez, FJ, Anzueto, A, Barnes, PJ, Bourbeau, J, Celli, BR, Chen, R, Decramer, M, Fabbri, LM, Frith, P, Halpin, DM, Lopez Varela, MV, Nishimura, M, Roche, N, Rodriguez-Roisin, R, Sin, DD, Singh, D, Stockley, R, Vestbo, J, Wedzicha, JA & Agusti, A (2017). Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease 2017 report: Gold executive summary. *Am J Respir Crit Care Med*, **195:** 557-582.

Waatevik, M, Johannessen, A, Gomez Real, F, Aanerud, M, Hardie, JA, Bakke, PS & Lind Eagan, TM (2016). Oxygen desaturation in 6-min walk test is a risk factor for adverse outcomes in COPD. *Eur Respir J*, **48:** 82-91.

Wagner, BD, Grunwald, GK, Zerbe, GO, Mikulich-Gilbertson, SK, Robertson, CE, Zemanick, ET & Harris, JK (2018). On the use of diversity measures in longitudinal sequencing studies of microbial communities. *Front Microbiol*, **9:** 1037.

Wang, Y & Salazar, JK (2016). Culture-independent rapid detection methods for bacterial pathogens and toxins in food matrices. *Comprehensive Reviews in Food Science and Food Safety*, **15:** 183-205.

Washko, GR (2010). Diagnostic imaging in COPD. *Semin Respir Crit Care Med*, **31:** 276-285.

Waugh, C, Cromer, D, Grimm, A, Chopra, A, Mallal, S, Davenport, M & Mak, J (2015). A general method to eliminate laboratory induced recombinants during massive, parallel sequencing of cDNA library. *Virol J*, **12:** 55.

Weinstock, GM (2012). Genomic approaches to studying the human microbiota. *Nature*, **489:** 250-256.

Wilkinson, TM, Donaldson, GC, Hurst, JR, Seemungal, TA & Wedzicha, JA (2004). Early therapy improves outcomes of exacerbations of chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*, **169:** 1298-1303.

Williams, OW, Sharafkhaneh, A, Kim, V, Dickey, BF & Evans, CM (2006). Airway mucus: From production to secretion. *Am J Respir Cell Mol Biol*, **34:** 527-536.

Willis, AD (2019). Rarefaction, alpha diversity, and statistics. *Front Microbiol*, **10:** 2407.

Wilson, KH & Blitchington, RB (1996). Human colonic biota studied by ribosomal DNA sequence analysis. *Appl Environ Microbiol*, **62:** 2273-2278.

Winter, SE & Baumler, AJ (2014). Dysbiosis in the inflamed intestine: Chance favors the prepared microbe. *Gut Microbes*, **5:** 71-73.

Woese, CR & Fox, GE (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc Natl Acad Sci U S A*, **74:** 5088-5090.

Wommack, KE, Bhavsar, J, Polson, SW, Chen, J, Dumas, M, Srinivasiah, S, Furman, M, Jamindar, S & Nasko, DJ (2012). Virome: A standard operating procedure for analysis of viral metagenome sequences. *Stand Genomic Sci*, **6:** 427-439.

Wong, RG, Wu, JR & Gloor, GB (2016). Expanding the UniFrac toolbox. *PLoS One*, **11:** e0161196.

Wylie, KM, Weinstock, GM & Storch, GA (2012). Emerging view of the human virome. *Transl Res*, **160:** 283-290.

Xia, Y & Sun, J (2017). Hypothesis testing and statistical analysis of microbiome. *Genes Dis*, **4:** 138-148.

Yamasaki, K & Eeden, SFV (2018). Lung macrophage phenotypes and functional responses: Role in the pathogenesis of COPD. *Int J Mol Sci*, **19:** 582-594.

Zaura, E (2012). Next-generation sequencing approaches to understanding the oral microbiome. *Adv Dent Res*, **24:** 81-85.

Zaura, E, Nicu, EA, Krom, BP & Keijser, BJ (2014). Acquiring and maintaining a normal oral microbiome: Current perspective. *Front Cell Infect Microbiol*, **4:** 85.

Zhao, N, Chen, J, Carroll, IM, Ringel-Kulka, T, Epstein, MP, Zhou, H, Zhou, JJ, Ringel, Y, Li, H & Wu, MC (2015). Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *Am J Hum Genet*, **96:** 797-807.

Zhao, X, Zhong, J, Wei, C, Lin, CW & Ding, T (2017). Current perspectives on viable but non-culturable state in foodborne pathogens. *Front Microbiol*, **8:** 580.

Zhou, J, He, Z, Yang, Y, Deng, Y, Tringe, SG & Alvarez-Cohen, L (2015). High-throughput metagenomic technologies for complex microbial community analysis: Open and closed formats. *MBio*, **6**.

Zhou, Y, Gao, H, Mihindukulasuriya, KA, La Rosa, PS, Wylie, KM, Vishnivetskaya, T, Podar, M, Warner, B, Tarr, PI, Nelson, DE, Fortenberry, JD, Holland, MJ, Burr, SE, Shannon, WD, Sodergren, E & Weinstock, GM (2013). Biogeography of the ecosystems of the healthy human body. *Genome Biol*, **14:** R1.

Zoetendal, EG, Collier, CT, Koike, S, Mackie, RI & Gaskins, HR (2004). Molecular ecological analysis of the gastrointestinal microbiota: A review. *J Nutr*, **134:** 465-472.

CHAPTER 3

Basic overview of the methods used in the statistical analysis of microbiome studies

The editorial style of Critical Reviews in Microbiology was followed in this chapter

(Excerpts from Chapter 2 can be found in Chapter 3)

Abstract

The decreasing cost of sequencing has increased the number of researchers studying the microbiome and the amount of data that is generated. The rising number of microbiome studies warrants a thorough understanding of the statistical methods that are used to analyse microbiome data, to ensure transparency, quality and generalisability of results.

A microbiome study has methodological steps whereby sequencing reads are generated and analysed. The final output from sequence analysis programs (after quality control and clustering) is an operational taxonomic unit (OTU) table, which displays the abundance of each OTU. With the advance of sequencing techniques, data outputs have expanded from simple descriptive observations to diversity measures that determine the differences in the microbiome within groups (alpha diversity) and between groups (beta diversity).

This review provides a critical overview of the appropriate application of the various statistical methods that are used in microbiome studies. Guidance on the use of the different alpha and beta diversity measures is provided, highlighting the advantages and disadvantages of each measure, followed by a discussion of multivariate analysis of microbiome data. The review is concluded with the observation that a large variety of statistical measures is used and that further standardisation of analysis methods is warranted. While other reviews have discussed these topics in detailed, this review is the first review to provide a basic overview of the different methods used in the analysis of microbiome studies for readers with no statistical background knowledge.

Keywords: Microbiome; statistical analysis; alpha diversity; beta diversity; multivariate analysis;

3.1 Introduction

Microorganisms are among the most abundant organisms on Earth, with bacteria accounting for 15% of the Earth's biomass and can persist in a wide variety of habitats including the human body [1, 2]. The microbiome is defined as all the microorganisms along with their genetic material (i.e. genomes) that are found in a specific environment, for example in the human gastrointestinal tract or lungs [3-6]. These microbiomes are complex and often provide essential functions for that particular environment [7-9]. In humans, the microbiome is an important component of host immunity and host metabolism [7-9]. The microbiome, in turn, is influenced by several host factors, such as the local environment (e.g. available nutrients, pH, temperature, etc. in the human lung) and movement of microorganisms in and out of the environment (e.g. movement of microorganisms out of the human gastrointestinal tract and into human lung) [10-13].

The advance of next-generation sequencing (NGS) technologies has enabled in-depth analysis of microbiomes; the study of microbial communities can be referred to as microbial ecology [14-17]. Microbial ecology has two important components: i) the diversity of the community and ii) the function of the community [14]. There are two types of studies that can be conducted to determine the diversity of the microbial environment: i) targeted metagenomics studies or ii) shotgun metagenomics studies; shotgun metagenomics has the added benefit of being able to determine the function of the microbial environment as well (if required) [18].

Several different analysis and statistical measures are available, which can be used to determine the diversity of the microbiome. However, if researchers do not have an adequate understanding of the methods, choosing an appropriate approach, is difficult and may result in incorrect conclusions [19].

This review provides an overview of the factors that should be taken into consideration when analysing microbiome data. The statistical approaches to the different microbiome diversity measures are discussed, including the advantages and disadvantages of each method, and guidance is provided on the appropriate use of certain measures. Statistical tools, including multivariate analysis, for microbiome data are reviewed. A roadmap with detailed steps to guide and support researchers to conduct specific types of microbiome analysis is provided (Figure 3.1). Table 3.1 provides a glossary of the terms used in this study.

3.2 Conducting a microbiome study

The first step to ensure that a study generates meaningful data is developing an appropriate research question with clear measurable objectives [20-22]. The research hypothesis drives the experimental design; one of the most important concerns in a microbiome study [20-22]. Several factors should be taken into consideration for a high-quality experimental design including the use of appropriate controls (positive and negative controls), sample collection, metadata, possible confounders and DNA extraction procedures (technical variation) [20-22]. Each of the above factors introduces technical variation in the study (and in some instances biological variation), which has been shown to impact on the bacterial composition that is measured in the microbiome and downstream analyses [23]. Two major approaches can be used to study the microbiome: i) the targeted approach where a specific region of the microbial genome is studied, i.e. targeted metagenomics approach or ii) an untargeted approach where all the microbial genetic material is analysed, i.e. shotgun metagenomics approach [21, 24, 25].

3.3 Analysis of microbiome data

Any microbiome study will include the following steps: i) sample collection and storage, ii) total bacterial DNA extraction (direct extraction from the sample) and iii) sequencing analysis [22]. In the case of the targeted metagenomics approach, a variable region of the 16S rRNA gene, such as the V1-V3 region is sequenced [26]. This region of the 16S rRNA gene has the highest similarity to the full-length sequence of the 16S rRNA gene and is, therefore, one of the more popular choices for sequencing using the 27F and 518R primers and the MiSeq platform (Illumina, USA) [27-29]. Several other primers and platforms can be used for 16S rRNA sequencing as summarised by Tremblay *et al.* (2015) [26]. With the untargeted approach i.e. shotgun metagenomics, all DNA present is sequenced [30].

3.3.1 Analysis of data generated from the targeted metagenomics approach

With the targeted approach, all bacterial DNA in the sample is extracted and specific regions, such as the V1-V3 region are amplified [22, 31]. After amplification using the specific primers, the library preparation is performed and this step includes the addition of adapters, indices and barcodes [32, 33]. Library preparation is followed by sequencing [32]. Next-generation sequencing, such as sequencing using the MiSeq platform (Illumina, USA) uses the adapters added during the library preparation phase to bind to oligonucleotides present on the company's propriety flow cell, enzymes add nucleotides to the fragments of DNA on the cell and reversible

dye-terminator nucleotides are briefly washed over the cell (with enough time to attach and with the excess nucleotides washed away); this process is repeated for several cycles [34].

After sequencing is completed, the data are available in one of four formats: i) demultiplexed single-end fastq file, ii) multiplexed single-end fastq file, iii) demultiplexed paired-end fastq file or iv) multiplexed paired-end fastq file [35-37]. Using programs, such as quantitative insights into microbial ecology (QIIME), QIIME2 or mothur, these sequences are demultiplexed (if multiplexed) or joined (if paired-end) [31, 38-41]. The overall data quality is improved by trimming the length of low-quality sequences, discarding short sequences and removal of chimeric sequences (sequences that formed from two different microbes) and singletons (sequence only observed once) [31, 38-43]. These processes result in a final output of the analysis, which is an operational taxonomic unit (OTU) table (with QIIME 2, this table is also referred to as a feature table) [40, 44, 45]. This table shows the abundance of each OTU within each sample in the dataset and is generated in a Biological Observation Matrix (BIOM) format (default file format for QIIME) [40, 44, 46-48]. Several software packages can use this .biom file to perform further analysis using diversity measures and to visualise the data as described in more detail later (see Figure 3.1).

3.3.2 Analysis of data generated using a shotgun metagenomics approach

As mentioned previously with a shotgun metagenomics approach (untargeted approach), all the DNA present in a sample is sequenced [30]. Similar to the targeted approach, all DNA and/or RNA in a sample is extracted [32, 49]. Extraction is followed by cDNA synthesis (for single-stranded DNA and RNA) and fragmentation [32, 49]. After fragmentation, library preparation (with adapters and barcodes) and sequencing are performed [32].

The analysis of metagenomics sequencing data is an expanding field, however, there is no standardisation and the workflow for the analysis of shotgun metagenomics (is different from the targeted metagenomics (16S rRNA sequencing) [15, 20, 21, 50, 51]. However, the first step for both methods is the same: quality control; with the shotgun metagenomics approach, this includes the filtering of low-quality sequences, demultiplexing and removal of adaptors [21, 50, 52, 53]. One of the problems with shotgun metagenomic sequencing is the presence of host DNA [54]. The removal of reads associated with host DNA is often the second step in the shotgun metagenomic analysis. After the removal of host DNA, sequences can be analysed in one of two ways: i) read-based profiling (a sequence read refers to the DNA characters in the

sequence) i.e. comparative assembly or ii) assembly-based profiling i.e. *de novo* assembly [20, 32, 52, 53, 55, 56]. The read-based profiling performs taxonomic classification by mapping the reads directly to genomes or marker genes, whereas the assembly-based profiling first assembles the shorter reads into contigs (longer continuous sequences) that are subsequently binned (sorted) by similarity and assembled to genomes or annotated contigs (in databases) [20, 52, 57]. Several programs can be used to analyse metagenomics data including metagenomic rapid annotations using subsystems technology (MG-RAST) (phylogenetic and functional analysis of metagenomes), CLC genomics workbench (using the microbial genomics module), MEGAHIT (assembly) and METAGEN (binning) and Kraken (taxonomic classification/binning) [20, 21, 52, 53, 57, 58]. These programs generate matrices (tables with rows and columns) that can be used to determine alpha and beta diversity [20, 59, 60].

3.3.3 Challenges of microbiome data

There are several aspects of microbiome data that makes analysis challenging [61]. These aspects include the following: i) each sample may have a different library size i.e. a different number of sequences are present in the samples, ii) there may be zero counts present in the data and iii) the total number of reads does not accurately reflect the absolute number of microorganisms present [61]. One of the ways in which these “problems” are overcome is by normalising the data [61].

3.4 Normalisation of data and rarefaction

One such method of normalisation is rarefaction. Rarefaction adjusts for differences in library sizes by selecting a threshold (equal to or less than the smallest number of reads) and randomly discards sequences from the larger samples (with more reads); this method is essentially random subsampling [62, 63]. However, the usefulness of rarefaction has been questioned and has been considered unnecessary for microbiome studies; McMurdie and Holmes (2014) have stated that the use of rarefaction is inadmissible as it omits valid data [62, 63].

There are several other ways in which the data can be normalised. These methods include i) scaling the read counts by the total number of reads and ii) converting the data to relative abundance [20].

3.5 Diversity measures used in microbiome studies

There are two diversity measures of importance in microbiome studies: alpha diversity and beta diversity [22, 46, 47, 64-66]. Alpha diversity refers to the bacterial diversity within a single sample while beta diversity describes the diversity between samples [20]. The alpha diversity provides information on how complex a sample is, i.e. the more bacteria there is in a sample (higher alpha diversity), the more interactions occur within the sample, whereas beta diversity shows how similar the different samples are to each other in terms of their bacterial composition [6, 67, 68].

The research question determines which diversity measure(s) is appropriate for data analysis [46]. Selection of the appropriate measure(s) for analysis is based on the following study characteristics: i) is the aim of the study to test for alpha diversity or beta diversity? ii) is the presence/absence of particular taxa the only information required or is the abundance important? (qualitative measures vs quantitative measures) and iii) are all taxa regarded as equally related to each other or are the taxa considered divergently related; i.e. not all species are equally related to each other [species (taxon)-based measures vs divergent (phylogenetic)-based measures] [64, 69].

3.5.1 Alpha diversity

Alpha diversity measures provide information on how diverse a single sample is and this can be compared to other samples; it is useful when comparing a diseased individual to a healthy individual to determine if the diseased individual's microbiome is less or more diverse [64]. However, even if two communities have similar alpha diversity measures, it does not mean that the two communities share the same taxa [70]. Beta diversity measures show the number of shared species between communities [64]. When deciding whether to use qualitative (presence/absence) or quantitative measures, the following points should be taken into consideration: i) quantitative measures are most useful when the data has a strong environmental filter (if subtle changes occur, qualitative measures are unable to take note of the difference) and ii) qualitative measures are most useful when rare species are present; with presence/absence data rare species are given the same weight as common species and as a result rare species are emphasised [71, 72]. A phylogenetic approach would provide more evolutionary information; however, when studying a new environment, there may be a new taxon whose lineage has not been defined [73, 74]. In this instance, it would be more appropriate to use a taxon-based approach [73, 74]

The most used statistical measures used for alpha diversity are Chao1, the Shannon index and the Simpson index [75]. According to Morris *et al.* (2014), an ideal alpha diversity measure does not exist and each alpha diversity measure interprets results differently, however, by using more than one alpha diversity measure, a more complete understanding of the interactions within the community may be possible [75]. The Chao1 measure gives more weight to rare species, whereas the Shannon index and the Simpson index give more weight to the common species [76, 77]. The Shannon index is more sensitive to the number of different species in the community (richness) whereas the Simpson index is more sensitive to the relative abundance of the species (evenness) [78]. Table 3.2 summarises the advantages and disadvantages of each statistical method to measure alpha diversity.

However, these measures (particularly the Shannon and Simpson indices) are not very intuitive and are often difficult to compare and interpret [79]. A solution to this problem is the usage of Hill numbers; these were created by Hill (1973) and were re-introduced to the field of microbial ecology by Jost (2007) [80-82]. The advantage of Hill numbers is that these numbers: i) obey the replication principle, ii) are intuitive, iii) can easily convert the Shannon and Simpson indices and iv) allow comparisons between studies [81, 83]. Additionally, these numbers are more sensitive to rare species/OTUs [79]. The Hill numbers use a scaling parameter i.e. q, that is referred to as the order of diversity. Three q values are important: i) when q=0, this is equivalent to species richness and rare OTUs are favoured with this value, ii) when q=1, this is equivalent to the exponential Shannon index, both abundant and rare species are given equal value and ii) when q=2, this is equivalent to the inverse of the Simpson index and abundant OTUs are favoured with this value [79]. While the Hill numbers are most used with Shannon and Simpson indices, these effective numbers have also been applied to phylogenetic alpha diversity measures and beta diversity measures [81, 84]

3.5.2 Beta diversity

The Bray-Curtis, unweighted UniFrac and weighted UniFrac are the preferred statistical tools for the measurement of beta diversity, used in conjunction with multivariate analysis [19, 85]. Table 3.3 shows the various beta-diversity measures that can be used to study the microbiome and Figure 3.2 provides information on how to choose a beta diversity measure in the context of different study designs.

Beta diversity measures provide information on whether there are variations in microbial composition between different populations or groups, but this measure is unable to identify the factors that are responsible for such variation [86, 87]. Variations between populations, if present, may be caused by i) biological interactions within the community, ii) environmental conditions (another variable) or iii) random variation (no known cause for the variation) [88]. The best approach to understanding the variation in beta diversity is to perform multivariate analysis [86].

3.6 Multivariate analysis of microbiome data to understand variation in beta diversity

In the literature, the term multivariate analysis is used interchangeable with the term multivariable analysis [89]. However, the two terms have different meanings [89]. Multivariate analysis involves the analysis of multiple outcomes whereas multivariable analysis involves the analysis of a single outcome with multiple variables. One of the reasons that these terms are used is interchangeably that microbiome data is inherently multivariate [90]. However, for the purpose of this review the term multivariate analysis will be used as most of the methods mention in this review have been referred to a multivariate analysis. Multivariate analysis of microbiome data can be performed in two ways: i) the distance-based approach that uses distance/dissimilarity matrices (beta diversity measures) such as the Bray-Curtis measure, or ii) the canonical approach that uses raw data i.e. OTU table [19, 91, 92]. The distance-based approach is discussed in more detail below. The canonical approach uses the OTU table and requires that some assumptions be made on the relationship between the groups (linear, unimodal, etc.), i.e. how the data will be distributed [19, 93]. Choosing the appropriate approach (and tests) for multivariate analysis can be complicated for researchers who do not have a thorough understanding of statistical analytical methods and as such the risk of making the incorrect conclusions is higher [19]. To help researchers understand multivariate analysis and to choose the right tools, PL Buttigieg and A Ramette (2014) developed an interactive website called GUSTA ME (https://sites.google.com/site/mb3gustame/home), that acts as a resource tool for microbial ecologist and other researchers studying the microbiome [19]. Table 3.4 summarises the various distance-based and canonical multivariate tests that are available.

3.6.1 Distance-based approaches

In the distance-based approach, the first step is to ensure that all the data is in the same scale and format [93, 94]. This is achieved by standardising and normalising the data [93, 94]. The second step is to choose a distance measure to be used, e.g. Bray-Curtis [93, 94]. The third step

is to visualise the similarity and dissimilarity between objects using cluster analysis or ordination [93, 94]. Patterns in a dataset may be observed using either cluster analysis or ordination [93, 94]. The more similar the samples are, the closer the samples will cluster [95].

3.6.1.1 Clustering methods

There are two types of multivariate clustering: hierarchical and *k*-means clustering (user-defined clustering; the user decides how many groups the data should be clustered into) [19, 93, 94]. Hierarchical clustering is more appropriate for small datasets whereas *k*-means clustering is the most suitable tool for large datasets [19, 96]. There are several different hierarchical clustering methods, including i) single-linkage clustering (also known as nearest neighbour clustering) e.g. minimum spanning tee (MST), ii) complete-linkage clustering e.g. and iii) average-linkage e.g. unweighted pair-group method with arithmetic mean (UPGMA) clustering [19, 91]. The user-defined method, *k*-means clustering uses an algorithm which requires three parameters from the user: i) the number of clusters, which is defined as *k*, ii) cluster initialisation (choosing initial clusters) and iii) a distance matrix [19, 93, 97-99].

3.6.1.2 Ordination

The term ordination can be defined as "the arrangement of units in some order" [91]. In ecology, ordination is used to visualise objects on reference axes [91, 93]. Ideally, each descriptor in the study should be plotted as an axis; however, if there are more than three descriptors, it is not possible to visualise on paper [91]. As a result, the axes are chosen based on descriptors that the researchers are interested in [91]. As the graph(s) represent the variability in a reduced space (dimensionally), these methods are referred to as ordination in reduced space [91]. An example of an ordination method is principal coordinate analysis (PCoA) [19, 91, 93, 100]. Clustering can be combined with ordination in a method called non-metric dimensional scaling (NMDS) [19, 91, 93, 100].

3.6.1.3 Test for statistical significance

The last step in the distance approach (for multivariate analysis) is to test for the significant differences between the groups [93, 94]. Several test statistics can be used including analysis of similarities (ANOSIM), the Mantel test and permutational multivariate analysis of variance (PERMANOVA) [93, 100]. The most popular test statistics is the PERMANOVA method, in part due to the fact it can be used in studies which have a small sample size [101]. Each of these methods tests a different null hypothesis [102],

3.7 Differential abundance analysis of microbiome data

An alternative analysis approach is to compare the abundance of the microorganisms across the different groups studied e.g. control vs treatment [20]. However, determining the differences between the communities is difficult as microbial data is inherently compositional. The relative abundance of the microbiome is considered compositional as it sums to one [61]. There are, however, several challenges with the analysis of compositional data: i) analysis does not work with data that contains zero counts (as mentioned previously); however, microbiome data usually have zeros (often presence/absence of an OTU) [103]. In addition to the zero counts, microbial data are often overdispersed [104]. To overcome these challenges, negative binomial and zero-inflated models have been used [104]. The zero-inflated models include zero-inflated poisson (ZIP), zero-inflated gaussian (ZIG) and zero-inflated negative binomial (ZINB). Other methods that have been used in differential abundance analysis include machine learning e.g. random forest regression, log-ratio transformation [additive (alr), centered (clr) and isometric(ilr)], generalised linear model (GLM) [20, 61, 105]. The R software has several packages (tools) which can be used to analyse differential abundances which are listed in Table 3.5.

3.8 Conclusions

In this rapidly expanding field of microbiome research, large amounts of data are generated. The study of this data forms part of a field of study referred to as microbial ecology. One of the main components of microbial ecology is studying the diversity of microbial communities. There are two ways in which this diversity can be measured and analysed: (i) the diversity within the communities (alpha diversity) and (ii) the diversity between communities (beta diversity), including determination of factors that explain differences between populations using multivariate analysis. There is a large variety of statistical measures available to analyse the microbiome and in this review, guidance has been provided on how, where and when to use these appropriately. The number of measures that can be used to study diversity is continuously increasing and is compounding the difficulty in choosing the appropriate statistical measure. The most important factor is the research question of the study; followed by sample size and the environment, such as the human lung being studied. While this review aims to provide a guide on the analysis of microbiome data, guidelines and consensus for microbiome studies from sample collection to statistical analysis are still needed. The way forward is for microbiome analysis to be standardised, with clear guidelines. However, this is extremely difficult as each study may have different considerations that need to be taken into account and

different research questions to be answered. The authors recommend the use of Hill numbers for alpha diversity analysis (however, this methodology has a steep learning curve) as these effective numbers can be compared across different measures and studies and for beta diversity analysis, the authors recommend the use of multivariate analysis with either a phylogenetic or non-phylogenetic beta diversity measure.

LIST OF ABBREVIATIONS

ANOSIM:	Analysis of group similarities
CA:	Correspondence analysis
CCA:	Canonical correspondence analysis
CCorA:	Canonical correlation analysis
DCA:	Detrended correspondence analysis
DFA/LDA:	Discriminatory function analysis
HCA:	Hierarchical clustering
MST:	Minimum spanning tree
N/A:	Not available
NGS:	Next generation sequencing
NMDS:	Nonmetric multidimensional scaling
OPLS-DA:	Orthogonal projections to latent structure discriminant analysis
PA:	Procrustes analysis
PCA:	Principal component analysis
PCoA:	Principal coordinate analysis
PERMANOVA:	Multivariate analysis of variance with permutation
RDA:	Redundancy analysis
UPGMA:	Unweighted pair-group method with arithmetic mean

Acknowledgements

The authors would like to thank the National Research Foundation (NRF) for providing a bursary; helping make this research possible and the National Health Laboratory Service (NHLS) for providing the research grant.

Declarations

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and materials: Not applicable

Competing interest: The authors declare that they have no competing interest

Funding: National Health Laboratory Service of South Africa (NHLS) Research Trust (Grant number: GRANT004 94626)

Author contributions: TGM wrote the manuscript; RPH and MME were involved in manuscript design and outline and edited the manuscript; MMK edited the manuscript

Biographical notes:

Tanweer Goolam Mahomed

Miss Goolam Mahomed is currently a PhD student in Medical Microbiology at the Department of Medical Microbiology, University of Pretoria. Her primary research focus has been on studying bacteria in respiratory diseases, focusing on bacterial strain typing, whole-genome sequencing, microbiome analysis and bioinformatics.

Remco PH Peters

Prof R.P.H. Peters is an expert clinician, epidemiologist and researcher that co-supervised Miss Goolam Mahomed's PhD. He works as Head of Clinical Research for the Foundation for Professional Development in East London, South Africa, and is an affiliated extraordinary professor at the Department of Medical Microbiology at the University of Pretoria and the Maastricht University Medical Centre. His research interests are focused on public health microbiology and infectious disease epidemiology, especially in the field of HIV, tuberculosis and sexually transmitted infections.

Marleen M Kock

Prof MM Kock is an associate professor at the Department of Medical Microbiology, University of Pretoria and collaborator on Miss Goolam Mahomed's PhD. Her research interests include sexually transmitted pathogens and antibiotic-resistant bacteria (focusing on Gram-negative pathogens). Her expertise is in the field of molecular microbiology.

Marthie M Ehlers

Prof MM Ehlers is a full professor at the Department of Medical Microbiology, University of Pretoria, South Africa, that supervised Miss Goolam Mahomed's PhD. Her primary research interests include antibiotic resistance and molecular characterisation of clinically important Gram-positive and Gram-negative bacteria (with a One Health focus) and the microbiome of respiratory diseases. Throughout her years as a researcher, Prof MM Ehlers has gained expertise in a variety of microbiology disciplines including virology and epidemiology.

References

1. Bar-On YM, Phillips R, Milo R: **The biomass distribution on Earth**. *Proc Natl Acad Sci U S A* 2018, **115**(25):6506-6511.

2. Dash HR, Das S: **Molecular methods for studying microorganisms from atypical environments**. In: *Microbiology of atypical environments*. 2018: 89-122.

3. Human Microbiome Project Consortium: **A framework for human microbiome research**. *Nature* 2012, **486**(7402):215-221.

4. Martin R, Miquel S, Langella P, Bermudez-Humaran LG: **The role of metagenomics in understanding the human microbiome in health and disease**. *Virulence* 2014, **5**(3):413-423.

5. Marchesi JR, Ravel J: **The vocabulary of microbiome research: a proposal**. *Microbiome* 2015, **3**:31.

6. Mammen MJ, Sethi S: **COPD and the microbiome**. *Respirology* 2016, **21**(4):590-599.

7. Foster KR, Schluter J, Coyte KZ, Rakoff-Nahoum S: **The evolution of the host microbiome as an ecosystem on a leash**. *Nature* 2017, **548**(7665):43-51.

8. Amon P, Sanderson I: **What is the microbiome?** *Arch Dis Child Educ Pract Ed* 2017, **102**(5):257-260.

9. Layeghifard M, Hwang DM, Guttman DS: **Disentangling interactions in the microbiome: A network perspective**. *Trends Microbiol* 2017, **25**(3):217-228.

10. Dickson RP, Erb-Downward JR, Freeman CM, McCloskey L, Beck JM, Huffnagle GB, Curtis JL: **Spatial variation in the healthy human lung microbiome and the adapted island model of lung biogeography**. *Ann Am Thorac Soc* 2015, **12**(6):821-830.

11. Lopes SP, Azevedo NF, Pereira MO: **Microbiome in cystic fibrosis: Shaping polymicrobial interactions for advances in antibiotic therapy**. *Crit Rev Microbiol* 2015, **41**(3):353-365.

12. Taylor SL, Wesselingh S, Rogers GB: **Host-microbiome interactions in acute and chronic respiratory infections**. *Cell Microbiol* 2016, **18**(5):652-662.

13. O'Dwyer DN, Dickson RP, Moore BB: **The lung microbiome, immunity, and the pathogenesis of chronic lung disease**. *J Immunol* 2016, **196**(12):4839-4847.

14. Madigan MT, Bender KS, Buckley DH, Sattley WM, Stahl DA: **Brock biology of microorganisms, global edition**. In., 15th eds. Harlow, United Kingdom: Pearson Education Limited; 2017.

15. Malla MA, Dubey A, Kumar A, Yadav S, Hashem A, Abd_Allah EF: **Exploring the human microbiome: The potential future role of next-generation sequencing in disease diagnosis and treatment**. *Front Immunol* 2019, **9**:2868.

16. Gupta S, Mortensen MS, Schjørring S, Trivedi U, Vestergaard G, Stokholm J, Bisgaard H, Krogfelt KA, Sørensen SJ: **Amplicon sequencing provides more accurate microbiome information in healthy children compared to culturing**. *Commun Biol* 2019, **2**(1).

17. Cao Y, Fanning S, Proos S, Jordan K, Srikumar S: **A review on the applications of next generation sequencing technologies as applied to food-related microbiome studies**. *Front Microbiol* 2017, **8**:1829.

18. Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, Lawley TD, Finn RD: **A new genomic blueprint of the human gut microbiota**. *Nature* 2019, **568**(7753):499-504.

19. Buttigieg PL, Ramette A: **A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses**. *FEMS Microbiol Ecol* 2014, **90**(3):543-550.

20. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciolek T, McCall LI, McDonald D *et al*: **Best practices for analysing microbiomes**. *Nat Rev Microbiol* 2018, **16**(7):410-422.

21. Poussin C, Sierro N, Boue S, Battey J, Scotti E, Belcastro V, Peitsch MC, Ivanov NV, Hoeng J: **Interrogating the microbiome: experimental and computational**

considerations in support of study reproducibility. *Drug Discov Today* 2018, **23**(9):1644-1657.

22. Goodrich JK, Di Rienzi SC, Poole AC, Koren O, Walters WA, Caporaso JG, Knight R, Ley RE: **Conducting a microbiome study**. *Cell* 2014, **158**(2):250-262.

23. Panek M, Cipcic Paljetak H, Baresic A, Peric M, Matijasic M, Lojkic I, Vranesic Bender D, Krznaric Z, Verbanac D: **Methodology challenges in studying human gut microbiota - effects of collection, storage, DNA extraction and next generation sequencing technologies**. *Sci Rep* 2018, **8**(1):5143.

24. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F: **Laboratory procedures to generate viral metagenomes**. *Nat Protoc* 2009, **4**(4):470-483.

25. Wommack KE, Bhavsar J, Polson SW, Chen J, Dumas M, Srinivasiah S, Furman M, Jamindar S, Nasko DJ: **VIROME: a standard operating procedure for analysis of viral metagenome sequences**. *Stand Genomic Sci* 2012, **6**(3):427-439.

26. Tremblay J, Singh K, Fern A, Kirton ES, He S, Woyke T, Lee J, Chen F, Dangl JL, Tringe SG: **Primer and platform effects on 16S rRNA tag sequencing**. *Front Microbiol* 2015, **6**:771.

27. Mekuto L, Ntwampe SKO, Mudumbi JBN, Akinpelu EA, Mewa-Ngongang M: **Metagenomic data of free cyanide and thiocyanate degrading bacterial communities**. *Data Brief* 2017, **13**:738-741.

28. Kim M, Chun J: **16S rRNA gene-based identification of bacteria and archaea using the EzTaxon server**. In: *New Approaches to Prokaryotic Systematics.* 2014: 61-74.

29. Wang Z, Liu H, Wang F, Yang Y, Wang X, Chen B, Stampfli MR, Zhou H, Shu W, Brightling CE *et al*: **A refined view of airway microbiome in chronic obstructive pulmonary disease at species and strain-levels**. *Frontiers in Microbiology* 2020, **11**.

30. Brumfield KD, Huq A, Colwell RR, Olds JL, Leddy MB: **Microbial resolution of whole genome shotgun and 16S amplicon metagenomic sequencing using publicly available NEON data**. *PLoS One* 2020, **15**(2):e0228899.

31. Amato KR: **An introduction to microbiome analysis for human biology applications**. *Am J Hum Biol* 2017, **29**(1).

32. Forbes JD, Knox NC, Peterson CL, Reimer AR: **Highlighting clinical metagenomics for enhanced diagnostic decision-making: A step towards wider implementation**. *Comput Struct Biotechnol J* 2018, **16**:108-120.

33. Illumina: **16S Metagenomic sequencing library preparation**. In.; 2013.

34. Hodkinson BP, Grice EA: **Next-generation sequencing: A review of technologies and tools for wound microbiome research**. *Adv Wound Care (New Rochelle)* 2015, **4**(1):50-58.

35. Kumar R, Eipers P, Little RB, Crowley M, Crossman DK, Lefkowitz EJ, Morrow CD: **Getting started with microbiome analysis: sample acquisition to bioinformatics**. *Curr Protoc Hum Genet* 2014, **82**:18.18.11–18.18.29.

36. Escalona M, Rocha S, Posada D: **A comparison of tools for the simulation of genomic next-generation sequencing data**. *Nat Rev Genet* 2016, **17**(8):459-469.

37. Calle ML: **Statistical analysis of metagenomics data**. *Genomics Inform* 2019, **17**(1):e6.

38. Mohsen A, Park J, Chen YA, Kawashima H, Mizuguchi K: **Impact of quality trimming on the efficiency of reads joining and diversity analysis of Illumina paired-end reads in the context of QIIME1 and QIIME2 microbiome analysis frameworks**. *BMC Bioinformatics* 2019, **20**(1):581.

39. Lopez-Garcia A, Pineda-Quiroga C, Atxaerandio R, Perez A, Hernandez I, Garcia-Rodriguez A, Gonzalez-Recio O: **Comparison of Mothur and QIIME for the analysis of rumen microbiota composition based on 16S rRNA amplicon sequences**. *Front Microbiol* 2018, **9**:3010.

40. Chappidi S, Villa EC, Cantarel BL: **Using Mothur to determine bacterial community composition and structure in 16S ribosomal RNA datasets**. *Curr Protoc Bioinformatics* 2019, **67**(1):e83.

41. Di Bella JM, Bao Y, Gloor GB, Burton JP, Reid G: **High throughput sequencing methods and analysis for microbiome research**. *J Microbiol Methods* 2013, **95**(3):401-414.

42. Auer L, Mariadassou M, O'Donohue M, Klopp C, Hernandez-Raquet G: **Analysis of large 16S rRNA Illumina data sets: Impact of singleton read filtering on microbial community description**. *Mol Ecol Resour* 2017, **17**(6):e122-e132.

43. Valverde JR, Mellado RP: **Analysis of metagenomic data containing high biodiversity levels**. *PLoS One* 2013, **8**(3):e58118.

44. Kaszubinski SF, Pechal JL, Schmidt CJ, Jordan HR, Benbow ME, Meek MH: **Evaluating bioinformatic pipeline performance for forensic microbiome analysis**. *J Forensic Sci* 2020, **65**(2):513-525.

45. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F *et al*: **Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2**. *Nat Biotechnol* 2019, **37**(8):852-857.

46. Navas-Molina JA, Peralta-Sanchez JM, Gonzalez A, McMurdie PJ, Vazquez-Baeza Y, Xu Z, Ursell LK, Lauber C, Zhou H, Song SJ *et al*: **Advancing our understanding of the human microbiome using QIIME**. *Methods Enzymol* 2013, **531**:371-444.

47. Kuczynski J, Stombaugh J, Walters WA, Gonzalez A, Caporaso JG, Knight R: **Using QIIME to analyze 16S rRNA gene sequences from microbial communities**. *Curr Protoc Bioinformatics* 2011, **36**:10.17.11-10.17.20.

48. McDonald D, Clemente JC, Kuczynski J, Rideout JR, Stombaugh J, Wendel D, Wilke A, Huse S, Hufnagle J, Meyer F *et al*: **The biological observation matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome**. *Gigascience* 2012, **1**(1):7.

49. Wylezich C, Papa A, Beer M, Hoper D: **A versatile sample processing workflow for metagenomic pathogen detection**. *Sci Rep* 2018, **8**(1):13108.

50. Mulcahy-O'Grady H, Workentine ML: **The challenge and potential of metagenomics in the clinic**. *Front Immunol* 2016, **7**:29.

51. Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D, Knight R: **Experimental and analytical tools for studying the human microbiome**. *Nat Rev Genet* 2011, **13**(1):47-58.

52. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N: **Shotgun metagenomics, from sampling to analysis**. *Nat Biotechnol* 2017, **35**(9):833-844.

53. Ju F, Zhang T: **Experimental design and bioinformatics analysis for the application of metagenomics in environmental sciences and biotechnology**. *Environ Sci Technol* 2015, **49**(21):12628-12640.

54. Forbes JD, Knox NC, Ronholm J, Pagotto F, Reimer A: **Metagenomics: The next culture-independent game changer**. *Front Microbiol* 2017, **8**:1069.

55. Reinert K, Langmead B, Weese D, Evers DJ: **Alignment of next-generation sequencing reads**. *Annu Rev Genomics Hum Genet* 2015, **16**:133-151.

56. Ghurye JS, Cepeda-Espinoza V, Pop M: **Metagenomic assembly: Overview, challenges and applications**. *Yale J Biol Med* 2016, **89**(3):353-362.

57. Breitwieser FP, Lu J, Salzberg SL: **A review of methods and databases for metagenomic classification and assembly**. *Brief Bioinform* 2019, **20**(4):1125-1136.

58. Xing X, Liu JS, Zhong W: **MetaGen: reference-free learning with multiple metagenomic samples**. *Genome Biol* 2017, **18**(1):187.

59. Plummer E, Twin J: **A comparison of three bioinformatics pipelines for the analysis of preterm gut microbiota using 16S rRNA gene sequencing data**. *J Proteomics Bioinform* 2015, **8**(12):283-291.

60. Niu SY, Yang J, McDermaid A, Zhao J, Kang Y, Ma Q: **Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes**. *Brief Bioinform* 2018, **19**(6):1415-1429.

61. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vazquez-Baeza Y, Birmingham A *et al*: **Normalization and microbial differential abundance strategies depend upon data characteristics**. *Microbiome* 2017, **5**(1):27.

62. Willis AD: **Rarefaction, alpha diversity, and statistics**. *Front Microbiol* 2019, **10**:2407.

63. McMurdie PJ, Holmes S: **Waste not, want not: why rarefying microbiome data is inadmissible**. *PLoS Comput Biol* 2014, **10**(4):e1003531.

64. Lozupone CA, Knight R: **Species divergence and the measurement of microbial diversity**. *FEMS Microbiol Rev* 2008, **32**(4):557-578.

65. Ju F, Zhang T: **16S rRNA gene high-throughput sequencing data mining of microbial diversity and interactions**. *Appl Microbiol Biotechnol* 2015, **99**(10):4119-4129.

66. Morgan XC, Huttenhower C: **Chapter 12: Human microbiome analysis**. *PLoS Comput Biol* 2012, **8**(12):e1002808.

67. Finotello F, Mastrorilli E, Di Camillo B: **Measuring the diversity of the human microbiota with targeted next-generation sequencing**. *Brief Bioinform* 2018, **19**(4):679-692.

68. Stubbendieck RM, Vargas-Bautista C, Straight PD: **Bacterial communities: interactions to scale**. *Front Microbiol* 2016, **7**:1234.

69. Hamady M, Knight R: **Microbial community profiling for human microbiome projects: Tools, techniques, and challenges**. *Genome Res* 2009, **19**(7):1141-1152.

70. Wagner BD, Grunwald GK, Zerbe GO, Mikulich-Gilbertson SK, Robertson CE, Zemanick ET, Harris JK: **On the use of diversity measures in longitudinal sequencing studies of microbial communities**. *Front Microbiol* 2018, **9**:1037.

71. Podani J, Ricotta C, Schmera D: **A general framework for analyzing beta diversity, nestedness and related community-level phenomena based on abundance data**. *Ecol Complex* 2013, **15**:52-61.

72. Jovel J, Patterson J, Wang W, Hotte N, O'Keefe S, Mitchel T, Perry T, Kao D, Mason AL, Madsen KL *et al*: **Characterization of the gut microbiome using 16S or shotgun metagenomics**. *Front Microbiol* 2016, **7**:459.

73. Chao A, Chiu C-H, Jost L: **Phylogenetic diversity measures and their decomposition: A framework based on Hill numbers**. In: *Biodiversity conservation and phylogenetic systematics: preserving our evolutionary heritage in an extinction crisis.* Edited by Pellens R, Grandcolas P. Cham: Springer International Publishing; 2016: 141-172.

74. Zaura E: **Next-generation sequencing approaches to understanding the oral microbiome**. *Adv Dent Res* 2012, **24**(2):81-85.

75. Morris EK, Caruso T, Buscot F, Fischer M, Hancock C, Maier TS, Meiners T, Muller C, Obermaier E, Prati D *et al*: **Choosing and using diversity indices: insights for ecological applications from the German biodiversity exploratories**. *Ecol Evol* 2014, **4**(18):3514-3524.

76. Daly A, Baetens J, De Baets B: **Ecological diversity: Measuring the unmeasurable**. *Mathematics* 2018, **6**(7).

77. Kim BR, Shin J, Guevarra R, Lee JH, Kim DW, Seol KH, Lee JH, Kim HB, Isaacson R: **Deciphering diversity indices for a better understanding of microbial communities**. *J Microbiol Biotechnol* 2017, **27**(12):2089-2093.

78. Johnson KV, Burnet PW: **Microbiome: Should we diversify from diversity?** *Gut Microbes* 2016, **7**(6):455-458.

79. Alberdi A, Gilbert MTP: **A guide to the application of Hill numbers to DNA-based diversity analyses**. *Mol Ecol Resour* 2019, **19**(4):804-817.

80. Jost L: **Partitioning diversity into independent alpha and beta components**. *Ecology* 2007, **88**(10):2427-2439.

81. Chao A, Chiu C-H, Jost L: **Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through**

Hill numbers. *Annual Review of Ecology, Evolution, and Systematics* 2014, **45**(1):297-324.

82. Hill M: **Diversity and evenness: A unifying notation and its consequences**. *Ecology Letters* 1973, **54**(2):427-432.

83. Ma ZS, Li L: **Measuring metagenome diversity and similarity with Hill numbers**. *Mol Ecol Resour* 2018, **18**(6):1339-1355.

84. Hsieh TC, Chao A: **Rarefaction and extrapolation: Making fair comparison of abundance-sensitive phylogenetic diversity among multiple assemblages**. *Syst Biol* 2017, **66**(1):100-111.

85. Zhao N, Chen J, Carroll IM, Ringel-Kulka T, Epstein MP, Zhou H, Zhou JJ, Ringel Y, Li H, Wu MC: **Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test**. *Am J Hum Genet* 2015, **96**(5):797-807.

86. Tuomisto H, Ruokolainen K: **Analyzing or explaining beta diversity? Understanding the targets of different methods of analysis**. *Ecology* 2006, **87**(11):2697-2708.

87. Legendre P: **Studying beta diversity: ecological variation partitioning by multiple regression and canonical analysis**. *J Plant Ecol* 2007, **1**(1):3-8.

88. Legendre P, Borcard D, Peres-Neto PR: **Analyzing beta diversity: Partitioning the spatial variation of community composition data**. *Ecol Monogr* 2005, **75**(4):435-450.

89. Hidalgo B, Goodman M: **Multivariate or multivariable regression?** *Am J Public Health* 2013, **103**(1):39-40.

90. Ebrahimi Kalan M, Jebai R, Zarafshan E, Bursac Z: **Distinction Between Two Statistical Terms: Multivariable and Multivariate Logistic Regression**. *Nicotine Tob Res* 2020.

91. Legendre P, Legendre LF: **Numerical ecology**, vol. 24, Third eds. Oxford, UK: Elsevier; 2012.

92. **GUSTA ME** [https://mb3is.megx.net/gustame/home]

93. Ramette A: **Multivariate analyses in microbial ecology**. *FEMS Microbiol Ecol* 2007, **62**(2):142-160.

94. Anderson MJ: **A new method for non-parametric multivariate analysis of variance**. *Austral Ecology* 2001, **26**(1):32-46.

95. Frades I, Matthiesen R: **Overview on techniques in cluster analysis**. In: *Bioinformatics methods in clinical research.* Humana Press; 2010: 81-107.

96. Rodriguez MZ, Comin CH, Casanova D, Bruno OM, Amancio DR, Costa LDF, Rodrigues FA: **Clustering algorithms: A comparative approach**. *PLoS One* 2019, **14**(1):e0210236.

97. Jain AK: **Data clustering: 50 years beyond *k*-means**. *Pattern Recognit Lett* 2010, **31**(8):651-666.

98. Khan SS, Ahmad A: **Cluster center initialization algorithm for *k*-means clustering**. *Pattern Recognit Lett* 2004, **25**(11):1293-1302.

99. Bai L, Liang J, Dang C, Cao F: **A cluster centers initialization method for clustering categorical data**. *Expert Syst Appl* 2012, **39**(9):8022-8029.

100. Paliy O, Shankar V: **Application of multivariate statistical techniques in microbial ecology**. *Mol Ecol* 2016, **25**(5):1032-1057.

101. Tang ZZ, Chen G, Alekseyenko AV: **PERMANOVA-S: association test for microbial community composition that accommodates confounders and multiple distances**. *Bioinformatics* 2016, **32**(17):2618-2625.

102. Anderson MJ, Walsh DCI: **PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing?** *Ecol Monogr* 2013, **83**(4):557-574.

103. Quinn TP, Erb I, Richardson MF, Crowley TM: **Understanding sequencing data as compositions: an outlook and review**. *Bioinformatics* 2018, **34**(16):2870-2878.

104. Xia Y, Sun J: **Hypothesis testing and statistical analysis of microbiome**. *Genes Dis* 2017, **4**(3):138-148.

105. Tsilimigras MC, Fodor AA: **Compositional data analysis of the microbiome: fundamentals, tools, and challenges**. *Ann Epidemiol* 2016, **26**(5):330-335.

106. McGill BJ, Etienne RS, Gray JS, Alonso D, Anderson MJ, Benecha HK, Dornelas M, Enquist BJ, Green JL, He F *et al*: **Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework**. *Ecol Lett* 2007, **10**(10):995-1015.

107. Travlos IS, Cheimona N, Roussis I, Bilalis DJ: **Weed-species abundance and diversity indices in relation to tillage systems and fertilization**. *Front Environ Sci* 2018, **6**.

108. Tyler AD, Smith MI, Silverberg MS: **Analyzing the human microbiome: a "how to" guide for physicians**. *Am J Gastroenterol* 2014, **109**(7):983-993.

109. Legendre P, Legendre L: **Canonical analysis**. In: *Numerical ecology.* 2012: 625-710.

110. Legendre P, Legendre L: **Cluster analysis**. In: *Numerical ecology.* 2012: 337-424.

111. Wright ES, Vetsigian KH: **Quality filtering of Illumina index reads mitigates sample cross-talk**. *BMC Genomics* 2016, **17**(1):876.

112. Girardot C, Scholtalbers J, Sauer S, Su SY, Furlong EE: **Je, a versatile suite to handle multiplexed NGS libraries with unique molecular identifiers**. *BMC Bioinformatics* 2016, **17**(1):419.

113. Kadariya J, Smith TC, Thapaliya D: ***Staphylococcus aureus* and staphylococcal food-borne disease: an ongoing challenge in public health**. *Biomed Res Int* 2014, **2014**:827965.

114. Brown MT, Wicker LR: **Discriminant analysis**. In: *Handbook of applied multivariate statistics and mathematical modeling.* 2000: 209-235.

115. Swingland IR: **Biodiversity, definition of**. In: *Encyclopedia of Biodiversity.* 2001: 377-391.

116. Battisti C, Bazzichetto M, Poeta G, Pietrelli L, Acosta ATR: **Measuring non-biological diversity using commonly used metrics: Strengths, weaknesses and caveats for their application in beach litter management**. *J Coast Conserv* 2017, **21**(2):303-310.

117. Magurran AE, McGill BJ: **Biological diversity: frontiers in measurement and assessment**. Oxford, United Kingdom: OUP Oxford; 2010.

118. Kitsios GD, Morowitz MJ, Dickson RP, Huffnagle GB, McVerry BJ, Morris A: **Dysbiosis in the intensive care unit: microbiome science coming to the bedside**. *J Crit Care* 2017, **38**:84-91.

119. Cui L, Morris A, Huang L, Beck JM, Twigg HL, 3rd, von Mutius E, Ghedin E: **The microbiome and the lung**. *Ann Am Thorac Soc* 2014, **11 Suppl 4**:S227-232.

120. Gaspar JM: **NGmerge: merging paired-end reads via novel empirically-derived models of sequencing errors**. *BMC Bioinformatics* 2018, **19**(1):536.

121. Shankar V, Agans R, Paliy O: **Advantages of phylogenetic distance based constrained ordination analyses for the examination of microbial communities**. *Sci Rep* 2017, **7**(1):6481.

122. Kulski JK: **Next-generation sequencing — An overview of the history, tools, and "omic" applications**. In: *Next generation sequencing - advances, applications and challenges.* 2016.

123. McCombie WR, McPherson JD, Mardis ER: **Next-generation sequencing technologies**. *Cold Spring Harb Perspect Med* 2019, **9**(11).

124. Legendre P, Legendre L: **Ordination in reduced space**. In: *Numerical ecology*. 2012: 425-520.

125. Wildi O: **Evaluating the predictive power of ordination methods in ecological context**. *Mathematics* 2018, **6**(12).

126. Sokal RR, Sneath PHA: **Principles of numerical taxonomy**. San Francisco: W.H. Freeman; 1963.

127. Schloss PD, Westcott SL: **Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis**. *Appl Environ Microbiol* 2011, **77**(10):3219-3226.

128. Porter TM, Hajibabaei M: **Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis**. *Mol Ecol* 2018, **27**(2):313-338.

129. Zhang J, Kobert K, Flouri T, Stamatakis A: **PEAR: a fast and accurate Illumina Paired-End reAd mergeR**. *Bioinformatics* 2014, **30**(5):614-620.

130. Madigan M, Bender K, Buckley D, Sattley W, Stahl D: **Microbial evolution and systematics**. In: *Brock biology of microorganisms.* Global Edition eds: Pearson Education Limited; 2017.

131. Strait DS: **Phylogeny**: John Wiley & Sons; 2018.

132. Washburne AD, Morton JT, Sanders J, McDonald D, Zhu Q, Oliverio AM, Knight R: **Methods for phylogenetic analysis of microbiome data**. *Nat Microbiol* 2018, **3**(6):652-661.

133. **Encyclopedia of research design**: SAGE Publications; 2010.

134. Legendre P, Legendre L: **Complex ecological data sets**. In: *Numerical ecology.* 2012: 1-57.

135. den Besten HMW, Amezquita A, Bover-Cid S, Dagnas S, Ellouze M, Guillou S, Nychas G, O'Mahony C, Perez-Rodriguez F, Membre JM: **Next generation of microbiological risk assessment: Potential of omics data for exposure assessment**. *Int J Food Microbiol* 2018, **287**:18-27.

136. **Random variation**. In: *Encyclopedia of public health.* Edited by Kirch W. Dordrecht: Springer Netherlands; 2008: 1232-1232.

137. Yegnasubramanian S: **Explanatory chapter: next generation sequencing**. *Methods Enzymol* 2013, **529**:201-208.

138. Rogers GB, Shaw D, Marsh RL, Carroll MP, Serisier DJ, Bruce KD: **Respiratory microbiota: addressing clinical questions, informing clinical practice**. *Thorax* 2015, **70**(1):74-81.

139. Chun J, Oren A, Ventosa A, Christensen H, Arahal DR, da Costa MS, Rooney AP, Yi H, Xu XW, De Meyer S *et al*: **Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes**. *Int J Syst Evol Microbiol* 2018, **68**(1):461-466.

140. Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ: **Counting the uncountable: statistical approaches to estimating microbial diversity**. *Appl Environ Microbiol* 2001, **67**(10):4399-4406.

141. Ashton JJ, Beattie RM, Ennis S, Cleary DW: **Analysis and interpretation of the human microbiome**. *Inflamm Bowel Dis* 2016, **22**(7):1713-1722.

142. Chao A: **Nonparametric estimation of the number of classes in a population**. *Scand J Stat* 1984, **11**(4):265-270.

143. Lemos LN, Fulthorpe RR, Triplett EW, Roesch LF: **Rethinking microbial diversity analysis in the high throughput sequencing era**. *J Microbiol Methods* 2011, **86**(1):42-51.

144. Magurran AE: **Measuring biological diversity**. Hoboken, United Kingdom: John Wiley & Sons, Incorporated; 2013.

145. Chazdon RL, Colwell RK, Denslow JS, Guariguata MR: **Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of Northeastern Costa Rica**. In: *Forest biodiversity research, monitoring and modeling: conceptual background and old world case studies.* 1998: 285-309.

146. Faith DP: **Conservation evaluation and phylogenetic diversity**. *Biol Conserv* 1992, **61**(1):1-10.

147. Lean C, Maclaurin J: **The value of phylogenetic diversity**. In: *Biodiversity conservation and phylogenetic systematics: Preserving our evolutionary heritage in an extinction crisis.* Edited by Pellens R, Grandcolas P. Cham: Springer International Publishing; 2016: 19-37.

148. Allen B, Kon M, Bar-Yam Y: **A new phylogenetic diversity measure generalizing the Shannon index and its application to phyllostomid bats**. *Am Nat* 2009, **174**(2):236-243.

149. Shannon CE: **A mathematical theory of communication**. *The Bell System Technical Journal* 1984, **27**(3):379-423.

150. Simpson EH: **Measurement of diversity**. *Nature* 1949, **163**:688.

151. Martin AP: **Phylogenetic approaches for describing and comparing the diversity of microbial communities**. *Appl Environ Microbiol* 2002, **68**(8):3673-3682.

152. Palmer MW: **The estimation of species richness by extrapolation**. *Ecology* 1990, **71**(3):1195-1198.

153. Heltshe JF, Forrester NE: **Estimating species richness using the Jackknife procedure**. *Biometrics* 1983, **39**(1):1-11.

154. Sørensen T: **A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons**. *Biol Skr* 1948, **5**:1-34.

155. Koleff P, Gaston KJ, Lennon JJ: **Measuring beta diversity for presence–absence data**. *J Anim Ecol* 2003, **72**(3):367-382.

156. Li K, Bihan M, Methe BA: **Analyses of the stability and core taxonomic memberships of the human microbiome**. *PLoS One* 2013, **8**(5):e63139.

157. Chao A, Chazdon RL, Colwell RK, Shen TJ: **Abundance-based similarity indices and their estimation when there are unseen species in samples**. *Biometrics* 2006, **62**(2):361-371.

158. Jaccard P: **The distribution of the flora in the Alpine zone**. *New Phytol* 1912, **11**(2):37-50.

159. Chang Q, Luan Y, Sun F: **Variance adjusted weighted UniFrac: a powerful beta diversity measure for comparing communities based on phylogeny**. *BMC Bioinformatics* 2011, **12**:118.

160. Lozupone C, Knight R: **UniFrac: a new phylogenetic method for comparing microbial communities**. *Appl Environ Microbiol* 2005, **71**(12):8228-8235.

161. Lozupone CA, Hamady M, Kelley ST, Knight R: **Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities**. *Appl Environ Microbiol* 2007, **73**(5):1576-1585.

162. Schroeder PJ, Jenkins DG: **How robust are popular beta diversity indices to sampling error?** *Ecosphere* 2018, **9**(2):e02100.

163. Horn HS: **Measurement of "overlap" in comparative ecological studies**. *Am Nat* 1966, **100**(914):419-424.

164. Rempala GA, Seweryn M: **Methods for diversity and overlap analysis in T-cell receptor populations**. *J Math Biol* 2013, **67**(6-7):1339-1368.

165. Morisita M: **Measuring of interspecific association and similarity between communities**. *Mem Fac Sci Kyushu Univ Series E* 1959, **3**:65-80.

166. Evans SN, Matsen FA: **The phylogenetic Kantorovich-Rubinstein metric for environmental sequence samples**. *J R Stat Soc Series B Stat Methodol* 2012, **74**(3):569-592.

167. Wong RG, Wu JR, Gloor GB: **Expanding the UniFrac toolbox**. *PLoS One* 2016, **11**(9):e0161196.

168. **Overview of QIIME 2 plugin workflows** [https://docs.qiime2.org/2019.10/tutorials/overview/#fun]

169. Sedlar K, Videnska P, Skutkova H, Rychlik I, Provaznik I: **Bipartite graphs for visualization analysis of microbiome data**. *Evol Bioinform Online* 2016, **12** (Suppl 1):17-23.

170. Huse SM, Mark Welch DB, Voorhis A, Shipunova A, Morrison HG, Eren AM, Sogin ML: **VAMPS: a website for visualization and analysis of microbial population structures**. *BMC Bioinformatics* 2014, **15**:41.

171. Dhariwal A, Chong J, Habib S, King IL, Agellon LB, Xia J: **MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data**. *Nucleic Acids Res* 2017, **45**(W1): W180-W188.

172. Buza TM, Tonui T, Stomeo F, Tiambo C, Katani R, Schilling M, Lyimo B, Gwakisa P, Cattadori IM, Buza J *et al*: **iMAP: an integrated bioinformatics and visualization pipeline for microbiome data analysis**. *BMC Bioinformatics* 2019, **20**(1):374.

Table 3.1: Glossary of terms used in the analysis of microbiome

Term	Definition	References
Abundance	Number of observed clustered sequences (or OTUs)	[61, 106, 107]
Alpha diversity	The average diversity within an environment e.g. sputum specimen.	[37, 108]
Beta diversity	Comparison of the diversity between different environments/samples	[37, 108]
Biom	Biological observation matrix	[48]
Canonical analysis	Direct comparison between matrices in their simplest form e.g. OTU table	[109]
Clustering (also referred to as binning)	Grouping similar objects together by partitioning into subsets	[110]
Demultiplexed reads	Reads that were previously in the same run and already have their barcodes removed	[111, 112]
Discriminatory analysis	Analysis that discriminates data into groups i.e. classifies data. It measures predictor/discriminant variables against mutually exclusive groups (grouping variables)	[100, 113, 114]
Diversity	A range of differences (variability) within, among or between groups	[83, 115]
Diversity measures	Measures/indices used to quantify diversity	[116]
Evenness	A measure of the relative abundance of different taxonomic units (OTUs) in a community	[6, 117-119]
Exploratory analysis	Analysis that is used to find patterns in data. It measures an object (e.g. sample) against a variable (e.g. abundance of OTUs)	[93, 100]
Fastq file	Read sequences with a quality file	[120]
Interpretive analysis	Analysis that interprets relationships between data. It measures explanatory variables (independent variables), such as environmental factors, different sample groups or patient metadata against response variables (variables of interest) such as OTU table.	[93, 100, 121]
Microbiome	All microorganisms living in a habitat, such as the human lungs, their genetic material and the surrounding environmental conditions	[5, 118]
Multiplexed reads	Reads from multiple samples joined in a single run with each sample having a unique barcode	[111, 112]
Next generation sequencing	High throughput rapid parallel sequencing. Also known as high throughput sequencing (HTS)	[6, 122, 123]
Ordination	Arrangement of data points across a reduced number of axes (one or more), whilst keeping trends and preserving distances between objects (data points). Visualised as two- or three-dimension plots	[22, 37, 100, 124, 125]
OTU	Operational taxonomic unit. Group of similar DNA sequences (often at 97% similarity)	[118, 126-128]
Paired-end reads	Reads generated from DNA sequenced using forward and reverse primers i.e. at both ends. Have two outputs (read files)	[36, 129]
Phylogeny	The evolutionary history of the microorganism i.e. how the microorganism diversified over time	[130-132]
Qualitative	Non-ordered data that is observed and has mutually exclusive categories	[133-135]
Quantitative	Ordered data that is measured	[134]
Random variation	A variation which has no known explanation or root cause	[136]
Read	A string of sequences (base pairs) generated by the next generation sequencing instrument/platform	[137]
Richness	Number of unique OTUs in a community	[6, 117-119, 138]
Single-end read	Read generated from DNA only being sequenced from one end. Has one output (read file)	[36, 129]
Taxa (taxon)	The taxonomic rank of a microorganism (or any organism).	[128]
Taxonomy	Hierarchical classification and identification	[130, 132, 139]
Unimodal distribution	Data that has only one peak on the variable density plot	[111, 112]

Table 3.2: Summary of characteristics of alpha diversity measures that can be used in microbiome studies

Statistical tool	Taxon/ Phylogenetic	Equations	Advantages	Disadvantages	References
Qualitative					
Chao1	Taxon	$S_{Chao1} = S_{obs} + \frac{n_1^2}{2n_2}$ where S_{obs} is the number of observed species, n_1 is the number of singletons (single reads) and n_2 is the number of doubletons	Precise	All species are regarded as equally related. Requires abundance data (e.g. OTU table)	[64, 140-142] [117, 143, 144]
Abundance-base coverage (ACE)	Taxon	$S_{ACE} = S_{abund} + \frac{S_{rare}}{C_{ACE}} + \frac{F_1}{C_{ACE}} \gamma_{ACE}^2$ where S_{abund} is the number of abundant species, S_{rare} is the number of rare species, $C_{ACE}=1-F_1/N_{rare}$ (F_1 is the number of species with i individuals) and $N_{rare} = \sum_{i=1}^{10} iF_i$	Considers both rare and abundant species	All species are regarded as equally related. Only provides information on the species observed	[64, 140, 141, 143-145]
Phylogenetic Diversity (PD)	Phylogenetic	$PD = (N-1) + \text{no. of internal nodes of the minimum spanning path}$, where N is the size of the taxa	Provides both branch length and topographical information	Requires a phylogenetic tree; More weight is given to richness (over evenness); analysis is difficult with populations of different sample sizes	[64, 144, 146, 147]
Quantitative					
Shannon's Index	Taxon	$H = -\sum_i p_i \ln p_i$; where p_i is the number of individuals in species s_i	Confounds species richness and evenness; sensitive to rarer species	All species are regarded as equally related; Sensitive to sample size; Values have no absolute meaning	[62, 64, 76, 143, 148, 149]
Simpson's Index	Taxon	$D = -\sum_i p_i^2$; where p_i is the number of individuals in species s_i	Suitable for smaller sample sizes; robust	All species are regarded as equally related; Requires abundance data; not intuitive; Values have no absolute meaning; does not account for unobserved species	[62, 64, 76, 143, 144, 148, 150]
Theta (θ)	Phylogenetic	$\theta(\pi) = \sum_{i=1}^{k} \sum_{j<i} p_i p_j d_{ij}$ where k is the number of distinct sequences, p_i is the frequency of the first (*i*th) sequence, p_j is the frequency of the second sequence (*j*th)and d_{ij} is the number of (nucleotide) differences between the two sequences	Provides a phylogenetic measurement	Richness is not considered	[64, 151]
Jackknife	Unknown	$JACK1 = SO + \frac{r1(n-1)}{n}$; where SO is the number of species observed in *n* quadrants and *r*1 is the number of species present in one quadrant	Precise; useful in populations where there is resampling	Sensitive to sample size	[66, 144, 152, 153]

Table 3.3: Summary of characteristics of beta diversity measures that are used in microbiome studies

Statistical tool	Taxon/ Phylogenetic	Equations	Input	Output (results)	Interpretation of results	Pros and Cons	References
Qualitative							
Sorenson Index/ Dice's coefficient	Taxon	$\boldsymbol{\beta_{sor}} = \frac{2a}{\alpha_1+\alpha_2}$; where a is the total number of species that occur in both populations, α_1 is the total number of species in population 1 and α_2 is the total number of species in population 2	OTU table	A value between 0 and 1	The closer the number is to one, the more similar the samples are	Pros: Simple and intuitive Cons: All species are regarded as equally related	[64, 143, 154-157]
Jaccard	Taxon	$\boldsymbol{\beta_j} = \frac{a}{\alpha_1+\alpha_2-a}$; where a is the total number of species that occur in both populations, α_1 is the total number of species in population 1 and α_2 is the total number of species in population 2	OTU table	A value between 0 and 1	The closer the number is to one, the more similar the samples are	Pros: Simple and intuitive Cons: All species are regarded as equally related	[64, 143, 155, 157, 158]
Unweighted UniFrac	Phylogenetic	$\boldsymbol{U} = \frac{\sum_i^n \boldsymbol{bi}\lvert A_i - B_i\rvert}{\sum_i^n \boldsymbol{bi}}$; where b_i is the branch length from branch i, A_i is the number of sequences/reads from branch i in population A and B_i is the number of sequences/reads from branch i in population B	Phylogenetic tree	A phylogenetic tree which indicates from which sample the sequences are from at the end of the node (from one sample, both samples, etc.)	If a node is shared between samples; the branch length will be shared indicating a similarity.	Pros: can compare samples from different conditions Cons: Gives to much weight to rare OTUs	[64, 104, 159-161]
Quantitative							
Sorenson quantitative index/ Bray-Curtis Index	Taxon	$\boldsymbol{BC_{ij}} = \frac{S_i+S_j-C_{ij}}{S_i+S_j}$; where S_i is the number of species in population i, S_j is the number of species in population j and C_{ij} is the total number of species (at the location with the fewest species)	OTU table	A value between 0 and 1	The closer the number is to one, the more similar the samples are	Pros: Robust Cons: sensitive to sample size; samples populations must be the same size	[64, 66, 117, 156, 157, 162]
Morisita-Horn measures	Taxon	$\boldsymbol{C_{MH}} = \frac{2\sum_{i=1}^{s} \boldsymbol{pi_1 pi_2}}{\sum_{i=1}^{s} \boldsymbol{pi_1^2} + \sum_{i=1}^{s} \boldsymbol{pi_2^2}}$; where pi_1 is the proportional abundance (percentage) of species in i in population 1 and pi_2 and pi_1 is the proportional abundance (percentage) of species in i in population 2	OTU table	A value between 0 and 1	The closer the number is to one, the more similar the samples are	Pros: Not sensitive to sample size Cons: can overlook rarer OTUs	[64, 70, 117, 144, 157, 163-165]
Weighted UniFrac	Phylogenetic	$\boldsymbol{U} = \sum_i^n \boldsymbol{bi}\left\lvert\frac{A_i}{A_T} - \frac{B_i}{B_T}\right\rvert$; where b_i is the branch length from branch i, A_i is the number of sequences/reads from branch i in population A, A_T is the total number of sequences/reads in population A, B_i is the number of sequences/reads from branch i in population B and B_T is the total number of sequences/reads in population B	Phylogenetic tree	A phylogenetic tree	A weight is given to the sequences based on their relative abundance. The width of the branch indicates the weight	Pros: can compare samples from different conditions Cons: Gives too much weight to more abundant OTUs	[64, 104, 161, 166, 167]

Table 3.4: Examples of multivariate tests to analyse microbiome data [43]

Test	Abbreviation	Raw data/Distance-based	Type of assumed relationship	Exploratory/ Interpretive/ Discriminatory	Ordination/ Clustering
Orthogonal projections to latent structure discriminant analysis	OPLS-DA	Raw data	Linear	Discriminatory	Ordination
Discriminatory function analysis	DFA/ LDA	Raw data	Linear	Discriminatory	Ordination
Hierarchical clustering	HCA	Distance-based	N/A	Exploratory	Clustering
k-means clustering	N/A	Distance-based	N/A	Exploratory	Clustering
Principal coordinate analysis	PCoA	Distance-based	N/A	Exploratory	Ordination
Nonmetric multidimensional scaling	NMDS	Distance-based	N/A	Exploratory	Ordination
Principal component analysis	PCA	Raw data	Linear	Exploratory	Ordination
Correspondence analysis	CA	Raw data	Unimodal	Exploratory	Ordination
Detrended correspondence analysis	DCA	Raw data	Unimodal	Exploratory	Ordination
Procrustes analysis	PA	Any data	N/A	Interpretive	Ordination
Canonical correspondence analysis	CCA	Raw data	Unimodal	Interpretive	Ordination
Redundancy analysis	RDA	Raw data	Linear	Interpretive	Ordination
Canonical correlation analysis	CCorA	Raw data	Linear	Interpretive	Ordination
Hypothesis Tests					
Multivariate analysis of variance with permutation	PERMANOVA	Distance-based	N/A	Interpretive	N/A
Analysis of group similarities	ANOSIM	Distance-based	N/A	Interpretive	N/A
Mantel test	N/A	Distance-based	N/A	Interpretive	N/A

N/A- Not available/not applicable

Table 3.5: Different tools available in R for differential abundance analysis

Name of the tool	Type of transformation (Normalisation)	Zero Handling	Statistical test(s)	Additional information	Reference
ALDEx	Log-ratio	Dirichlet distribution (Monte-Carlo instances)	Welch's t-test Wilcoxon Kruskal-Wallis (two or more groups)	Requires a large number of samples	[103]
DESeq2	Negative binomial GLM	Bayesian shrinkage	Wald Test	N/A	[61]
metagenomeSeq	Zero-inflated Gaussian (ZIG)	N/A	N/A	N/A	[61]
edgeR	Negative binomial GLM	Bayesian shrinkage	Unknown	More conservative than DESeq2	[61]
ANCOM	Log-ratio	N/A	Mann-Whitney	N/A	[61]

N/A- Not available/not applicable

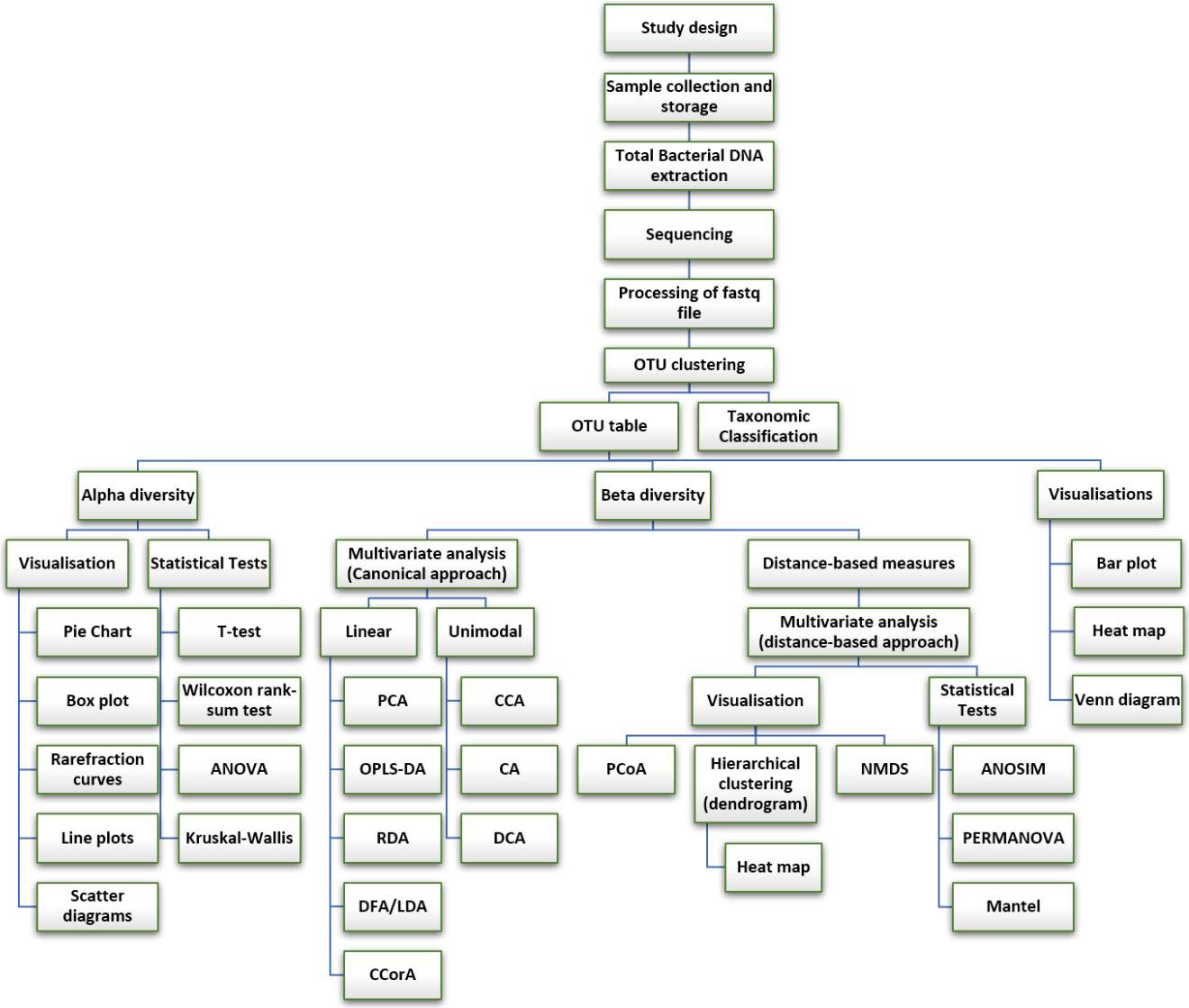

Figure 3.1: **Flow diagram summarising the steps required in microbiome analysis using the targeted approach. Abbreviations: OTU: Operational taxonomic unit; CCA: Canonical correspondence analysis; PCA: Principal component analysis CA: Correspondence analysis; DCA: Detrended correspondence analysis; PCoA: Principal coordinate analysis; NMDS: Nonmetric multidimensional scaling; OPLS-DA: Orthogonal projections to latent structure discriminant analysis; RDA: Redundancy analysis; DFA/LDA: Discriminatory function analysis; CCorA: Canonical correlation analysis; PERMANOVA: Multivariate analysis of variance with permutation; ANOSIM: Analysis of group similarities; ANOVA: Analysis of variance; analysis of similarities (ANOSIM) [20-22, 46, 88, 104, 168-172].**

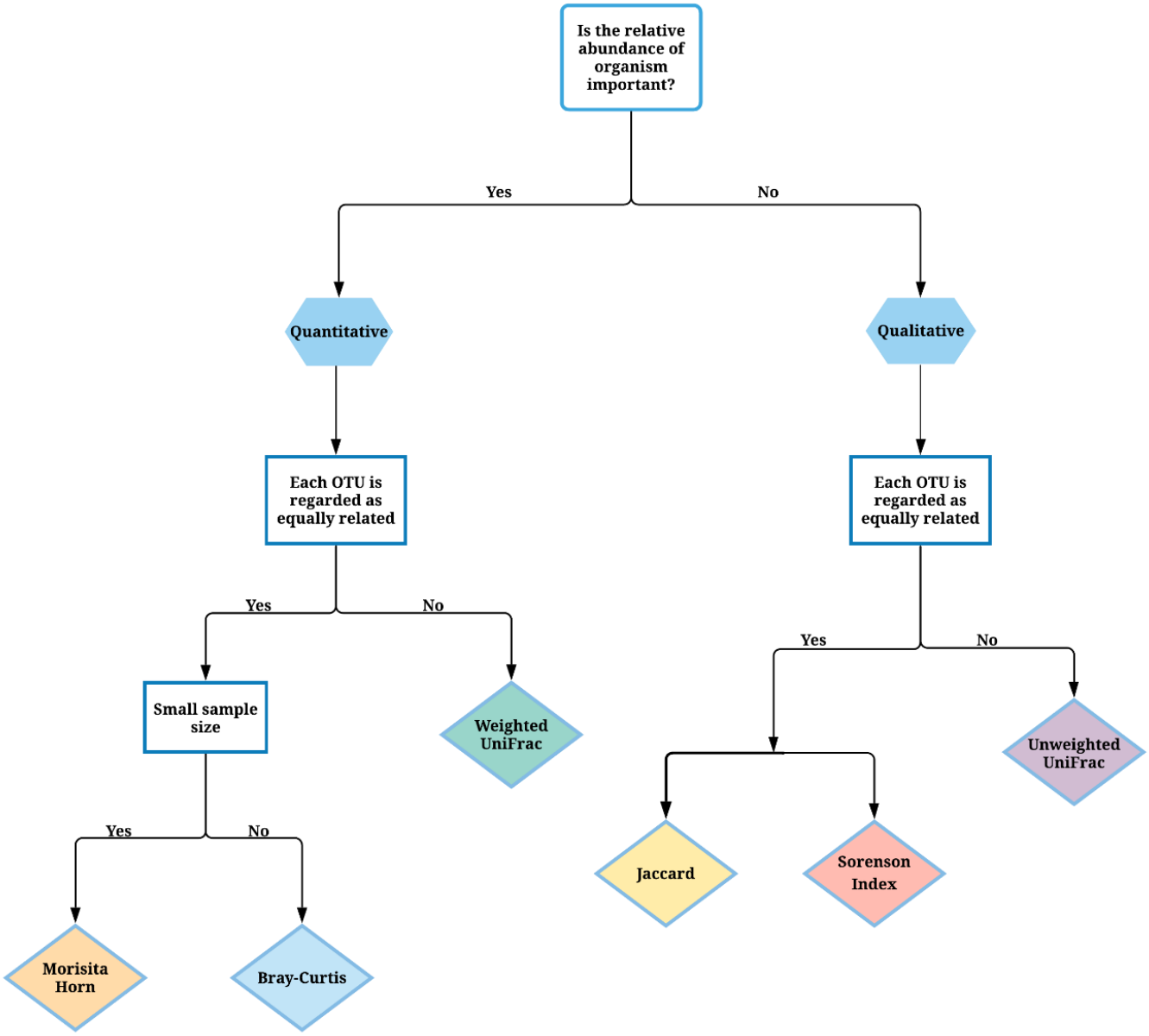

Figure 3.2: **Algorithm to guide the choice of statistical measures to determine beta diversity in microbiome studies. Step 1 is choosing between a quantitative or a qualitative measure. Step 2 is deciding whether to consider the phylogenetic relationship between operational taxonomic units (OTUs). Other considerations, such as sample size, help inform the final decision on which measure to use [64, 66, 70, 104, 117, 143, 144, 155-157, 159, 161, 164, 166, 167].**

CHAPTER 4

Lung microbiome of stable and exacerbated COPD patients in Pretoria, South Africa

The editorial style of the Microbiome Journal was followed in this chapter

Abstract

Background

Chronic obstructive pulmonary disease (COPD) is a progressive lung disease characterised by the occurrence of exacerbations triggered by bacterial or viral infections. The aim of this study was to determine the composition of the lung microbiome and lung virome in patients with COPD in an African setting and to compare their composition between participants with stable and exacerbation states of disease.

Methods

Twenty-four adult patients with COPD were recruited from three hospitals in Tshwane Health District, South Africa. Spontaneously expectorated sputum was collected for microbiological analysis. Bacterial DNA was extracted using the Isolate II Genomic DNA Kit (Bioline, UK). Targeted metagenomics was performed to determine the microbiome composition and analysed using quantitative insights into microbial ecology 2 software. Viral DNA and RNA were extracted from selected samples using the Isolate II Genomic DNA Kit (Bioline, UK). and the QIAmp Viral RNA Kit (Qiagen, Germany) followed by conversion to cDNA. Shotgun metagenomics sequencing (virome) of pooled DNA and RNA was performed on the MiSeq platform and analysed using Kraken 2 software.

Results

The most abundant phyla across all microbiome samples were *Firmicutes* (ranging from 41% to 91%), *Proteobacteria* (ranging from 3% to 62%), *Bacteroidetes* (ranging from 3% to 22%) and *Actinobacteria* (ranging from 1% to 22%). The following genera were most prevalent: *Haemophilus*, *Streptococcus*, *Veillonella*, *Prevotella* and *Granulicatella*. Both Chao1 [median values of 147.06 and 115.56, interquartile (IQR) values of 63.67 and 17.92, p-value= 0.58] and Simpson diversity measures (median values of 0.84 and 0.86, IQR values of 0.13 and 0.08, p-value=0.72) of the microbiome did not differ significantly between participants with stable (n=18) and exacerbation states (n=6) of COPD. No distinct clusters were observed using PCoA

and weighted UniFrac measures for beta diversity. However, a difference in the abundances between stable and exacerbation states of disease was observed for the following genera: i) *Actinomyces* (lower), ii) *Granulicatella* (higher), iii) *Haemophilus* (higher) and iv) *Veillonella* (lower). Virome analysis showed a high abundance of the BeAn 58058 virus, a member of the *Poxviridae* family, in all six samples (abundances ranged from 90% to 94% across the samples).

Conclusions

This study is among the first to report lung microbiome composition in COPD patients from Africa. Compared to the other settings relatively high frequencies of *Haemophilus* and low frequencies of *Streptococcus* genera (although this genus was present in all samples) were observed. In this small sample set, no differences in alpha or beta diversity between stable and exacerbation disease states were observed, but an unexpectedly high frequency of BeAn 58058 virus was observed. These observations highlight the need for further research of the lung microbiome of COPD patients in African settings.

4.1 Background

Chronic obstructive pulmonary disease (COPD) is a progressive lung disease that results in progressive airflow limitation (i.e. obstruction) [1, 2]. Chronic obstructive pulmonary disease is one of the world's leading causes of death (ranked as the third leading cause of death) [3]. Symptoms of COPD include a chronic cough, dyspnoea and sputum production [4, 5]. These symptoms affect the quality of life of the individual suffering from this disease [6]. There is limited data concerning the prevalence of COPD in the African continent; the last reported prevalence data on COPD in South Africa was in 2005 (19% in men and women over 40 years of age); this data was from a single city i.e. Cape Town and may not be representative of the whole country [7-10]. This disease has been linked to smoking, exposure to occupational dust (e.g. working in a mine), burning of biomass and fossil fuels, previous *Mycobacterium tuberculosis* (TB) infection and to HIV; all of these risk factors are highly prevalent in South Africa [10].

Exacerbation of airway inflammation and its associated symptoms are other factors that affects the quality of life for these individuals [10]. Patients suffering from COPD often move between a stable state of disease (where symptoms are absent to mild) to an exacerbation state of disease (defined as worsening of symptoms, respiratory and/or non-respiratory) [11, 12]. The frequency of these exacerbations increases over the course of the disease, as the lung damage due to COPD progresses [13]. Exacerbations can be triggered by: i) environmental pollutants, ii) an unknown cause or iii) infection with bacteria and/or viruses [14]. Bacterial and viral infections account for between 30% to 50% of all exacerbations [15]. However, bacteria have been detected in the stable state of disease as well and the association between these microorganisms and disease is unclear [16, 17].

To better understand the role of microorganisms in COPD disease, the use of next-generation sequencing (NGS) can be employed to study the microbiome (defined as the genetic material of the microorganism in the community) [18]. Next-generation sequencing is high-throughput, parallel sequencing technology [19, 20]. It has been used to sequence whole genomes of bacteria and viruses, perform transcriptomics (studying the complete set of RNA transcripts produced by the genomes) and to study the microbiome/metagenome [19]. The advantage of NGS over culturing and other molecular methods is that it can detect unculturable bacteria and provide information regarding the diversity, composition and functional roles of members of the microbiome [21, 22]. An important drawback is that the cost of sequencing is still relatively

high, especially in the African continent [23]. Next-generation sequencing can be employed in one of two ways: i) using a targeted approach or ii) using a shotgun metagenomics approach [24, 25].

The targeted approach is commonly used to study the microbiome and is employed by targeting the 16S rRNA gene [26, 27]. This gene is useful for studying the bacterial microbiome as it is universally present and conserved within all bacteria [28-30]. Studying the virome, i.e. viral component of the microbiome is more challenging as: i) most viruses are difficult to culture, ii) there is no consensus sequence to study viruses and iii) viruses are diverse and may be ssDNA, ssRNA, dsDNA or dsRNA [31-33]. By using shotgun metagenomics (i.e. random sequencing of the DNA from the microbial community) along with cDNA synthesis to study the virome, these challenges can be overcome [34-36].

In South Africa, there is no data on the composition of the lung microbiome in COPD patients. Previous studies on the lung microbiome of COPD patients were conducted in Europe and the USA [37-39]. Furthermore, there have been limited studies on the lung virome in COPD [40, 41]. It is important to study not only the microbiome in the African continent in countries, such as South Africa but also the virome as local environmental conditions, e.g. climate and clinical comorbidities, e.g. HIV (which is highly prevalent in sub-Saharan Africa) have the potential to affect the microbiome. The aim of this study was to determine the composition of the lung microbiome and the lung virome in the sputum of COPD patients and to compare their composition between stable and exacerbation states of disease.

4.2 Methods

4.2.1 Study setting and patient recruitment criteria

Chronic obstructive pulmonary disease (COPD) patients admitted to or attending clinics (for scheduled check-ups at the lung unit, HIV clinics or at the private practice) at one of three hospitals in the Tshwane Health district (one academic, one district and one private) were invited to participate in the study. Written informed consent was obtained from all participants if the inclusion and exclusion criteria were met (Table 4.1). The planned patient groups were as follows: i) stable state COPD in HIV-positive individual, ii) stable state COPD in HIV-negative individuals, iii) exacerbation state COPD in HIV-positive individuals and iv) exacerbation state COPD in HIV-negative individuals. Healthy controls were not included as sputum specimens are difficult to obtain from healthy individuals. The sample size per group

was determined as follows: a sample size of 20 per group was considered more than adequate to identify meaningful shifts and differences in the microbiome. All participants that met the inclusion and exclusion criteria were included in the study. Participants were classified as either in the stable or in the exacerbation state based on the definition by Vogelmeier *et al.* (2017). An exacerbation state was defined as acute worsening of respiratory symptoms and any patient not in an exacerbation state was considered stable. Ethical approval was granted from the Research Ethics Committee, Faculty of Health Sciences, University of Pretoria (REC no: 237/2017). All aspects of the research were conducted by the candidate unless otherwise stated.

Table 4.1: Inclusion and exclusion criteria for COPD patients in this study

Stable state	
Inclusion criteria	**Exclusion criteria**
HIV patients on antiviral therapy (ART)	Active tuberculosis infection (receiving treatment)
Over 40 years of age	Receiving immunosuppressants
Able to provide informed consent	Cancer
	Lung surgery within the last six months
	Unable to answer the questionnaire (CDQ)
	Antibiotics within last month
Exacerbation state	
Inclusion criteria	**Exclusion criteria**
HIV patients on antiviral therapy (ART)	Active tuberculosis infection (receiving treatment)
Over 40 years of age	Receiving immunosuppressants
Able to provide informed consent	Cancer
Increased/worsening of respiratory symptoms 48 h before the visit	Lung surgery within the last six months
	Unable to answer the questionnaire (CDQ)
	Unable to give informed consent
	Antibiotics therapy 24 h before admission
	Antibiotic therapy administered for more than 12 h after admission

ART: Antiviral therapy
CDQ: Chronic obstructive pulmonary disease diagnostic questionnaire
h: hour
HIV: Human immunodeficiency virus

4.2.2 Extraction of DNA and RNA and cDNA synthesis

Spontaneously expectorated sputum specimens were collected from participants at a single time point, transported on ice and stored at -80°C (Innova U535 Upright, Eppendorf, Germany) until batch processing could occur (no preservation medium was used). The sputum specimens were treated with an equal volume of 0.1% dithiothreitol (DTT) (Roche Diagnostics, Switzerland) to reduce sputum viscosity and were homogenised for 30 seconds (sec) (Vortex-Genie® 2; Scientific Industries Inc., USA) [42-44]. The samples were split into three aliquots for: i) bacterial DNA extraction (aliquot 1), ii) viral DNA and RNA extraction (aliquot 2) and iii) storage at -80°C (aliquot 3, for future studies) (Innova U535 Upright, Eppendorf, Germany).

The bacterial extraction aliquot (aliquot 1) was centrifuged (Spectrafuge™ 24D, Labnet International Inc., USA) at 4 000 x *g* for 30 min before extraction. The pellet was used for extraction and bacterial DNA was extracted using the Isolate II Genomic DNA Kit (Bioline, UK). The manufacturer's instructions (protocol 9.2) were followed with the addition of 10 mg/mL lysozyme (Sigma-Aldrich, USA), 3 U/µL lysostaphin (Sigma-Aldrich, USA) and 6.75 µL of 10 U/µL mutanolysin (Sigma-Aldrich, USA) to the hard-to-lyse buffer [20 mM Tris (Sigma-Aldrich, USA) pH 8.0; 1% Triton X-100 (Amresco, USA); 2 mM EDTA (Sigma-Aldrich, USA)].

The viral DNA and RNA aliquot was treated with DNase I to remove host (human) DNA [10 U/mL TURBO™ DNase (Ambion, USA)] at 37°C for 30 min (AccuBlock™ Digital Dry Bath, Labnet International Inc., USA), followed by inactivation with 15 mM ethylenediaminetetraacetic acid (EDTA) (Sigma-Aldrich, USA) at 75°C for 10 min (AccuBlock™ Digital Dry Bath, Labnet International Inc., USA) according to the manufacturer's instructions [45]. The samples (after processing) were further split into two separate aliquots for DNA and RNA extraction, respectively, i.e. aliquot 2 was split into two aliquots (aliquots 2.1 and 2.2.). The viral DNA aliquot (aliquot 2.1) was centrifuged (Spectrafuge™ 24D, Labnet International Inc., USA) at 4 000 x *g* for 30 min before extraction. The pellet was used for extraction with the Isolate II Genomic DNA Kit (Bioline, UK) according to the manufacturer's instructions (protocol 9.13). The RNA extraction (aliquot 2.2) was performed according to the manufacturer's instructions using the QIAmp Viral RNA kit (Qiagen, Germany). The RNA was converted to cDNA using the SuperScript First-Strand Synthesis System for RT-PCR (Invitrogen, USA) using the random hexamer primers supplied according to the manufacturer's instructions (BioRad T100™ Thermal Cycler, BioRad Laboratories Inc., USA). The second synthesis (to convert cDNA and ssDNA) was performed using Klenow Fragment (New England Biolabs, USA) (Bio-rad T100™ Thermal cycle, Bio-rad Laboratories Inc., USA). The converted cDNA and ssDNA (along with dsDNA) were amplified with KAPA HiFi polymerase (Roche, Switzerland) and the FR20RV primer as described previously (BioRad T100™ Thermal Cycler, BioRad Laboratories Inc., USA) [46]. All converted cDNA, ssDNA and dsDNA were pooled together.

4.2.3 Targeted and shotgun metagenomics approach

The targeted metagenomics was performed at Inqaba Biotechnical Industries, South Africa. Steps performed by the company included PCR amplification, library preparation, purification

of the products, indexing and sequencing of V1-V3 region of the 16S rRNA gene using the MiSeq platform (Illumina, USA) at Inqaba Biotec Industries, South Africa. After, the targeted approach, a subset of six (representative) samples were selected for virome sequencing (due to high cost of sequencing) according to the following criteria: i) samples should be from both states of disease and ii) samples should be representative of the diversity in the samples; based on the number of operational taxonomic units (OTUs) as follows: i)one for low diversity (<40 OTUs), ii) one for intermediate diversity (between 40 OTUs and 50 OTUs) and iii) one for high diversity (>50 OTUs). Shotgun metagenomics of the amplified and pooled virome samples was performed using the MiSeq platform (Illumina, USA) at the National Institute for Communicable Diseases of South Africa (NICD). Steps performed by the company included sample purification, library preparation, indexing and sequencing. The fragments of the 16S rRNA sequences were analysed using QIIME2 (using Greengenes database) following the moving picture tutorial (which included quality control steps done using Deblur that removed low-quality sequences and ensured that all sequences had the same read length). Human DNA was removed from the virome sequences using Bowtie2 (with Hg38 reference genome) and the virome sequences were analysed using Kraken 2 [on the Galaxy platform (with the virome (2019) databases] [47-49]. The viral sequencing results were compared to the virus-host database (https://www.genome.jp/virushostdb/view/) to determine the host of the viruses identified [50].

4.2.4 Statistical analysis and data visualisation

The data were analysed on R using the following packages: i) Qiime2R (to import QIIME2 data), ii) phyloseq (alpha diversity, beta diversity, statistical tests, principal coordinate analysis analysis (PCoA), hierarchical clustering and relative abundance of the taxa), iii) ggplot2 (for the plotting of all graphs) and iv) DESeq2 (to determine if there was a log2fold difference). A p-value < 0.05 was considered significant (for any of the statistical tests).

4.3 Results

4.3.1 Patient demographics

A total of 80 participants were planned to be included in the study, however due to the strict inclusion and exclusion criteria as well as the limited number of patients attending the clinic or being admitted to the hospital, this number could not be realised. A total of 24 participants were enrolled in the study; 18 males and six females aged from 50 years old to 82 years old (median= 60 years old with a standard deviation of 7.34). Only one of the participants was HIV-positive.

Participants were distributed across the three hospitals as follows: i) Hospital A (Tertiary Academic Hospital): 16 participants, ii) Hospital B (District Hospital): one participant and Hospital C (Private Hospital): seven participants. Eighteen of the participants were in the stable state of disease at the time of sampling and six of the participants were in the exacerbation state of disease at the time of sampling. Four of the participants had never smoked, nine of the participants were current smokers and 11 participants had stopped smoking.

4.3.2 The sputum microbiome

A total of 631 OTUs were identified across the 24 samples for the microbiome. These OTUs were divided into 14 phyla, 27 classes, 37 orders, 70 families and 77 genera. Twenty-two percent (140/631) of all OTUs could be classified to a species level. The relative abundance of unclassified species ranged from 32% to 94% between samples.

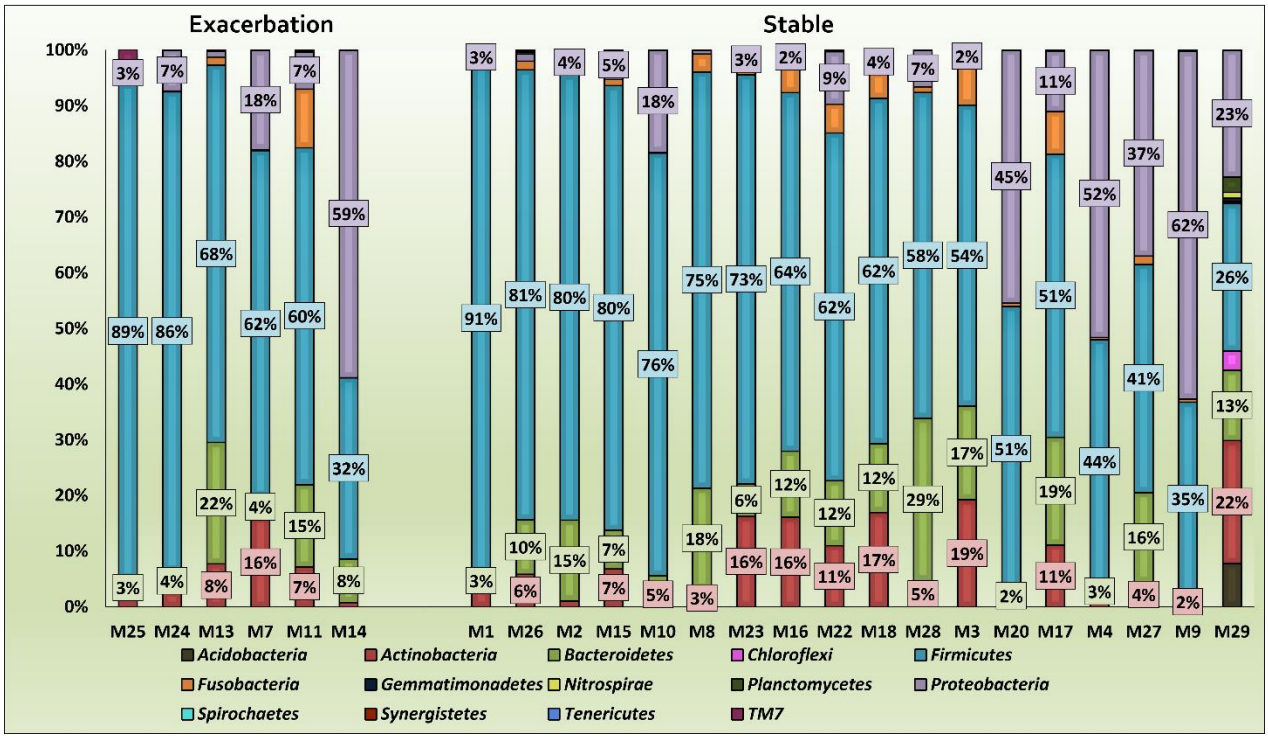

Figure 4.1: Bar plots showing the relative abundance of the differing phyla by disease state occurring in the sputum microbiome of 24 COPD participants using targeted metagenomics across the different samples. *Firmicutes* are shown in blue, *Proteobacteria* in purple, *Bacteroidetes* in green and *Actinobacteria* in red. The graph is separated into the exacerbation state (n=6) and stable state (n=18). The specimens are ordered according to the prevalence of *Firmicutes*.

The most abundant genera were *Streptococcus* (detected in all 24 samples, with abundances ranging from 19% to 82%), *Haemophilus* (detected in all 24 samples, with abundances ranging from 0.02% to 61%), *Prevotella* (detected in all 24 samples, with abundances ranging from 0.1% to 22%), *Veillonella* (detected in all 24 samples, with abundances ranging from 0.15% to 19%) and *Granulicatella* (detected in all 24 samples, with abundances ranging from 0.12% to 11%). The most abundant species in the 22% of the OTUs that could be classified to species level were: i) *Haemophilus influenzae* (detected in 21/24 samples, with abundance ranging from 0.01% to 61%), ii) *Haemophilus parainfluenzae* (detected in 22/24 samples, with abundance ranging from 0.01% to 16%), *Prevotella melaninogenica* (detected in all 24 samples, with abundance ranging from 0.08% to 15%), *Veillonella dispar* (detected in 21/24 samples, with abundance ranging from 0.02% to 9%) and *Veillonella parvula* (detected in 23/24 samples, with abundance ranging from 0.07% to 9%). Additionally, sample M20 showed a high abundance of *Serratia marcescens* (41%), sample M4 showed a high abundance of *Pseudomonas* spp. (49%) (could not be classified to a species level) and sample M26 showed a high abundance of *Staphylococcus aureus* (13%).

4.3.3 Comparison of exacerbation and stable states of disease for the microbiome

The relative abundance of the *Actinobacteria*, *Bacteroidetes*, *Firmicutes*, *Fusobacteria* and *Proteobacteria* phyla differed across the disease states; with a higher abundance of *Firmicutes* (63% in the exacerbation state and 61% in the stable state) and a lower abundance of *Actinobacteria* (5% in the exacerbation state and 8% in the stable state), *Bacteroidetes* (9% in the exacerbation state and 11% in the stable state) and *Proteobacteria* (17% in the exacerbation state and 19% in the stable state), during the exacerbation state (Figure 4.2). At a genus level (Figure 4.3), the exacerbation state showed changes in 75 genera; with 49 genera that had a lower relative abundance and 26 genera that had a higher abundance. Key genera that showed lower relative abundance during the exacerbation state included *Porphyromononas* (0.19% in the exacerbation state and 3.92% in the stable state), *Serratia* (0.00% in the exacerbation state and 2.99% in the stable state), *Staphylococcus* (0.00% in the exacerbation state and 1.02% in the stable state) and *Streptococcus* (47.88% in the exacerbation state and 49.61% in the stable state). Genera that showed a higher relative abundance in the exacerbation state included *Granulicatella* (5.30% in the exacerbation state and 3.06% in the stable state), *Haemophilus* (16.82% in the exacerbation state and 11.08% in the stable state), *Prevotella* (10.02% in the exacerbation state and 7.87% in the stable state) and *Veillonella* (6.92% in the exacerbation

state and 4.44% in the stable state). Although, the relative abundance differed across the disease state, with DESeq2 analysis no significant difference was observed.

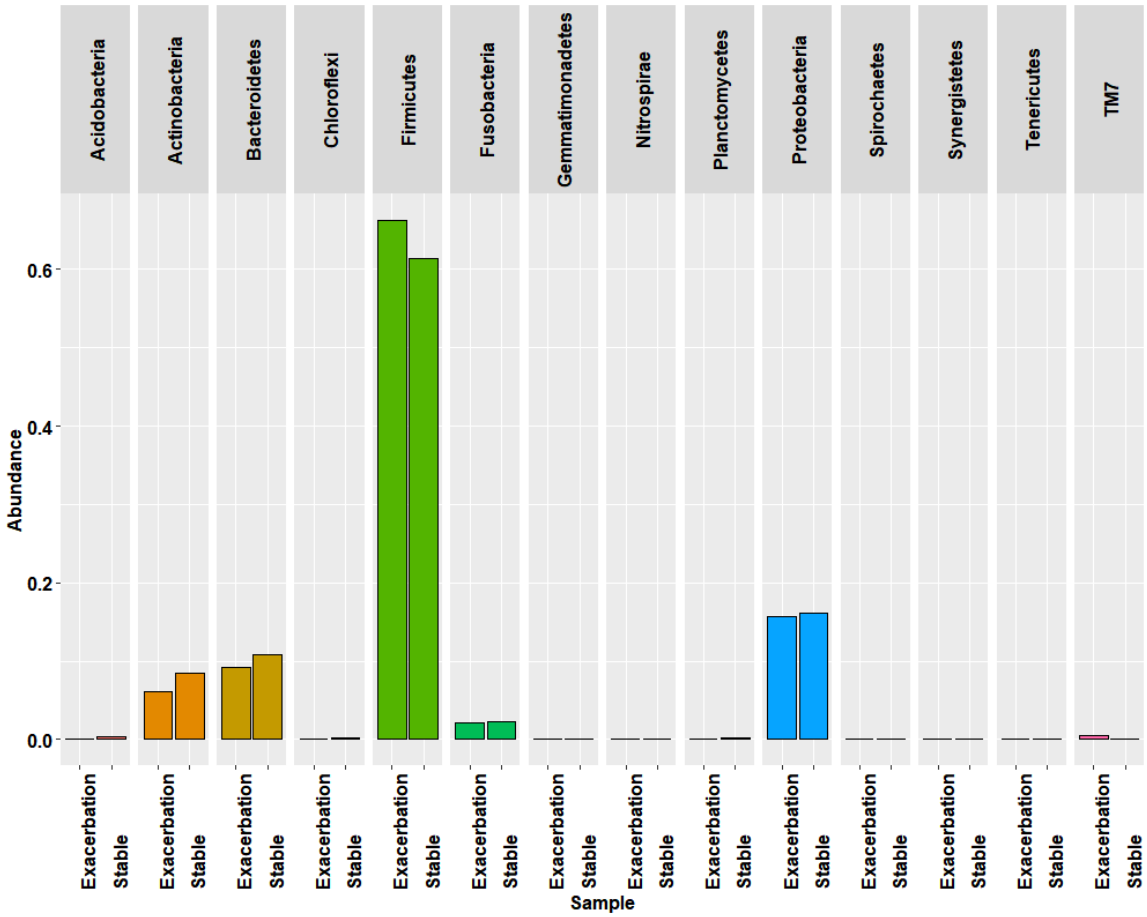

Figure 4.2: Bar plots showing the relative abundance of the different phyla in the sputum microbiome of COPD participants as determined by targeted metagenomics compared across the exacerbation state (n=6) and stable state (n=18). The relative abundance is shown as a proportion of total abundance for the disease state.

Figure 4.3: Bar plots showing the relative abundance of the genera in the sputum microbiome of COPD participants by disease state. The relative abundance is shown as a proportion of total abundance for the disease state.

There was no significant difference in the alpha diversity between disease states for the microbiome using the Wilcoxon sum rank test for both Chao1 (p-values=0.58) and Simpson diversity measures (p-value=0.72) (Figure 4.4). Beta diversity measures showed no distinct clustering for any of the variables using PCoA and the weighted UniFrac measures i.e. there was overlap between the two disease states (Figure 4.5).

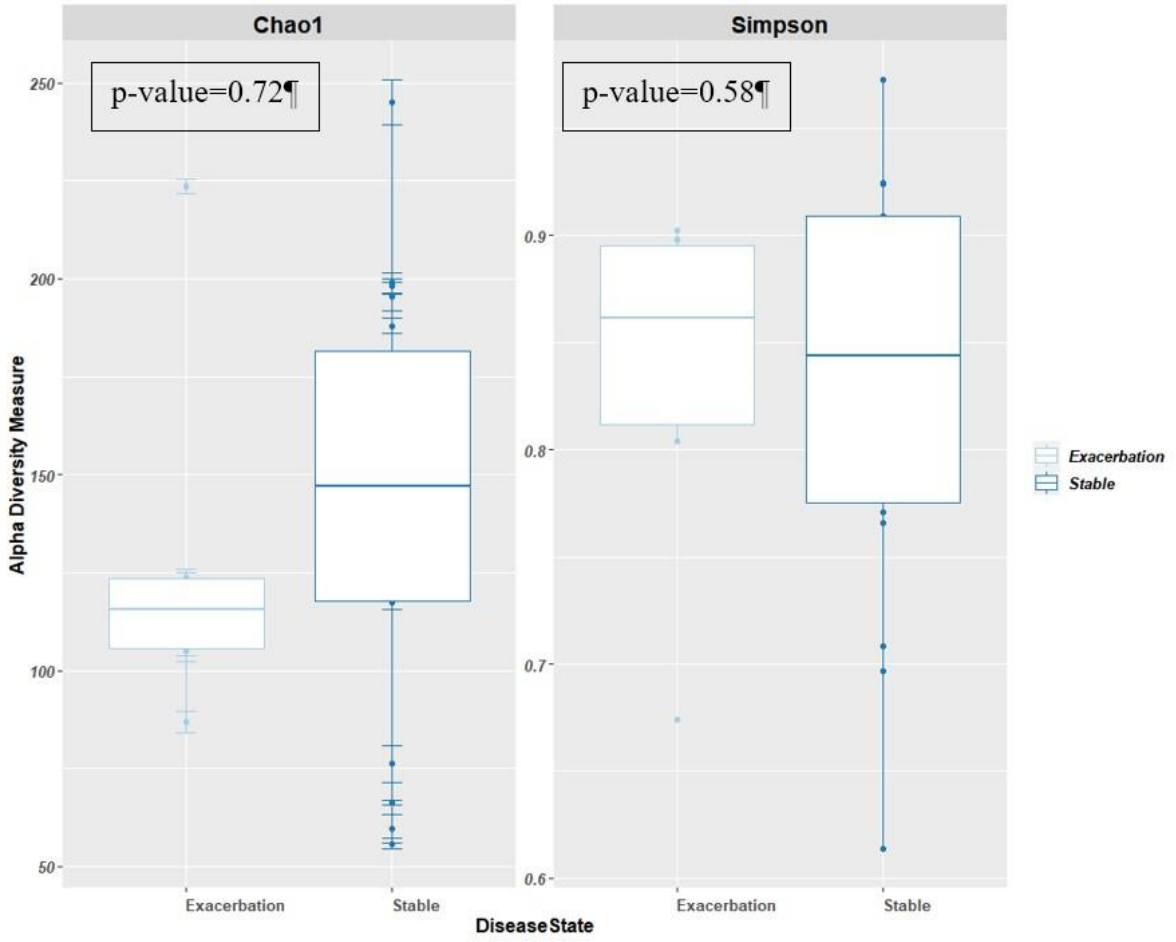

Figure 4.4: **The alpha diversity boxplot of the sputum microbiome compared across the exacerbation state (n=6) and stable state (n=18) of COPD using Chao1 and Simpson diversity measures. Each dot on the graph represents a sample. The boxes represent the interquartile range (IQR) and the horizontal line represents the median. The median values for the Chao1 diversity measure were as follows: i) stable state=147.06 and ii) exacerbation state=115.56. The median values for the Simpson diversity measures were as follows: i) stable state=0.84 and ii) exacerbation state=0.86. The IQR values for the Chao1 diversity measure were as follows: i) stable state=63.67 and ii) exacerbation state=17.92. The IQR values for the Simpson diversity measure were as follows: i) stable state=0.13 and ii) exacerbation state=0.08.**

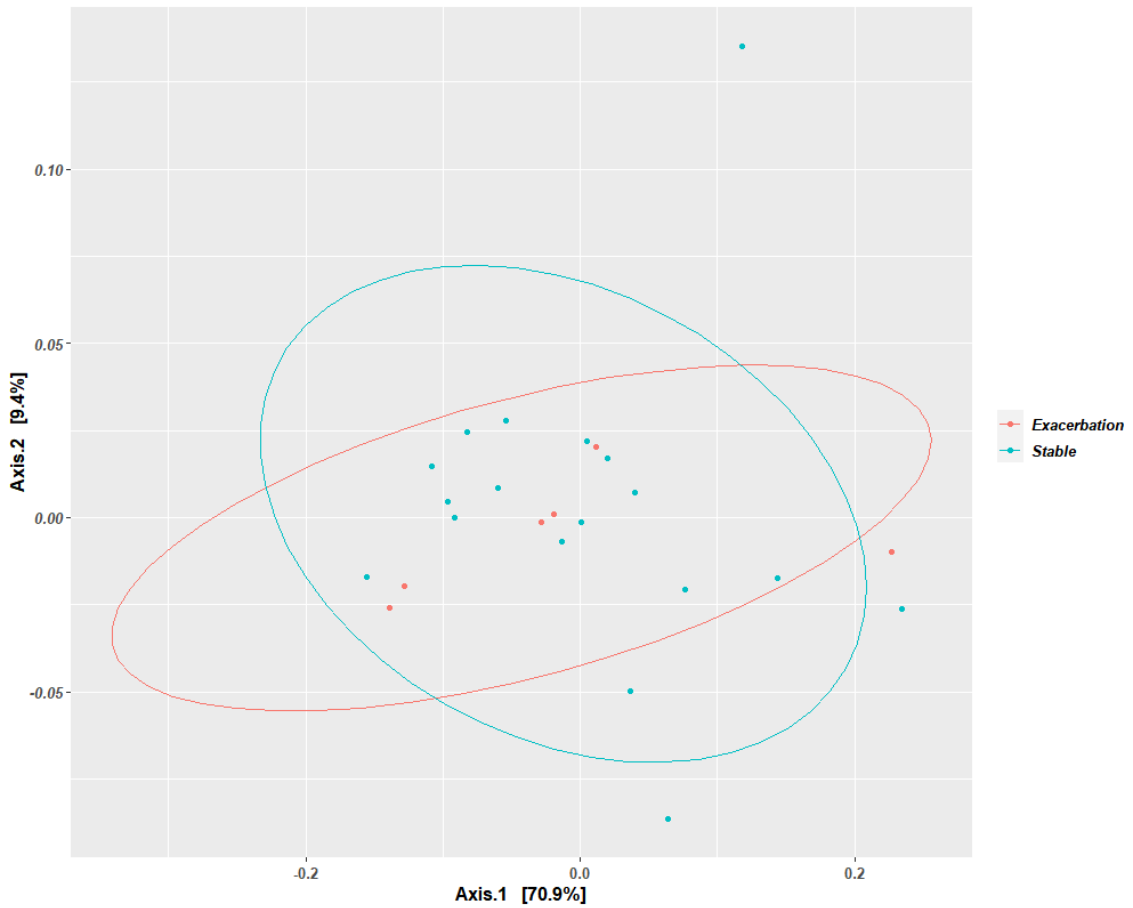

Figure 4.5: Principal coordinate analysis (PCoA) plot derived using the weighted UniFrac diversity measure comparing the different disease states of COPD in the sputum microbiome The ellipses show the different states of disease with the exacerbation state (n=6) indicated in red and the stable state (n=18) indicated in blue; with the dots represent in each sample.

4.3.4 The sputum virome

Six samples were selected for virome analysis as follows: i) one low diversity sample (<40 OTUs) from the exacerbation state of disease, ii)) one low diversity sample (<40 OTUs) from the stable state of disease, iii) one medium diversity sample (between 40 OTUs and 50 OTUs) from the exacerbation state of disease, iv) one medium diversity sample (between 40 OTUs and 50 OTUs) from the stable state of disease, v) one high diversity sample (>50 OTUs) from the exacerbation state of disease and vi) one high diversity sample (>50 OTUs) from the stable state of disease. A total of 3 480 operational taxonomic units (OTUs) were identified across the six samples for the virome. The taxonomic classification identified 16 phyla, 34 classes, 53 orders, 141 families and 826 genera. Most of the OTUs [95% (3 306/3 480)] could be classified up to a species level. The most abundant family across all samples was the *Poxviridae* family

(detected in all six samples, with abundances ranging from 90% to 93%), followed by the bacteriophage families *Myoviridae* (detected in all six samples, with abundances 0.63% to 2.11%) and *Siphoviridae* (detected in all six samples, with abundances 1.08% to 1.55%) and lastly by *Herelleviridae* (detected in all six samples, with a abundances ranging from 0.08% to 0.16%) (Figure 4.6).

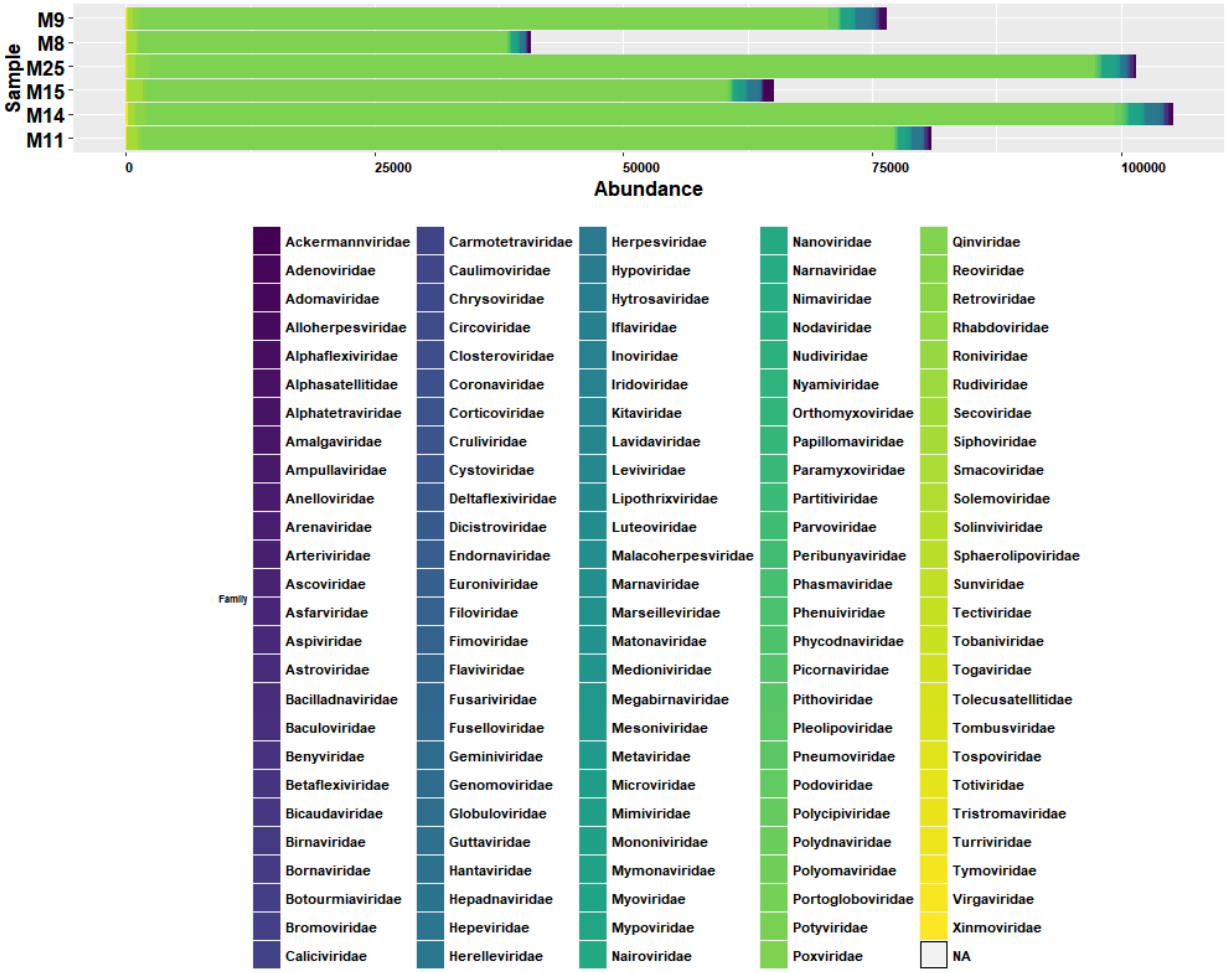

Figure 4.6: Bar plots showing the abundance of viruses at a family level; the most prevalent families were as follows: i) *Poxviridae* (indicated in light green), ii) *Siphoviridae* (indicated in green-yellow), iii) *Myoviridae* (indicated in dark green); iv) *Herelleviridae* (indicated in blue). Viruses that had no taxonomic designation at the phyla or family level are indicated by NA (not available). The abundance is shown as the number of operational taxonomic units.

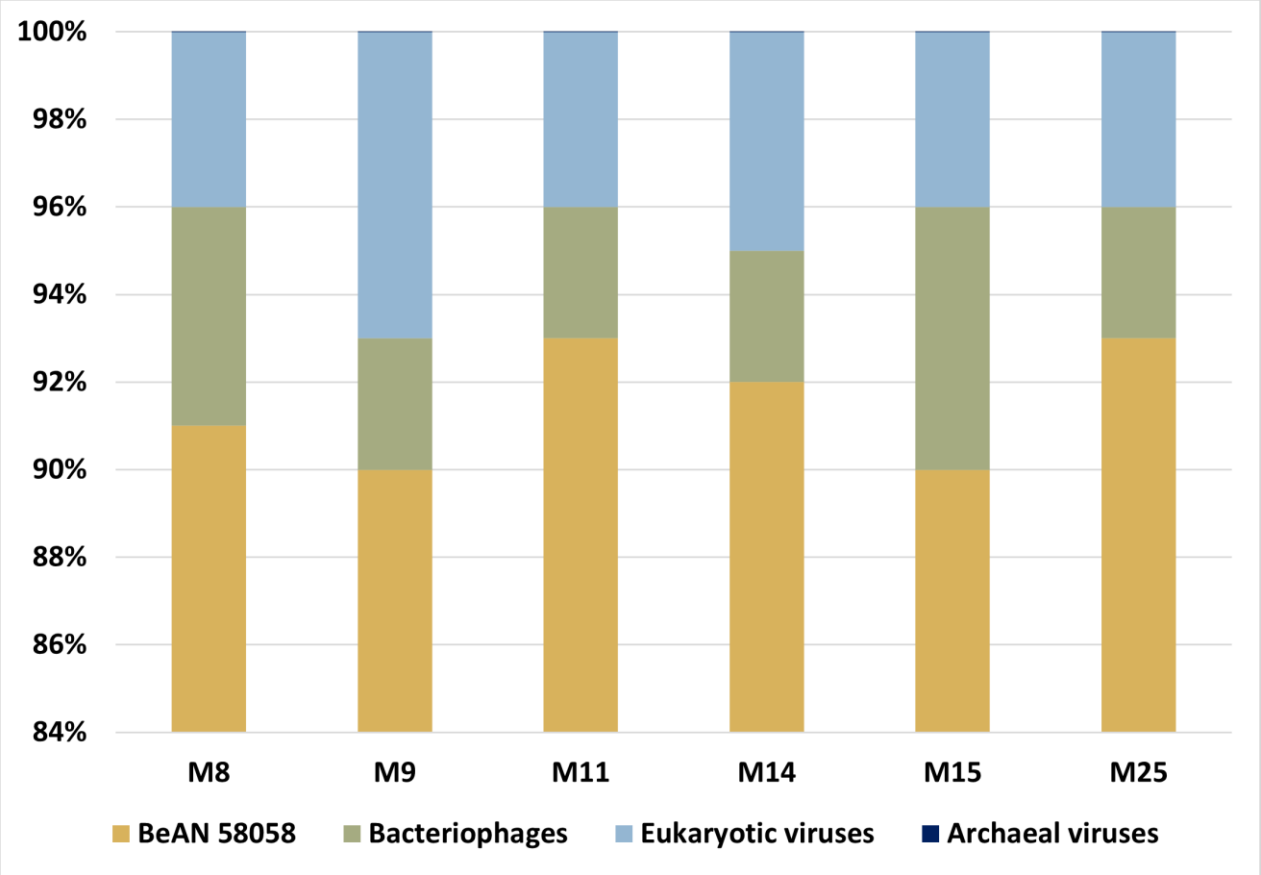

Figure 4.7: Bar plot showing the distribution of viruses across the different samples (n=6) of the sputum virome of COPD participants based on their hosts.

The most prevalent species was BeAn 58058, a member of the *Poxviridae* family that was detected in all specimens sent for virome sequencing (Figure 4.7) followed by bacteriophages (associated with both Gram-positive and Gram-negative bacteria). Most of the viruses identified were dsDNA viruses (ranging from 97.23% to 98.15%).

4.4 Discussion

In this study, the composition of the sputum microbiome of COPD participants was investigated and was compared between the different disease states, i.e. stable state of disease and exacerbation state of disease. Two phyla predominated, *Firmicutes* and *Proteobacteria*; with *Streptococcus* and *Haemophilus* being the most prevalent genera. However, this study observed no significant differences between the exacerbation and stable states of disease in COPD, in terms of relative abundance, alpha diversity and beta diversity for the sputum microbiome in COPD. With the virome, a high prevalence of the virus, BeAn 58058 was observed. In this study, there was difficulty in recruiting HIV-positive individuals with COPD and as a result,

only a single HIV-positive participant was recruited in this study. There were several possible reasons for the low recruitment rate of HIV-positive individuals suffering from COPD including: i) active TB cases were excluded from the study (no participants with HIV-TB overlap), ii) the HIV population at the hospitals may have been a younger population and iii) the HIV population in South Africa is mostly female and COPD is often underdiagnosed in the female population [51-58].

In both the stable state and exacerbation states of disease, the results showed that four phyla dominated, i.e. *Firmicutes* (ranging from 26% to 91%), *Proteobacteria* (ranging from 2% to 62%), *Bacteroidetes* (ranging from 2% to 29%) and *Actinobacteria* (ranging from 1% to 22%). This is in agreement with previous studies conducted on the lung microbiome (including the healthy lung and other disease states), that have observed that these four phyla are known to be dominant in the lung [59, 60; 68]. Similar to this study, those studies also had small sample sizes (less than 30 participants), however, these studies: i) had different patient groups (included asymptomatic smokers, asthmatics, healthy controls and younger patients), ii) used different specimen types, such as bronchoalveolar lavage (BAL) (invasive specimen) and iii) used different sequencing technologies, such as 454 pyrosequencing [59-62; 68]. Despite these differences, these four phyla have always dominated in the lung, although the prevalence of these phyla may differ in specific diseases, with some phyla, such as *Proteobacteria* being more prevalent in asthma and COPD [59-62]. However, the changes in the microbial composition of the COPD lung (e.g. the higher prevalence of *Proteobacteria*), occurs only once the disease has progressed; in mild COPD disease, the microbial composition is similar to that of the healthy lung as can be observed in this study where *Firmicutes* has a higher prevalence [63, 64]. In this study, when comparing the disease states, a higher abundance of the *Firmicutes* phylum (2% higher in the exacerbated state) and a lower abundance of the *Proteobacteria* (2% lower in the exacerbation state), *Actinobacteria* (3% lower in the exacerbation state) and *Bacteroidetes* phyla (2% lower in the exacerbation state) in the exacerbation state were observed. This is in agreement with studies that have compared stable and exacerbation states of COPD disease and have observed an increase in one or more phyla (either *Proteobacteria* or *Firmicutes*) often associated with a decrease in the other phyla (either *Proteobacteria* or *Firmicutes*) [65-70]. None of the studies specified the percentage increase of either phylum during exacerbations; however, these studies did indicate which phyla increased, except Millares *et al.* (2015) [65-70]. In most of these studies *Proteobacteria* were higher, however in the Jubinville *et al.* (2018) and Wang *et al.* (2020) studies, *Firmicutes* were higher as well [65-70]. All of these studies

were conducted using sputum specimens, had a variety of different sample sizes (ranging from nine participants to 281 participants), were conducted in USA, Europe and China, used different sequencing technologies (454 sequencing, MiSeq sequencing and PhyloChip) and targeted different regions of the 16S rRNA gene (V1-V3, V3-V5, V6-V8, V4, V3-V4 or full-length) [65-70]. No association were noted between the choice of primer pair and the most prevalent phyla.

The genera that showed the highest frequency in this study, across both disease states, were *Granulicatella (Firmicutes)*, *Haemophilus (Proteobacteria)*, *Prevotella (Bacteroidetes)*, *Streptococcus (Firmicutes)* and *Veillonella (Firmicutes)*. This is in agreement with previous studies conducted on the microbiome of COPD lung and the healthy lung, where these genera along with *Pseudomonas* and *Porphyrononas* are detected in high abundances (independent of the disease state) [71, 72]. Most of these studies were conducted in the USA or Europe using either sputum or BAL specimens and 454 sequencing. The genera identified in this study (during stable state and exacerbation state) were similar to a study conducted by Wang *et al.* (2016) [68]. However, the abundances of these genera differed when compared to the study by Wang *et al.* (2016): i) some genera, such *Haemophilus* had a higher prevalence [5.7% increase in this study and 3% increase in Wang *et al.* (2016)] and ii) some genera, such as *Streptococcus* [1.7% decrease in this study and 3% decrease in Wang *et al.* (2016)] had a lower prevalence. The differences in abundances of the genera could be attributed to the different study population and setting; the study by Wang *et al.* (2016), had a larger study population (n=87) compared to this study (n=24) and was conducted in the United Kingdom (UK) (developing country vs developed country). The difference in the sequencing methodology between this study and the study by Wang *et al.* (2016) could account for the differing prevalence as well; this study used targeted the V1-V3 regions of the 16S rRNA gene using MiSeq platform (Illumina, USA) whereas Wang *et al.* (2016) targeted the V3-V5 regions of the 16S rRNA gene using 454 sequencing (Roche Diagnostics, UK). Geographical location and local environmental conditions, such as air pollution have been shown to affect the lung microbiome and could explain the difference in relative abundance between the two studies [13, 73]. Additionally, seasonal variation may play a role in the bacteria identified [74]. Most of the exacerbation samples in this study were collected in either autumn or winter. In Pretoria, the dry season is in winter which is in contrast to the United Kingdom, where the dry season generally falls in summer.

Additionally, the bacteria that showed a higher prevalence (between 2% to 6% higher) during the exacerbation state of disease, i.e. *Granulicatella*, *Haemophilus*, *Prevotella* and *Veillonella*, have been associated with gastrointestinal reflux disease (GERD) [75]. As a result of COPD patients having a common cough, GERD is associated with COPD and is considered a comorbidity [76]. In fact, GERD has been observed to be a predictor of exacerbations in COPD and implies that a higher prevalence of these bacteria could be used as a potential indicator of COPD exacerbations [76, 77].

In this study, bacterial alpha diversity and beta diversity analysis showed no difference between disease states. This observation is in agreement with previous COPD studies except for a study by Jubinville *et al.* (2018) who observed a difference in alpha diversity when comparing paired samples, i.e. the diversity in the paired samples differed across the disease state with most exacerbation samples showing a higher diversity [65-67, 69]. All these studies were conducted in Europe (the UK and Spain) or Northern America (Canada and USA) using sputum specimens, with most studies having less than 30 participants and most studies used the 454 sequencing. The only difference between these studies and the study by Jubinville *et al.* (2018) was the diversity measure used; most of the other studies used the Shannon index (often combined with Chao1 and Faith PD diversity measure), whereas Jubinville *et al.* (2018) used the Simpson index. Unlike, the Shannon index, the Simpson index is affected more by the relative abundances (i.e. evenness) of the species in a sample; this suggests that during the exacerbation state of disease, the abundances of species/OTUs changes but not the number of species/OTUs (richness) [78].

In this study, the most prevalent viral family was *Poxviridae* followed by *Siphoviridae* and *Myoviridae*. When compared to the only two other studies that have focused on the COPD lung virome, this study differed in the relative abundance of the key families [40, 41, 79]. The study by Garcia-Nunez *et al.* (2018) used sputum specimens (n=10) from paired stable and exacerbation patients (n=5) in Spain. The study by van Rijn *et al.* (2019) used nasopharyngeal swabs (n=88) collected from exacerbation patients between 2006 and 2010 and was conducted in Norway. The most prevalent viral families in these studies were *Anelloviridae* (negative sense DNA virus with no known pathogenicity in humans) and *Siphoviridae* (double-stranded DNA bacteriophages that have been found in the lung virome of cystic fibrosis (CF) patients as well as in the gastro-intestinal tract virome and the oral virome) [40, 41, 79-84]. These bacteriophages i.e. *Siphoviridae* and *Myoviridae* may act as reservoirs for antibiotic resistance

genes (contain antibiotic resistance genes in their genomes), mobile genetic elements and may contain virulence genes and other genes that affect bacterial metabolic pathways [35, 85].

A high abundance of *Poxviridae* was observed in this study, particularly the BeAn 58085 virus (BAV). *Poxviridae* is a family of complex, double-stranded DNA (dsDNA) viruses that are often zoonotic [86]. The most well-known virus from this family is the causative agent of smallpox (which has been eradicated), *Variola virus* and the clinical presentation of most human infections of this family is skin lesions [86]. Only two other virome studies, one that studied ascetic fluid in the human body (conducted in Spain) and one that studied ocular adnexa (conducted in Denmark on samples collected between 2005 and 2014) detected the BeAn 58058 virus in humans [87, 88]. This virus (BeAn 58058) was originally isolated from rodents (*Oryzomys* sp.) in Brazil in 1963 [89]. According to the viral-host database, the only known host for the BeAn 58058 virus is the *Oryzomys* sp., however, other *Poxviridae* have been known to infect a wide variety of hosts including humans [50]. The BeAn 58085 virus is considered a variant of the *Vaccinia virus* [90, 91]. The *Vaccinia virus* is a close relative of the smallpox virus that was used as a vaccine vector for smallpox until 1970 [90, 91]. There are three possible explanations for the high abundance of BeAn 58058 virus detected in this study. The first theory is that the BeAn 58058 virus is an ancient virus that over time has incorporated as part of the human genome; the theory is supported by i) a study by Mollerup *et al.* (2019) conducted on the virome of the ocular adnexa, which showed that viral reads (i.e. the BeAn 58058 virus) identified had high sequence homology to sequences of human origin, ii) a study that was conducted on the human genome (studying structural variants) identified the BeAn 58058 virus as part of the genome and iii) *Poxviridae* are dsDNA viruses and can easily integrate into the double-stranded human genome [92]. The second theory is that BeAn 58058 is a DNA artefact of the smallpox vaccine (which was a live attenuated vaccine) received years earlier; evidence supporting this theory includes the following: i) the study population in this study were all over the age of 50 years and would have received the smallpox vaccine before the vaccination programme for the smallpox virus was terminated in South Africa (in 1970) and ii) the *Vaccinia virus*, which was used for the smallpox vaccine showed high homology with the BeAn 58058 virus [90, 91, 93]. The third theory is that the participants in this study encountered an environmental exposure from which the virus was contracted, e.g. rats and its similarity to the *Cotia virus*, which can infect human cells [94]. The fourth theory is that the BeAn 58058 is a contaminant (i.e. a sequence not truly in the sample) from the extraction kit, from animal cells, reagents used or even from a previous sequencing run [95, 96]. Further analysis of the lung

virome, as well as the human genome of healthy individuals (i.e. not suffering from any lung disease) across different geographical regions and age groups, should provide insight into this in the future.

Although this study had a small population size, did not include healthy controls and did not have paired samples for the different disease states, this study provided a good pilot overview of the sputum microbiome and the sputum virome of the COPD lung in a South African setting. Additionally there was a skewed representation of the different disease states that could have impact the results resulting in either inflated or decreased relative abundance of the phyla and genera when comparing the disease states. Different respiratory samples can be used to study the lung microbiome, each with their own limitations. A sputum specimen was the specimen chosen for this study (instead of BAL, which has been used by most studies on the COPD microbiome) as it is the most patient-friendly method, i.e. is non-invasive [97]. The different specimens target different regions of the respiratory system with the sputum having a mixture of the microbiomes from both the upper respiratory tract and the lower respiratory tract [97-99]. Additionally, sputum specimens have higher bacterial loads (unlike BAL which has low bacterial loads and therefore less likely to magnify any contaminants) and are better for longitudinal studies (as these specimens are non-invasive) [98]. A limitation of this study was that a longitudinal study could not be completed due to time constraints (a longitudinal study was not possible within the three-year period of the PhD and is costly). The choice of specimen affects the diversity within a specimen and can result in distinct microbiomes [100, 101]. As only a single HIV participant could be recruited into this study, no comparison between HIV-positive individuals and HIV-negative individuals could be performed for the sputum microbiome in COPD patients; this requires further research. A diverse microbiome was observed in this study in both states of disease; with a predominated *Proteobacteria* population in the exacerbation state of disease. Conversely, the virome was dominated by a single virus, the BeAn 58058 virus. However, the origins of this virus and its possible clinical relevance is unknown. Future studies into the virome would require further investigation into this virus by studying the lung virome in healthy individuals and other lung diseases in the South African and international context.

4.5 Conclusions

This study is among the first to report lung microbiome composition in COPD patients from Africa. No statistically significant differences in the microbiome of COPD patients during the different states of disease were observed in this study. However, this study did note differences in the frequencies of key phyla and genera when compared to other studies from Europe and the USA. However, the reason for this differing microbial profile is unknown and warrants further research. In the virome, a high frequency of the BeAn 58058 virus was observed in the six samples; the explanation for this observation is unclear. To conclude, the sputum microbiome in South African COPD patients is diverse, regardless of the disease state, while the sputum virome warrants further research.

References

1. Terzikhan N, Verhamme KM, Hofman A, Stricker BH, Brusselle GG, Lahousse L: **Prevalence and incidence of COPD in smokers and non-smokers: the Rotterdam Study**. *Eur J Epidemiol* 2016, **31**(8):785-792.

2. Owuor N, Nalamala N, Gimenes JA, Jr., Sajjan US: **Rhinovirus and COPD airway epithelium**. *Pulm Crit Care Med* 2017, **2**(3).

3. Lopez-Campos JL, Tan W, Soriano JB: **Global burden of COPD**. *Respirology* 2016, **21**(1):14-23.

4. Vogelmeier CF, Criner GJ, Martinez FJ, Anzueto A, Barnes PJ, Bourbeau J, Celli BR, Chen R, Decramer M, Fabbri LM *et al*: **Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease 2017 report: GOLD executive summary**. *Am J Respir Crit Care Med* 2017, **195**(5):557-582.

5. Lee SW, Kuan CS, Wu LS, Weng JT: **Metagenome and metatranscriptome profiling of moderate and severe COPD sputum in Taiwanese Han males**. *PLoS One* 2016, **11**(7):e0159066.

6. Deslee G, Burgel PR, Escamilla R, Chanez P, Court-Fortune I, Nesme-Meyer P, Brinchault-Rabin G, Perez T, Jebrak G, Caillaud D *et al*: **Impact of current cough on health-**

related quality of life in patients with COPD. *Int J Chron Obstruct Pulmon Dis* 2016, **11**:2091-2097.

7. Buist AS, McBurnie MA, Vollmer WM, Gillespie S, Burney P, Mannino DM, Menezes AMB, Sullivan SD, Lee TA, Weiss KB *et al*: **International variation in the prevalence of COPD (The BOLD Study): a population-based prevalence study**. *The Lancet* 2007, **370**(9589):741-750.

8. Salvi S: **The silent epidemic of COPD in Africa**. *The Lancet Global Health* 2015, **3**(1):e6-e7.

9. Viviers PJ, Van Zyl-Smit RN: **Chronic obstructive pulmonary disease – diagnosis and classification of severity**. *S Afr Med J* 2015, **105**(9).

10. Abdool-Gaffar MS, Calligaro G, Wong ML, Smith C, Lalloo UG, Koegelenberg CFN, Dheda K, Allwood BW, Goolam-Mahomed A, van Zyl-Smit RN: **Management of chronic obstructive pulmonary disease-A position statement of the South African Thoracic Society: 2019 update**. *J Thorac Dis* 2019, **11**(11):4408-4427.

11. Miravitlles M, Anzueto A: **Antibiotic prophylaxis in COPD: Why, when, and for whom?** *Pulm Pharmacol Ther* 2015, **32**:119-123.

12. Pavord ID, Jones PW, Burgel PR, Rabe KF: **Exacerbations of COPD**. *Int J Chron Obstruct Pulmon Dis* 2016, **11 Spec Iss**:21-30.

13. Bouquet J, Tabor DE, Silver JS, Nair V, Tovchigrechko A, Griffin MP, Esser MT, Sellman BR, Jin H: **Microbial burden and viral exacerbations in a longitudinal multicenter COPD cohort**. *Respir Res* 2020, **21**(1):77.

14. Global Initiative for Chronic Obstructive Lung Disease: **Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease (2020 Report)**. 2020.

15. Aaron SD: **Management and prevention of exacerbations of COPD**. *BMJ* 2014, **349**:g5237.

16. Doring G, Parameswaran IG, Murphy TF: **Differential adaptation of microbial pathogens to airways of patients with cystic fibrosis and chronic obstructive pulmonary disease**. *FEMS Microbiol Rev* 2011, **35**(1):124-146.

17. D'Anna SE, Balbi B, Cappello F, Carone M, Di Stefano A: **Bacterial-viral load and the immune response in stable and exacerbated COPD: significance and therapeutic prospects**. *Int J Chron Obstruct Pulmon Dis* 2016, **11**:445-453.

18. Clooney AG, Fouhy F, Sleator RD, A OD, Stanton C, Cotter PD, Claesson MJ: **Comparing apples and oranges?: Next generation sequencing and its impact on microbiome analysis**. *PLoS One* 2016, **11**(2):e0148028.

19. Kulski JK: **Next-generation sequencing — An overview of the history, tools, and "omic" applications**. In: *Next generation sequencing - advances, applications and challenges.* 2016.

20. Park ST, Kim J: **Trends in next-generation sequencing and a new era for whole genome sequencing**. *Int Neurourol J* 2016, **20**(Suppl 2): S76-83.

21. Ji B, Nielsen J: **From next-generation sequencing to systematic modeling of the gut microbiome**. *Front Genet* 2015, **6**:219.

22. Ito T, Sekizuka T, Kishi N, Yamashita A, Kuroda M: **Conventional culture methods with commercially available media unveil the presence of novel culturable bacteria**. *Gut Microbes* 2019, **10**(1):77-91.

23. Helmy M, Awad M, Mosa KA: **Limited resources of genome sequencing in developing countries: Challenges and solutions**. *Appl Transl Genom* 2016, **9**:15-19.

24. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F: **Laboratory procedures to generate viral metagenomes**. *Nat Protoc* 2009, **4**(4):470-483.

25. Wommack KE, Bhavsar J, Polson SW, Chen J, Dumas M, Srinivasiah S, Furman M, Jamindar S, Nasko DJ: **VIROME: a standard operating procedure for analysis of viral metagenome sequences**. *Stand Genomic Sci* 2012, **6**(3):427-439.

26. Kembel SW, Wu M, Eisen JA, Green JL: **Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance**. *PLoS Comput Biol* 2012, **8**(10):e1002743.

27. Martin C, Burgel PR, Lepage P, Andrejak C, de Blic J, Bourdin A, Brouard J, Chanez P, Dalphin JC, Deslee G *et al*: **Host-microbe interactions in distal airways: relevance to chronic airway diseases**. *Eur Respir Rev* 2015, **24**(135):78-91.

28. Hiergeist A, Glasner J, Reischl U, Gessner A: **Analyses of intestinal microbiota: culture versus sequencing**. *ILAR J* 2015, **56**(2):228-240.

29. Woese CR, Fox GE: **Phylogenetic structure of the prokaryotic domain: the primary kingdoms**. *Proc Natl Acad Sci U S A* 1977, **74**(11):5088-5090.

30. Gürtler V, Subrahmanyam G, Shekar M, Maiti B, Karunasagar I: **Chapter 12- bacterial typing and identification by genomic analysis of 16S–23S rRNA intergenic transcribed spacer (ITS) sequences**. In: *Methods in Microbiology.* Edited by Michael Goodfellow, Iain Sutcliffe, Chun J, vol. 41: Academic Press; 2014: 253-274.

31. King AM, Adams MJ, Carstens EB, Lefkowitz EJ: **Virus taxonomy**. *Ninth report of the International Committee on Taxonomy of Viruses* 2012:486-487.

32. Williams SC: **The other microbiome**. *Proc Natl Acad Sci U S A* 2013, **110**(8):2682-2684.

33. Wylie KM: **The virome of the human respiratory tract**. *Clin Chest Med* 2017, **38**(1):11-19.

34. Bragg L, Tyson GW: **Metagenomics using next-generation sequencing**. *Methods Mol Biol* 2014, **1096**:183-201.

35. Wylie KM, Weinstock GM, Storch GA: **Emerging view of the human virome**. *Transl Res* 2012, **160**(4):283-290.

36. Amato KR: **An introduction to microbiome analysis for human biology applications**. *Am J Hum Biol* 2017, **29**(1).

37. Cabrera-Rubio R, Garcia-Nunez M, Seto L, Anto JM, Moya A, Monso E, Mira A: **Microbiome diversity in the bronchial tracts of patients with chronic obstructive pulmonary disease**. *J Clin Microbiol* 2012, **50**(11):3562-3568.

38. Dickson RP, Martinez FJ, Huffnagle GB: **The role of the microbiome in exacerbations of chronic lung diseases**. *The Lancet* 2014, **384**(9944):691-702.

39. Sze MA, Dimitriu PA, Suzuki M, McDonough JE, Campbell JD, Brothers JF, Erb-Downward JR, Huffnagle GB, Hayashi S, Elliott WM *et al*: **Host response to the lung microbiome in chronic obstructive pulmonary disease**. *Am J Respir Crit Care Med* 2015, **192**(4):438-445.

40. Garcia-Nunez M, Gallego M, Monton C, Millares L, Pomares X, Monso E, Capilla S, Espasa M, Ferrari R, Moya A *et al*: **The respiratory virome in chronic obstructive pulmonary disease**. *Future Virol* 2018, **13**(7):457-466.

41. van Rijn AL, van Boheemen S, Sidorov I, Carbo EC, Pappas N, Mei H, Feltkamp M, Aanerud M, Bakke P, Claas ECJ *et al*: **The respiratory virome and exacerbations in patients with chronic obstructive pulmonary disease**. *PLoS One* 2019, **14**(10):e0223952.

42. Hamid Q, Kelly MM, Linden M, Louis R, Pizzichini MMM, Pizzichini E, Ronchi C, Van Overveld F, Djukanovic R: **Methods of sputum processing for cell counts, immunocytochemistry and *in situ* hybridisation**. *Eur Respir J* 2002, **20**(Supplement 37):19S-23S.

43. Terranova L, Oriano M, Teri A, Ruggiero L, Tafuro C, Marchisio P, Gramegna A, Contarini M, Franceschi E, Sottotetti S *et al*: **How to process sputum samples and extract bacterial DNA for microbiota analysis**. *Int J Mol Sci* 2018, **19**(10):3256-3568.

44. Stokell JR, Khan A, Steck TR: **Mechanical homogenization increases bacterial homogeneity in sputum**. *J Clin Microbiol* 2014, **52**(7):2340-2345.

45. de la Cruz Pena MJ, Martinez-Hernandez F, Garcia-Heredia I, Lluesma Gomez M, Fornas O, Martinez-Garcia M: **Deciphering the human virome with single-virus genomics and metagenomics**. *Viruses* 2018, **10**(3).

46. Allander T, Tammi MT, Eriksson M, Bjerkner A, Tiveljung-Lindell A, Andersson B: **Cloning of a human parvovirus by molecular screening of respiratory tract samples**. *Proc Natl Acad Sci U S A* 2005, **102**(36):12891-12896.

47. Wood DE, Lu J, Langmead B: **Improved metagenomic analysis with Kraken 2**. *Genome Biol* 2019, **20**(1):257.

48. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Gruning BA *et al*: **The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update**. *Nucleic Acids Res* 2018, **46**(W1): W537-W544.

49. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F *et al*: **Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2**. *Nature Biotechnology* 2019, **37**(8):852-857.

50. Mihara T, Nishimura Y, Shimizu Y, Nishiyama H, Yoshikawa G, Uehara H, Hingamp P, Goto S, Ogata H: **Linking virus genomes with host taxonomy**. *Viruses* 2016, **8**(3):66.

51. Simbayi L, Zuma K, Zungu N, Moyo S, Marinda E, Jooste S, Mabaso M, Ramlagan S, North A, Van Zyl J: **South African National HIV prevalence, incidence, behaviour and communication survey, 2017: towards achieving the UNAIDS 90-90-90 targets**. 2019.

52. Johnson LF, Mossong J, Dorrington RE, Schomaker M, Hoffmann CJ, Keiser O, Fox MP, Wood R, Prozesky H, Giddy J *et al*: **Life expectancies of South African adults starting antiretroviral treatment: collaborative analysis of cohort studies**. *PLoS Med* 2013, **10**(4):e1001418.

53. Butler I, MacLeod W, Majuba PP, Tipping B: **Human immunodeficiency virus infection and older adults: A retrospective single-site cohort study from Johannesburg, South Africa**. *South Afr J HIV Med* 2018, **19**(1):838.

54. Cornell M, Johnson LF, Schomaker M, Tanser F, Maskew M, Wood R, Prozesky H, Giddy J, Stinson K, Egger M *et al*: **Age in antiretroviral therapy programmes in South Africa: a retrospective, multicentre, observational cohort study**. *The Lancet HIV* 2015, **2**(9):e368-e375.

55. Barnes PJ: **Inflammatory mechanisms in patients with chronic obstructive pulmonary disease**. *J Allergy Clin Immunol* 2016, **138**(1):16-27.

56. Gut-Gobert C, Cavailles A, Dixmier A, Guillot S, Jouneau S, Leroyer C, Marchand-Adam S, Marquette D, Meurice JC, Desvigne N *et al*: **Women and COPD: do we need more evidence?** *Eur Respir Rev* 2019, **28**(151).

57. Chapman KR, Tashkin DP, Pye DJ: **Gender bias in the diagnosis of COPD**. *Chest* 2001, **119**(6):1691-1695.

58. Ancochea J, Miravitlles M, García-Río F, Muñoz L, Sánchez G, Sobradillo V, Duran-Tauleria E, Soriano JB: **Underdiagnosis of chronic obstructive pulmonary disease in women: quantification of the problem, determinants and proposed actions**. *Archivos de Bronconeumología (English Edition)* 2013, **49**(6):223-229.

59. Invernizzi R, Lloyd CM, Molyneaux PL: **Respiratory microbiome and epithelial interactions shape immunity in the lungs**. *Immunology* 2020, **160**(2):171-182.

60. Fabbrizzi A, Amedei A, Lavorini F, Renda T, Fontana G: **The lung microbiome: clinical and therapeutic implications**. *Intern Emerg Med* 2019, **14**(8):1241-1250.

61. Segal LN, Alekseyenko AV, Clemente JC, Kulkarni R, Wu B, Chen H, Berger KI, Goldring RM, Rom WN, Blaser MJ *et al*: **Enrichment of lung microbiome with supraglottic taxa is associated with increased pulmonary inflammation**. *Microbiome* 2013, **1**(1):19.

62. Hilty M, Burke C, Pedro H, Cardenas P, Bush A, Bossley C, Davies J, Ervine A, Poulter L, Pachter L *et al*: **Disordered microbial communities in asthmatic airways**. *PLoS One* 2010, **5**(1):e8578.

63. Huffnagle GB, Dickson RP, Lukacs NW: **The respiratory tract microbiome and lung inflammation: a two-way street**. *Mucosal Immunol* 2017, **10**(2):299-306.

64. Haldar K, George L, Wang Z, Mistry V, Ramsheh MY, Free RC, John C, Reeve NF, Miller BE, Tal-Singer R *et al*: **The sputum microbiome is distinct between COPD and health, independent of smoking history**. *Respir Res* 2020, **21**(1):183.

65. Jubinville E, Veillette M, Milot J, Maltais F, Comeau AM, Levesque RC, Duchaine C: **Exacerbation induces a microbiota shift in sputa of COPD patients**. *PLoS One* 2018, **13**(3):e0194355.

66. Millares L, Perez-Brocal V, Ferrari R, Gallego M, Pomares X, Garcia-Nunez M, Monton C, Capilla S, Monso E, Moya A: **Functional metagenomics of the bronchial microbiome in COPD**. *PLoS One* 2015, **10**(12):e0144448.

67. Wang Z, Singh R, Miller BE, Tal-Singer R, Van Horn S, Tomsho L, Mackay A, Allinson JP, Webb AJ, Brookes AJ *et al*: **Sputum microbiome temporal variability and dysbiosis in chronic obstructive pulmonary disease exacerbations: an analysis of the COPDMAP study**. *Thorax* 2018, **73**(4):331-338.

68. Wang Z, Bafadhel M, Haldar K, Spivak A, Mayhew D, Miller BE, Tal-Singer R, Johnston SL, Ramsheh MY, Barer MR *et al*: **Lung microbiome dynamics in COPD exacerbations**. *Eur Respir J* 2016, **47**(4):1082-1092.

69. Huang YJ, Sethi S, Murphy T, Nariya S, Boushey HA, Lynch SV: **Airway microbiome dynamics in exacerbations of chronic obstructive pulmonary disease**. *J Clin Microbiol* 2014, **52**(8):2813-2823.

70. Wang J, Chai J, Sun L, Zhao J, Chang C: **The sputum microbiome associated with different sub-types of AECOPD in a Chinese cohort**. *BMC Infect Dis* 2020, **20**(1):610.

71. Ubags NDJ, Marsland BJ: **Mechanistic insight into the function of the microbiome in lung diseases**. *Eur Respir J* 2017, **50**(3):1602467-1602489.

72. Faner R, Sibila O, Agusti A, Bernasconi E, Chalmers JD, Huffnagle GB, Manichanh C, Molyneaux PL, Paredes R, Perez Brocal V *et al*: **The microbiome in respiratory medicine: current challenges and future perspectives**. *Eur Respir J* 2017, **49**(4).

73. Rylance J, Kankwatira A, Nelson DE, Toh E, Day RB, Lin H, Gao X, Dong Q, Sodergren E, Weinstock GM *et al*: **Household air pollution and the lung microbiome of healthy adults in Malawi: a cross-sectional study**. *BMC Microbiol* 2016, **16**(1):182.

74. Kumpitsch C, Koskinen K, Schopf V, Moissl-Eichinger C: **The microbiome of the upper respiratory tract in health and disease**. *BMC Biol* 2019, **17**(1):87.

75. Park CH, Seo SI, Kim JS, Kang SH, Kim BJ, Choi YJ, Byun HJ, Yoon JH, Lee SK: **Treatment of non-erosive reflux disease and dynamics of the esophageal microbiome: a prospective multicenter study**. *Sci Rep* 2020, **10**(1):15154.

76. Lee AL, Goldstein RS: **Gastroesophageal reflux disease in COPD: links and risks**. *Int J Chron Obstruct Pulmon Dis* 2015, **10**:1935-1949.

77. Sanchez J, Schumann DM, Karakioulaki M, Papakonstantinou E, Rassouli F, Frasnelli M, Brutsche M, Tamm M, Stolz D: **Laryngopharyngeal reflux in chronic obstructive pulmonary disease - a multi-centre study**. *Respir Res* 2020, **21**(1):220.

78. Johnson KV, Burnet PW: **Microbiome: Should we diversify from diversity?** *Gut Microbes* 2016, **7**(6):455-458.

79. Mitchell AB, Oliver BG, Glanville AR: **Translational aspects of the human respiratory virome**. *Am J Respir Crit Care Med* 2016, **194**(12):1458-1464.

80. Zarate S, Taboada B, Yocupicio-Monroy M, Arias CF: **Human virome**. *Arch Med Res* 2017, **48**(8):701-716.

81. Freer G, Maggi F, Pifferi M, Di Cicco ME, Peroni DG, Pistello M: **The virome and its major component, *Anellovirus*, a convoluted system molding human immune defenses and possibly affecting the development of asthma and respiratory diseases in childhood**. *Front Microbiol* 2018, **9**:686.

82. Simmonds P, Sharp CP: ***Anelloviridae***. In: *Clinical Virology*. 2016: 701-711.

83. Fermin G: **Virion structure, genome organization, and taxonomy of viruses**. In: *Viruses*. 2018: 17-54.

84. Malathi VG, Renuka Devi P: **ssDNA viruses: key players in global virome**. *Virus disease* 2019, **30**(1):3-12.

85. Keen EC, Dantas G: **Close encounters of three kinds: bacteriophages, commensal bacteria, and host immunity**. *Trends Microbiol* 2018, **26**(11):943-954.

86. Isaacs SN, Buller RM: ***Poxviruses***. In: *Clinical Virology*. 2016: 385-413.

87. Mollerup S, Mikkelsen LH, Hansen AJ, Heegaard S: **High-throughput sequencing reveals no viral pathogens in eight cases of ocular adnexal extranodal marginal zone B-cell lymphoma**. *Exp Eye Res* 2019, **185**:107677.

88. Blanco-Picazo P, Fernandez-Orth D, Brown-Jaque M, Miro E, Espinal P, Rodriguez-Rubio L, Muniesa M, Navarro F: **Unravelling the consequences of the bacteriophages in human samples**. *Sci Rep* 2020, **10**(1):6737.

89. Wanzeller AL, Souza AL, Azevedo RS, Junior EC, Filho LC, Oliveira RS, Lemos PS, Junior JV, Vasconcelos PF: **Complete genome sequence of the BeAn 58058 virus isolated from *Oryzomys* sp. rodents in the Amazon region of Brazil**. *Genome Announc* 2017, **5**(9).

90. Marques JT, Trindade GD, Da Fonseca FG, Dos Santos JR, Bonjardim CA, Ferreira PC, Kroon EG: **Characterization of ATI, TK and IFN-alpha/betaR genes in the genome of the BeAn 58058 virus, a naturally attenuated wild *Orthopoxvirus***. *Virus Genes* 2001, **23**(3):291-301.

91. Silva DCM, Moreira-Silva EAdS, Gomes JdAS, Fonseca FGd, Correa-Oliveira R: **Clinical signs, diagnosis, and case reports of *vaccinia virus* infections**. *Braz J Infect Dis* 2010, **14**:129-134.

92. Oliveira GP, Rodrigues RAL, Lima MT, Drumond BP, Abrahao JS: ***Poxvirus* host range genes and virus-host spectrum: A critical review**. *Viruses* 2017, **9**(11).

93. Abrahao JS, Trindade Gde S, Ferreira JM, Campos RK, Bonjardim CA, Ferreira PC, Kroon EG: **Long-lasting stability of *vaccinia virus* strains in murine feces: implications for virus circulation and environmental maintenance**. *Arch Virol* 2009, **154**(9):1551-1553.

94. Haller SL, Peng C, McFadden G, Rothenburg S: ***Poxviruses* and the evolution of host range and virulence**. *Infect Genet Evol* 2014, **21**:15-40.

95. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ: **Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data**. *Microbiome* 2018, **6**(1):226.

96. Marti JM: **Recentrifuge: Robust comparative analysis and contamination removal for metagenomics**. *PLoS Comput Biol* 2019, **15**(4):e1006967.

97. Ditz B, Christenson S, Rossen J, Brightling C, Kerstjens HAM, van den Berge M, Faiz A: **Sputum microbiome profiling in COPD: beyond singular pathogen detection**. *Thorax* 2020, **75**(4):338-344.

98. Carney SM, Clemente JC, Cox MJ, Dickson RP, Huang YJ, Kitsios GD, Kloepfer KM, Leung JM, LeVan TD, Molyneaux PL *et al*: **Methods in lung microbiome research**. *Am J Respir Cell Mol Biol* 2020, **62**(3):283-299.

99. Sulaiman I, Schuster S, Segal LN: **Perspectives in lung microbiome research**. *Curr Opin Microbiol* 2020, **56**:24-29.

100. Hogan DA, Willger SD, Dolben EL, Hampton TH, Stanton BA, Morrison HG, Sogin ML, Czum J, Ashare A: **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**(3):e0149998.

101. Chang, Dela Cruz CS, Sharma L: **Challenges in understanding lung microbiome: It is NOT like the gut microbiome**. *Respirology* 2020, **25**(3):244-245.

CHAPTER 5

Comparison of targeted metagenomics and the IS-Pro method for analysing the lung microbiome

The editorial style of the Microbiome Journal was followed in this chapter

Abstract

Background

Targeted metagenomics and the IS-Pro (intergenic spacer profiling) method are two of the many methods that have been used to study the microbiome. Targeted metagenomics targets the hypervariable regions of the 16S rRNA gene and the IS-Pro method targets the intergenic spacer regions between the 16S rRNA and 23S rRNA gene regions. The aim of this study was to compare targeted metagenomics and IS-Pro methods for the ability to discern the microbial composition of the lung microbiome of COPD patients.

Methods

Spontaneously expectorated sputum specimens were collected from COPD patients in the Tshwane Health District, South Africa. Bacterial DNA was extracted from the specimens using Isolate II Genomic DNA kit and aliquoted. One aliquot was used for targeted metagenomics using V1-V3 primers of the 16S rRNA gene on the MiSeq platform and a second aliquot for the IS-Pro method according to the manufacturer's instructions. The analysis was performed using the QIIME2 bioinformatics pipeline and the commercial IS-Pro software for targeted metagenomics and the IS-Pro method, respectively. Additionally, a laboratory cost per isolate and time analysis was performed for each method.

Results

Statistically significant differences were observed in alpha diversity when targeted metagenomics and IS-Pro methods' data were compared using the Shannon diversity measure [median values of 2.732 and 2.183, interquartile range (IQR) values of 0.09 and 0.44, p-value=0.0006] but not with the Simpson diversity measure (median values of 0.866 and 0.851, IQR values of 0.13 and 0.06, p-value=0.84). Distinct clusters with no overlap between the two technologies were observed using PCoA plots and the Jaccard diversity measure for beta diversity. At a phylum level targeted metagenomics had a lower relative abundance of the

Proteobacteria (16% vs 38%), *Bacteroidetes* (10.27% vs 12.4%) and *Fusobacteria* (2.3% vs 6.6%) and higher relative abundance of *Actinobacteria* (12.3% vs 2.45%) and *Firmicutes* (57% vs 40.5%) when compared to the IS-Pro method. At a genus level, *Haemophilus*, *Prevotella* and *Streptococcus* were the most prevalent and were observed in similar abundances for both methods. Targeted metagenomics was only able to classify 23% (144/631) of all OTUs to a species level, compared to the IS-Pro method, which was able to classify 86% (55/64) of all OTUs to a species level. However, the unclassified OTUs accounted for a higher relative abundance when using the IS-Pro method (35%) compared to targeted metagenomics (5%). These unclassified OTUs from the IS-Pro method could be classified at the phylum level, with *Proteobacteria* (20%) accounting for the most unclassified sequences. The two methods performed comparably in terms of time; however, the IS-Pro method was more user-friendly.

Conclusions

It is essential to understand the value of different methods for characterisation of the microbiome. Targeted metagenomics and IS-Pro methods showed differences in their abilities to identify and characterise OTUs, in the diversity and microbial composition of the lung microbiome. The IS-Pro method might miss relevant species and could inflate the abundance of members of the *Proteobacteria*. However, the IS-Pro kit was able to identify most of the important lung pathogens, such as *Burkholderia* and *Pseudomonas* and may work well in a more diagnostics-orientated setting. Both methods were comparable in terms of cost and time; however, the IS-Pro method was easier to use.

5.1 Background

Microorganisms occur as communities and can play an important role in host metabolism [1-3]. This collective of microorganisms within a community (ecosystem) and their genetic material is referred to as a microbiome [4, 5]. Previously, culture-dependent techniques were used to study the microbiome, however, researchers have found that less than 1% of all bacteria can be cultured and that the microbiome is often more diverse than culture methods suggest [4, 6]. Culture-independent methods, such as denaturing gradient gel electrophoresis, fluorescence *in situ* hybridisation, microarrays, quantitative polymerase chain reaction and terminal length polymorphisms have since been used to study the microbiome [7-11]. However, the most popular approach to study the microbiome is sequencing analysis, either using Sanger or next-generation sequencing (NGS) technologies using a targeted approach [8-10, 12].

The most popular target of these sequencing methods is the 16S rRNA gene region [13, 14]. The 16S rRNA gene is useful for identifying bacteria and determining phylogenetics as this gene is present in all prokaryotes, i.e. it is universal, is easily isolated and is highly conserved (i.e. the sequences and the length of the genes change very little with time) [9, 15, 16]. Additionally, this 16S rRNA gene codes for part of the ribosome; in bacteria (and archaea) the 70S ribosome, is divided into two components: the 30S subunit and the 50S subunit [17]. The 30S subunit includes the 16S rRNA sequence (the Shine-Dalgarno sequence, required for protein translation, is complementary to 3' end of 16S rRNA) and proteins, whereas the 50S subunit includes the 23S rRNA and 5S rRNA [17-19]. The 16S ribosomal subunit consists of both hypervariable and conserved regions, with the sequencing primers that are commonly used targeting the conserved regions between the hypervariable regions [18, 20]. There are nine hypervariable (V1-V9) regions and nine conserved regions (which alternate) [20, 21]. Among the most common primers used for 16S rRNA gene are the 27F and 518R primers that cover the V1 to V3 hypervariable regions [22, 23]. This region, i.e. V1-V3 region of the 16S rRNA was shown to have the highest similarity with full-length sequences of the 16S rRNA gene [24].

The IS-Pro (intergenic spacer profiling) method, a targeted metagenomics method that targets the intergenic spacer (IS) region between the 16S rRNA and 23S rRNA was developed by Budding and colleagues in 2010 to identify all bacteria present in the sample, i.e. a clinical specimen. The intergenic spacer region was chosen due to its variability; this region is more variable than the hypervariable regions of the 16S rRNA [25, 26]. The IS region has species-specific differences in length and sequence polymorphisms, which are used to identify bacteria

and can be termed a profiling method [25, 26]. This method has been used to study the vaginal microbiome, the gastrointestinal tract microbiome and has been tested in a clinical setting (clinical microbiology laboratory) for the identification of bacteria from "sterile" body sites/fluids [25-44].

Studies that have investigated the lung microbiome have mostly used targeted metagenomics. To our knowledge, no studies have used the IS-Pro method to study the lung microbiome. The aim of this study was to compare the IS-Pro method to 16S rRNA sequencing in its ability to discern the microbial composition of the lung microbiome of COPD patients.

5.2 Methods

5.2.1 Study design and study participants

Patients suffering from COPD that were admitted or were attending a clinic at one of three hospitals (one academic, one district and one private) in the Tshwane Health District were invited to participate in the study. If the inclusion and exclusion criteria were met and written informed consent was obtained, participants were included in the study (Table 4.1). Ethical approval was granted from The Research Ethics committee, Faculty of Health Sciences, University of Pretoria (REC no: 237/2017). All aspects of the research were conducted by the candidate unless otherwise stated.

5.2.2 Sputum specimen processing and bacterial DNA extraction

Spontaneously expectorated sputum specimens were collected from all participants at a single time point. The specimens were transported on ice without any preservation media and stored at -80°C (Innova U535 Upright, Eppendorf, Germany) until batch processing could occur. Each sputum specimen was thawed (after all specimens were collected) and treated with an equal volume of 0.1% dithiothreitol (DTT) (Roche. Switzerland) (to reduce sputum viscosity) and were homogenised for 30 seconds (sec) (Vortex-Genie®2; Scientific Industries Inc., USA) [45-48]. An aliquot of the homogenised sputum (250 µL) was transferred to a new 2 mL microcentrifuge tube (Axygen, Corning, Germany) and centrifuged at 4 000 x *g* (Spectrafuge™ 24D, Labnet International Inc., USA) for 30 min before extraction. The pellet was used for extraction and bacterial DNA extraction was performed using the Isolate II Genomic DNA Kit (Bioline, UK). The manufacturer's instructions were followed with the addition of 10 mg/mL lysozyme (Sigma-Aldrich, USA), 3 U/µL lysostaphin (Sigma-Aldrich, USA) and 6.75 µL of 10 U/µL mutanolysin (Sigma-Aldrich, USA) to the hard-to-lyse buffer [20 mM Tris (Sigma-

Aldrich, USA) pH 8.0; 1% Triton X-100 (Amresco, USA); 2 mM EDTA(Sigma-Aldrich, USA)]. The extracted DNA was separated into three aliquots [in two new 2 mL microcentrifuge tubes (Axygen, Corning, Germany)] and stored at -20°C (Samsung, South Korea) until further usage. Aliquot 1 was used for targeted metagenomics and aliquot 2 was used for the IS-Pro method. The DNA concentration and purity were measured using the Genova Nano spectrophotometer (Jenway, UK).

5.2.3 Targeted metagenomics

Aliquot 1 of the extracted bacterial DNA (section 5.2.2) was sent to Inqaba Biotechnical Industries (Pretoria, South Africa), a commercial NGS service provider, for sequencing. Briefly, bacterial DNA was amplified using a PCR targeting the V1-V3 region of the 16S rRNA gene using the 27F and 518R primers [49]. The amplicons generated from the PCR assay were gel purified, end-repaired (removal of 3' overhangs) and the Illumina-specific adapter sequences were ligated to each amplicon using the NEBNext® Ultra™ II DNA library prep kit for Illumina® (New England Biolabs, USA) according to the manufacturer's instructions. After ligation (and quantification) the samples were indexed using the NEBNext® Multiplex Oligos for Illumina® (Index Primers Set 1) (New England Biolabs, USA), followed by purification with AMPure XP beads (Beckman Coulter, USA). The purified amplicons were sequenced using the MiSeq v3 platform (Illumina, USA) for 600 cycles. Each sample generated 300 bp paired-end reads. The resulting fastq files underwent quality control (QC) and were analysed using QIIME2 and the Greengenes database (13.8) [50].

5.2.4 The IS-Pro method to determine the microbiome

The IS-Pro kit (InBiome, the Netherlands) was used to amplify the previously extracted bacterial DNA (section 5.2.2; aliquot 2), according to the manufacturer's instructions and was performed at Synexa Life Sciences, Cape Town, South Africa. The kit components included two master mixes (PROTEO and FIRBAC), two control vials (one for *Proteobacteria* and one for *Firmicutes*/*Bacteroidetes*) and eMix (reference marker). The PROTEO master mix targets only the *Proteobacteria*, whereas the FIRBAC master mix targets the *Actinobacteria*, *Bacteroidetes*, *Firmicutes*, *Fusobacteria* and *Verrucomicrobia* phyla [26]. In a microtiter plate, for each sample (n=24), the positive control (included in the kit), the negative control [nuclease-free water, (Qiagen, Germany)] and the following was added: 12 µL of PROTEO master mix (supplied with kit) in a well and 12 µL of FIRBAC master mix (supplied with kit) to a separate well. To each well, 8 µL of extracted bacterial DNA was added. The PCR amplification

(Applied Biosystems GeneAmp PCR 9700, ThermoFisher Scientific, USA) conditions were as follows: 35 cycles of 94°C for 30 s, 56°C for 45 s, and 72°C for 1 min, followed by a final extension step at 72°C for 10 min. After amplification, 16 µL of the eMix was added to each well (for the number of samples and controls) in a new microtiter plate and 4 µL of each amplicon was added to a well, followed by denaturation at 94°C for 3 min. The samples were analysed on the Applied Biosystems 3730xL genetic analyser (ThermoFisher Scientific, USA) at the Central Analytical Facility (CAF) at Stellenbosch University, Cape Town, South Africa.

Data were analysed using the IS-Pro software suite (InBiome, The Netherlands), which generates microbial profiles. The colour of the peak generated is obtained from the labelled primers and provides information of which phyla had been amplified, whereas the length of the fragment obtained is used to identify the bacteria to lower taxonomic levels (genus, species or subspecies). Each peak within a profile is considered an operational taxonomic unit (OTU) and its intensity determined the abundance.

5.2.5 Statistical analysis and data visualisation

The program used for 16S rRNA analysis, QIIME2, generated the taxonomy table and OTU table in .qza format. The .qza files were converted into the correct format for phyloseq using the R package, QIIME2R. The IS-Pro data were converted to the required format for phyloseq manually (in Excel). Phyloseq requires two files to process and analyse data: i) a file containing the taxonomic information (of the microorganisms) and ii) a file containing the read/OTU counts present in each sample. The IS-Pro output was a single file that contained both taxonomic and OTU counts and therefore needed to be separated; as such the taxonomic data and the OTU counts were moved into two different files, which were used as the taxonomy table and OTU table respectively. The data were analysed in R using the following packages: i) phyloseq [alpha diversity, beta diversity, statistical tests, principal coordinate analysis (PCoA), and relative abundance of the taxa], ii) ggplot2 (for the plotting of all graphs) and iii) DESeq2 (to determine if there was a log2fold difference).

5.2.6 Cost per isolate and time analysis

Targeted metagenomics was compared to the IS-Pro method in terms of cost, time to analysis and user-friendliness. The cost calculated included estimates based on the procurement of resources in our laboratory at the Department of Medical Microbiology of the University of Pretoria, the cost for sample processing, DNA extraction, reagents for PCR assays and PCR

clean-up, consumables and the complete cost of sequencing (based on the quote generated by the company that performed sequencing and includes both labour cost and the benchtop cost). Time to analysis was calculated from the date of sequencing results were received up until the analyses were completed (including statistical analysis). The user-friendliness was determined based on the authors' experience with QIIME2 and the IS-Pro proprietary software.

5.3 Results

5.3.1 Patient demographics

A total of 80 participants were planned to be included in the study, however due to the strict inclusion and exclusion criteria as well as the limited number of patients attending the clinic or being admitted to the hospital, this number could not be realised. A total of 24 participants were enrolled in the study; 18 males and six females aged from 50 years old to 82 years old (median= 60 years old with a standard deviation of 7.34). Only one of the participants was HIV-positive. Participants were distributed across the three hospitals as follows: i) Hospital A (Tertiary Academic Hospital): 16 participants, ii) Hospital B (District Hospital): one participant and Hospital C (Private Hospital): seven participants. Eighteen of the participants were in the stable state of disease at the time of sampling and six of the participants were in the exacerbation state of disease at the time of sampling. Four of the participants had never smoked, nine of the participants were current smokers and 11 participants had stopped smoking.

5.3.2 Alpha and beta diversity analysis

One of the 24 samples was excluded from subsequent analysis as the sample did not meet the quality control requirements with the IS-Pro method; the concentration of the internal size marker was too low. This sample had generated data using targeted metagenomics. When alpha diversity was compared between targeted metagenomics and IS-Pro methods (Figure 5.1), a significant difference was observed using the Shannon diversity measure (using Wilcoxon sum rank test, p-value=0.0006, median values of 2.732 and 2.183); targeted metagenomics showed a higher alpha diversity than the IS-Pro method. No difference was observed with the Simpson diversity measure when comparing targeted metagenomics and IS-Pro methods (using Wilcoxon sum rank test, p-value=0.84, median values of 0.866 and 0.851).

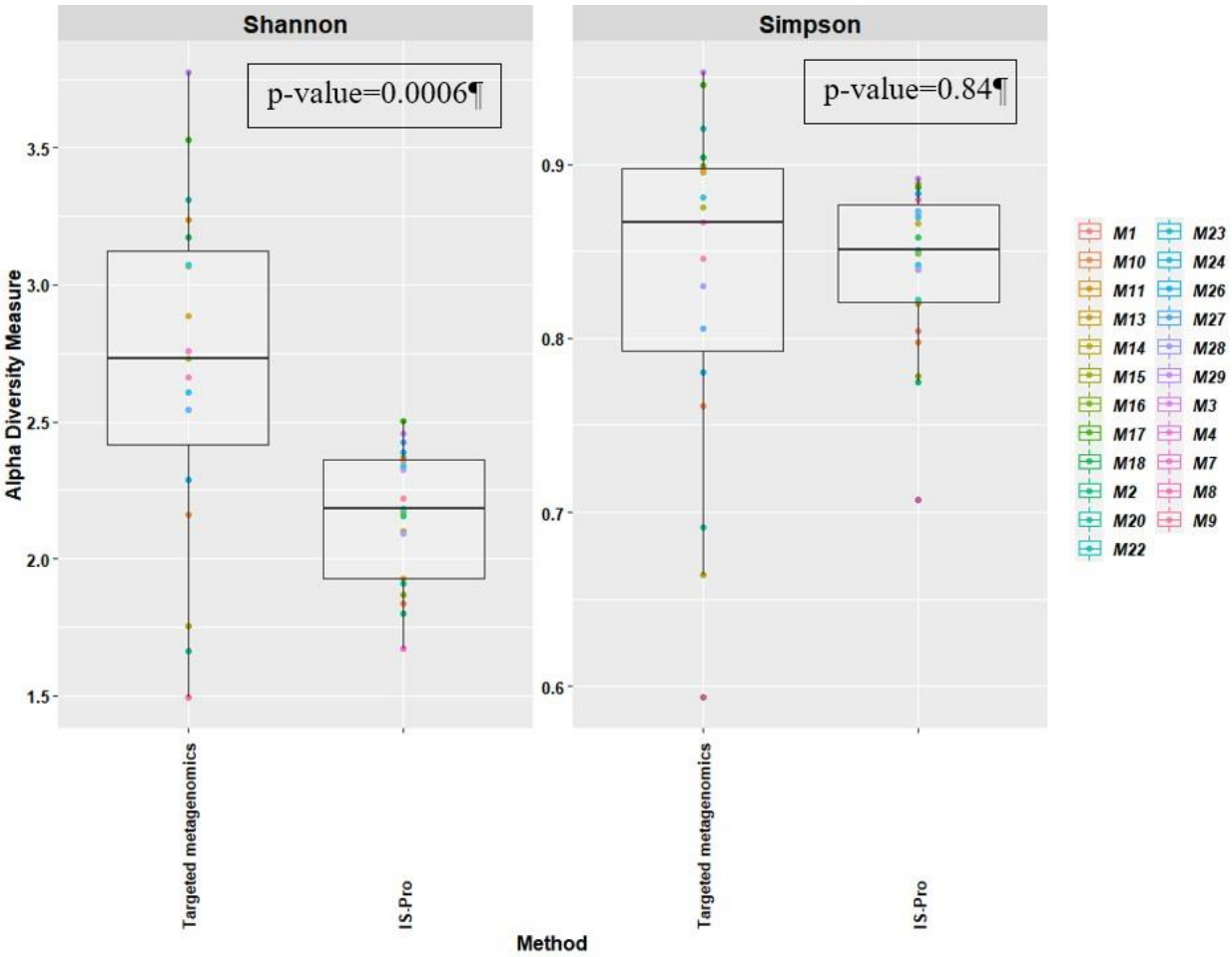

Figure 5.1: The alpha diversity boxplot of the sputum microbiome of COPD participants comparing the targeted metagenomics and IS-Pro methods (n=23) for Shannon and Simpson diversity measures. Each dot on the graph represents a sample. The boxes represent the interquartile range (IQR) and the horizontal line represents the median. The median values for the Shannon diversity measure were as follows: i) targeted metagenomics=2.732 and ii) IS-Pro method=2.183. The median values for the Simpson diversity measures were as follows: i) targeted metagenomics=0.866 and ii) IS-Pro method=0.851. The IQR values for the Shannon diversity measure were as follows: i) targeted metagenomics =0.09 and ii) IS-Pro method =0.44. The IQR values for the Simpson diversity measure were as follows: i) targeted metagenomics =0.13 and ii) IS-Pro method =0.06.

Beta diversity analysis (PCoA analysis) of the two methods (between the targeted metagenomics and IS-Pro methods) showed the isolates clustering according to the method (Figure 5.2). Both the Jaccard diversity and the Morisita Horn (not shown) measures showed the two methods forming distinct clusters with no overlap between the two methods. The targeted metagenomics isolates clustered further apart than the IS-Pro method isolates.

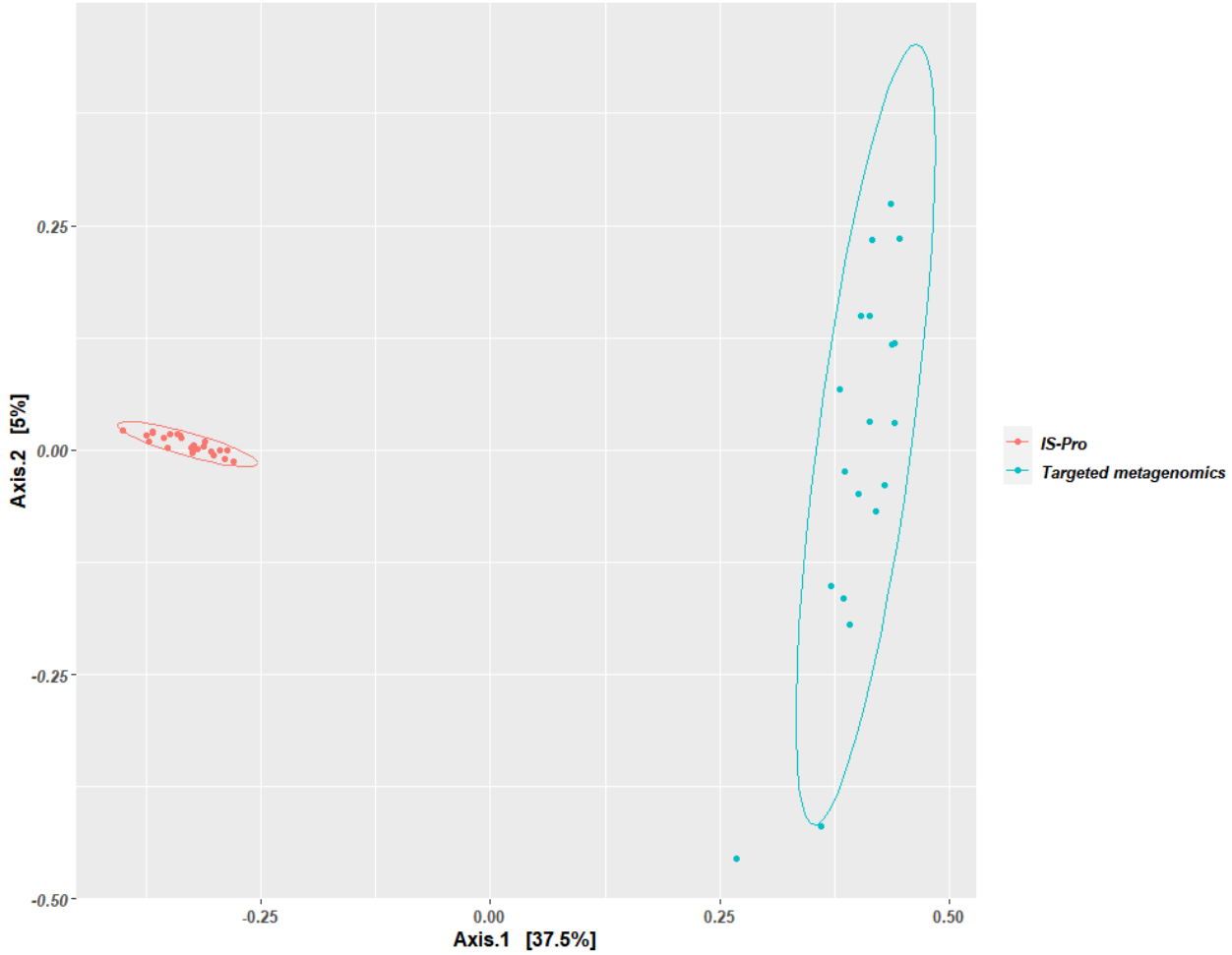

Figure 5.2: Principal coordinate analysis (PCoA) plot derived using the Jaccard diversity measure of the sputum microbiome of COPD participants. The PCoA plot compares the targeted metagenomics and IS-Pro methods; with the dots representing each sample.

5.3.3 Difference in relative abundance between targeted metagenomics and IS-Pro methods

The most prevalent phyla according to both methods were *Firmicutes* (57.1% for the targeted metagenomics and 40.5% for the IS-Pro method), *Proteobacteria* (16% for the targeted metagenomics and 38% for the IS-Pro method), *Bacteroidetes* (10.3% for the targeted metagenomics and 12.4% for the IS-Pro method), *Actinobacteria* (12.3% for the targeted

metagenomics and 2.5% for the IS-Pro method) and *Fusobacteria* (2.3% for the targeted metagenomics and 6.6% for the IS-Pro method) (Figure 5.3). The IS-Pro method, however, showed a higher relative abundance of the *Proteobacteria*, *Bacteroidetes* and *Fusobacteria* and lower relative abundance of *Actinobacteria*, and *Firmicutes*. At a sample level (with sample 29), the trend observed was similar except *Bacteroidetes* had a lower relative abundance and *Firmicutes* had a higher relative for the IS-Pro method (Figure 5.4).

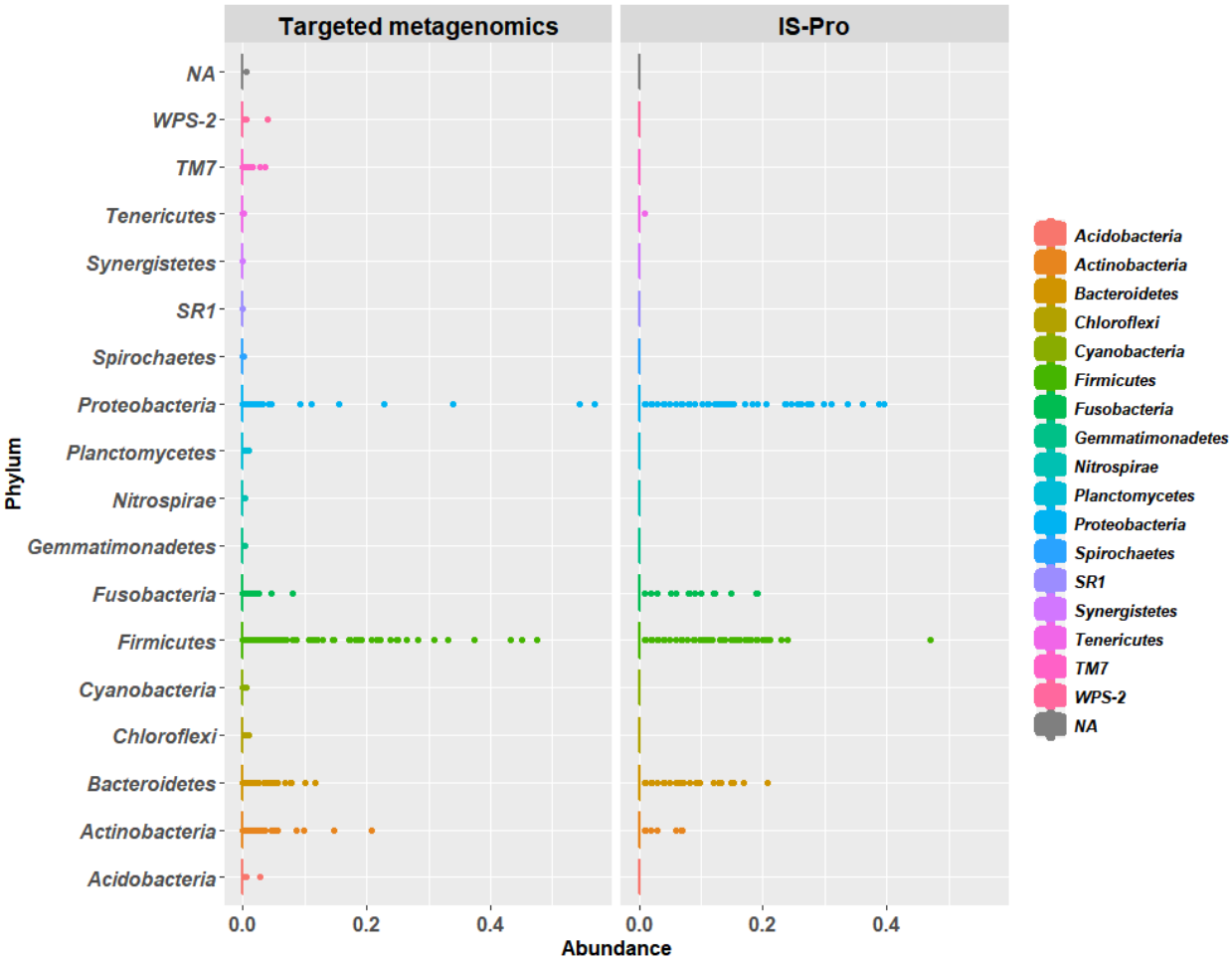

Figure 5.3: Relative abundance of specific phyla in the sputum microbiome of COPD participants as detected by the targeted metagenomics and IS-Pro methods (n=23). The dots represent the different abundances of each sample, according to the different phyla. Phyla that are depicted with a single line on the y-axis were not present in any samples for that method. The relative abundance is shown as a proportion of total abundance for the different methods.

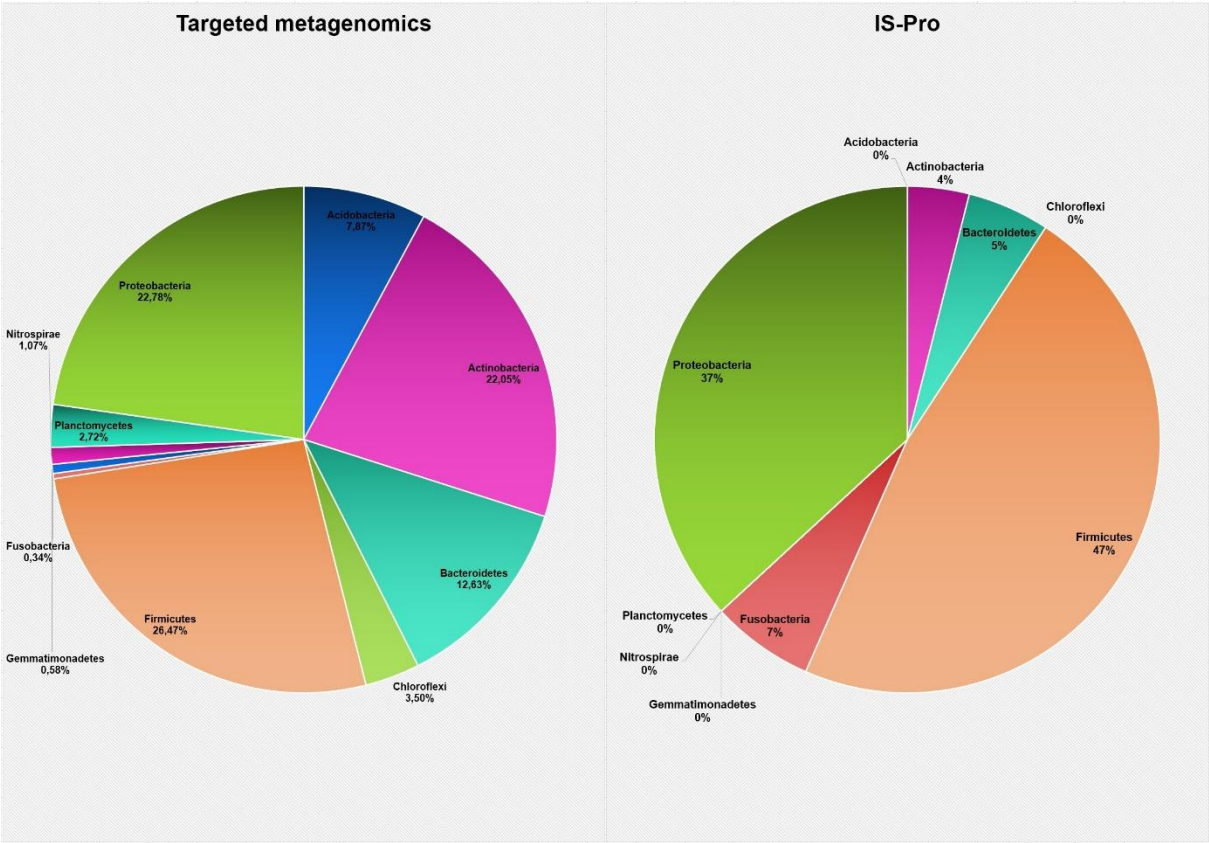

Figure 5.4: Relative abundance of specific phyla (depicted as pie graphs) in sample 29 as detected by the targeted metagenomics and IS-Pro methods.

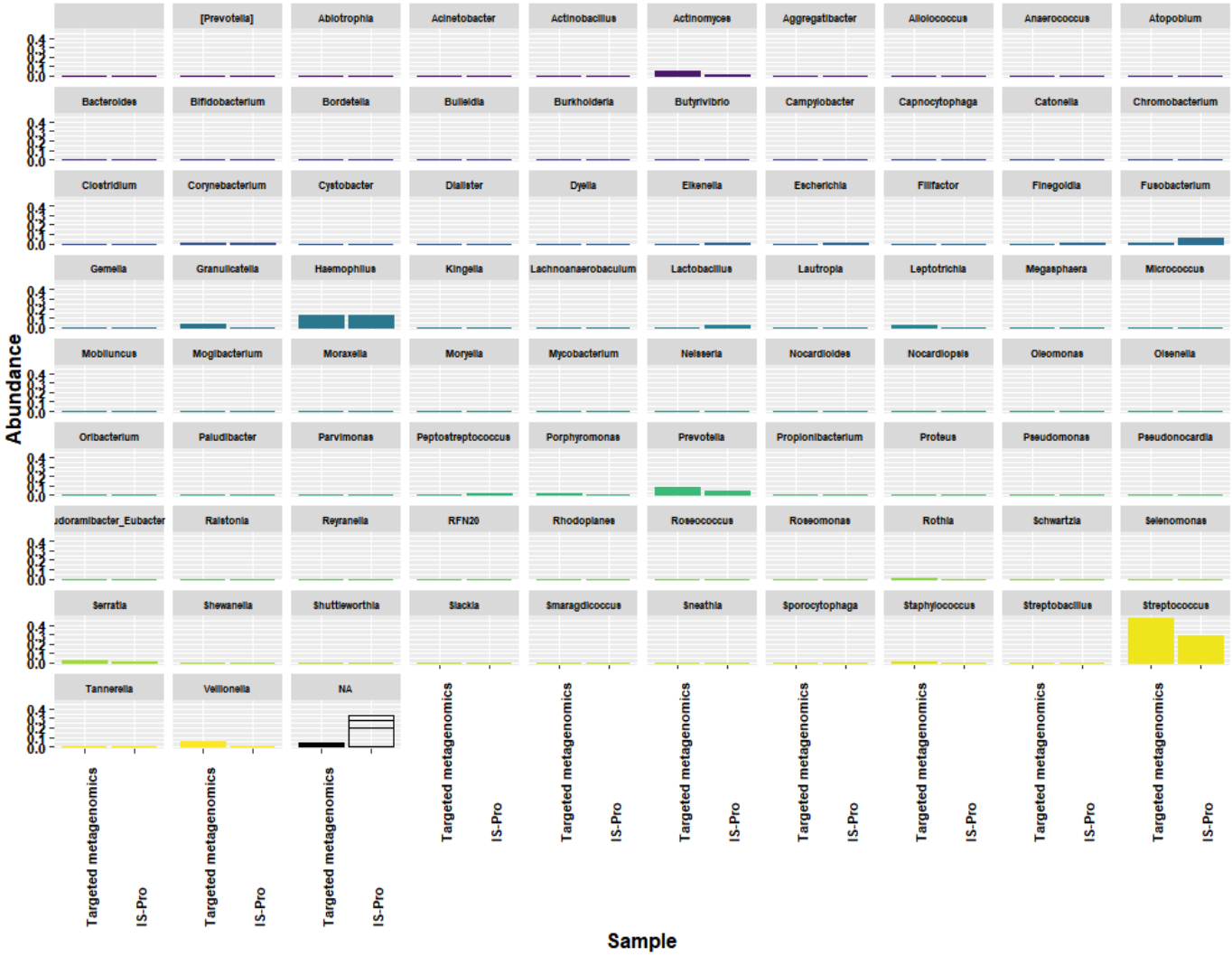

Figure 5.5: Bar plots showing the relative abundance of genera in the sputum microbiome of COPD participants as characterised by the targeted metagenomics and IS-Pro methods (n=23). The operational taxonomic units that could not be classified at a genus level are indicated as NA on the graph. The relative abundance is shown as a proportion of total abundance for the different methods.

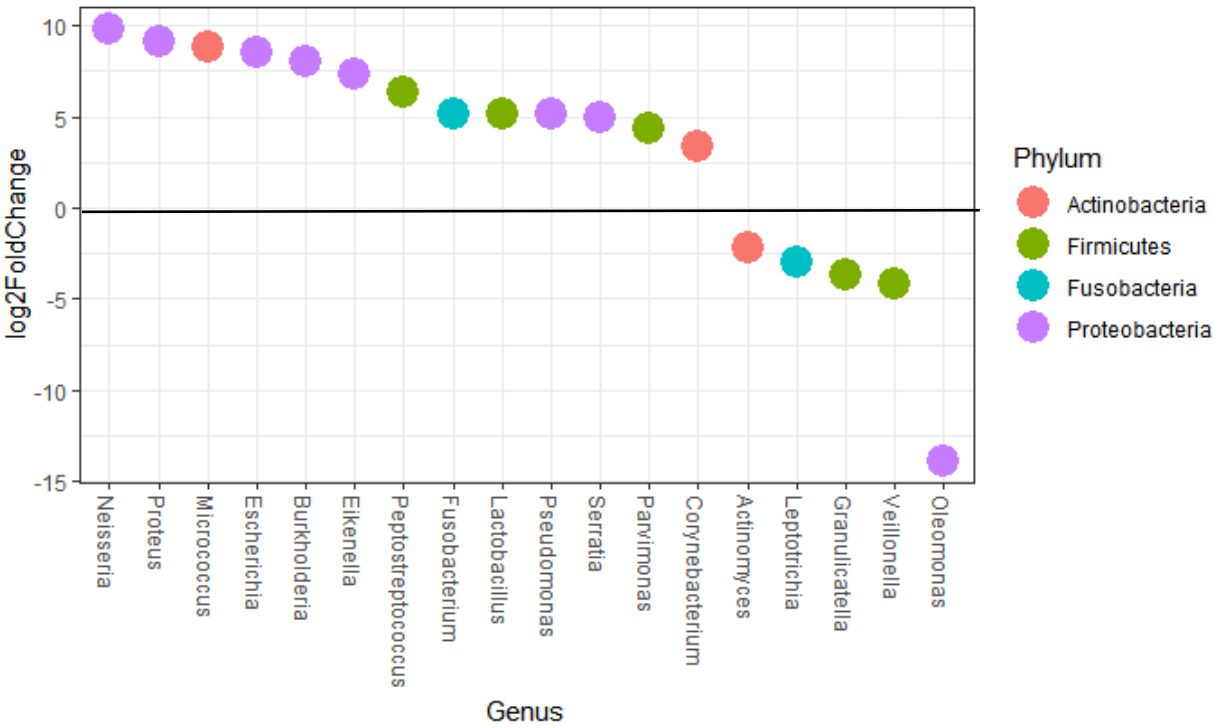

Figure 5.6: Graph of the DESeq2 analysis showing the log2fold differential abundance of the different genera between the targeted metagenomics and IS-Pro methods (n=23) in the sputum microbiome of COPD participants. Log2fold changes greater than zero indicated an increase in the relevant genera, whereas log2fold changes less than zero indicated a decrease in the relevant genera. All genera with dots above the zero line (indicated in black) had an increased relative abundance with the IS-Pro method when compared to the targeted metagenomics.

A comparison of the relative abundance of the targeted metagenomics and IS-Pro methods at genus level showed that the IS-Pro method had an increased abundance of 28 genera including *Burkholderia* (from 0.00% with targeted metagenomics to 0.82% with IS-Pro method), *Fusobacterium* (from 0.29% to 6.49%), *Lactobacillus* (from 0.10% to 2.64%), *Pseudomonas* (from 0.03% to 0.692%) and *Peptostreptococcus* (from 0.04% to 1.69%) (Figure 5.5) The IS-Pro method had a decreased abundance of 40 genera including *Streptococcus* (from 45.06% with targeted metagenomics to 29.39% with IS-Pro method), *Actinomyces* (from 5.72% to 0.74%), *Veillonella* (from 4.99% to 0.00%), *Prevotella* (from 8.77% to 4.89%), *Granulicatella* (from 3.59% to 0.00%), and *Leptotrichia* (from 2.44% to 0.00%). Further analysis showed that the IS-Pro method did not detect any *Veillonella*, *Granulicatella* or *Leptotrichia*. Using DESeq2 (Figure 5.6) to compare targeted metagenomics and IS-Pro methods showed a log2fold

difference in several genera; with thirteen genera observed in higher abundance with the IS-Pro method and five genera observed in lower abundance with the IS-Pro method. Approximately 50% (7/13) of the genera that were observed in higher abundances with the IS-Pro method belonged to the *Proteobacteria* phylum and included *Neisseria, Proteus, Escherichia, Burkholderia, Eikenella. Serratia* and *Pseudomonas*. Most of the genera that were observed in lower abundances with the IS-Pro method belonged to the *Firmicutes* phylum and included *Veillonella* and *Granulicatella*.

The IS-Pro method was able to classify more OTUs [86% (55/64)] to a species level than targeted metagenomics, which could classify only 23% (144/631) of the OTUs to a species level. However, the unclassified OTUs accounted for a higher relative abundance of the IS-Pro method (35%) than targeted metagenomics (5%) (Figure 5.5). The distribution of the unclassified phyla (at class level) for the IS-Pro method was as follows: 16% for *Firmicutes*, 23% for *Bacteroidetes* and 61% for *Proteobacteria* (Figure 5.7). Although not all the OTUs could be resolved at the genus level for targeted metagenomics, all could be classified at class level (Figure 5.7).

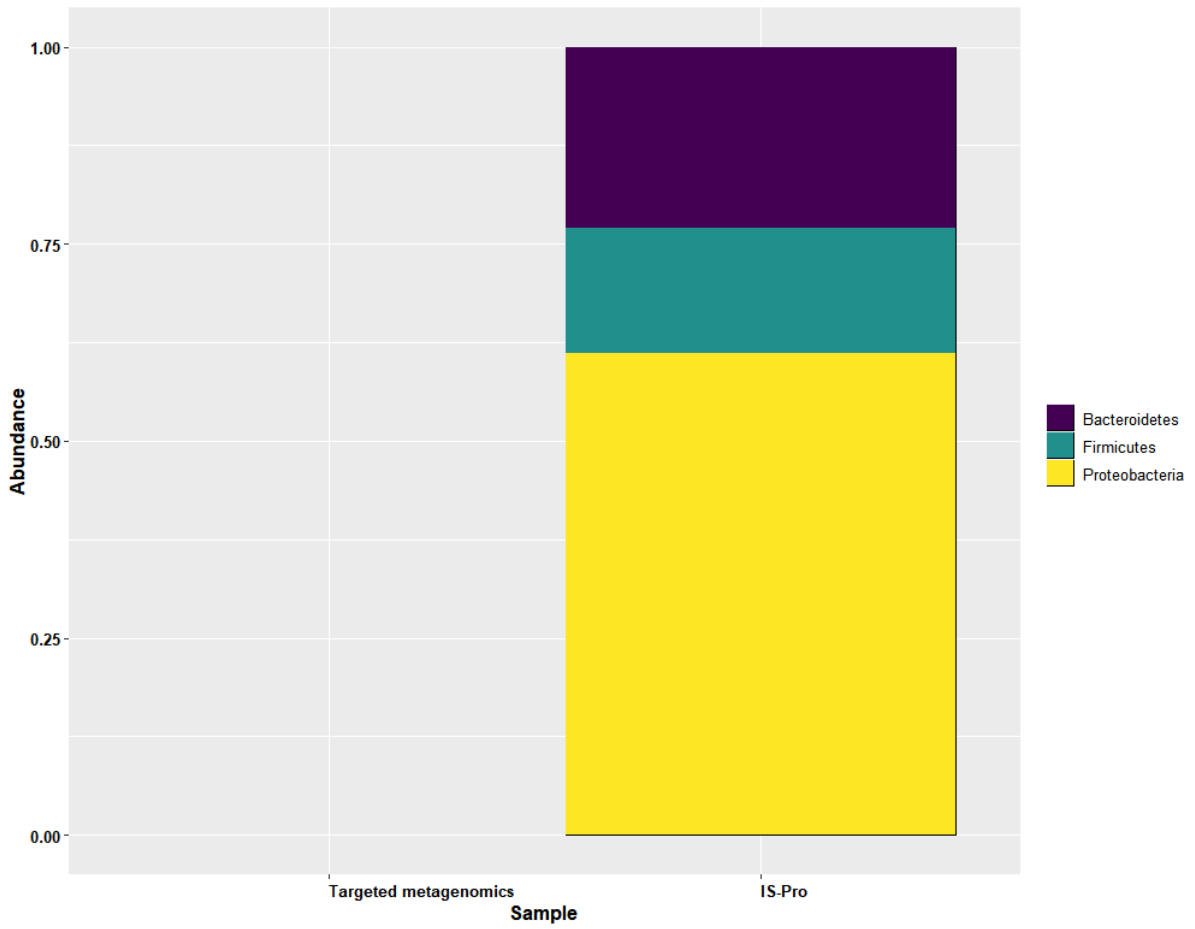

Figure 5.7: The distribution of the unclassified operational taxonomic units (OTUs) at a class level of the sputum microbiome of COPD participants for the targeted metagenomics and IS-Pro methods by phyla. At a class level, all the OTUs from targeted metagenomics could be classified.

5.3.4 Comparison of targeted metagenomics and IS-Pro methods in terms of cost-effectiveness, sample preparation and data analysis

The cost per isolate and time required for each technology is shown below (Table 5.1). The two technologies were compared in terms of cost, time and user-friendliness of data analysis software.

Table 5.1: Comparison of targeted metagenomics and IS-Pro methods in terms of cost, time and ease of use in our setting

Description	Targeted metagenomics	IS-Pro method
Laboratory cost per isolate*	$87.57(R 1 441.28)	$117.73 (R 1 937.85)
Turnaround time (from DNA extraction till statistical analysis)	9 days (user-dependent and platform-dependent)	7 days
Hands-on time (labour cost)	Laboratory: 5 days (1 day for DNA extraction, 4 days for sequencing and clean-up) Analysis: 4 days (3 days for analysis using QIIME and 1 day for statistical analysis)	Laboratory: 5 days (1 day for DNA extraction, 1 day for the IS-Pro PCR and 1 day for clean-up and 2 days for sequencing) Analysis: 2 days (1 day for analysis using IS-Pro proprietary software and 1 day for statistical analysis)
Steps involved	Bacterial DNA extraction PCR amplification of the target region Library preparation (and pooling of samples) Sequencing run Quality control analysis and generation of an OTU table using a program, such as QIIME2 Statistical analysis using a program, such as R	Bacterial DNA extraction PCR amplification using the IS-Pro kit Fragment analysis using a genetic analyser (uses capillary electrophoresis) Analysis of data and generation of an OTU table using IS-Pro proprietary software Statistical analysis using a program, such as R
Ease of use	Requires familiarity with Linux system	Easy to use (requires no prior knowledge of the IS-Pro propriety software)

*The cost is the cost at the time the study was conducted, is depicted in South African Rand and is dependent on international exchange rates (the cost of the dollar was based on the exchange rate on 04/10/2020)

Targeted metagenomics and IS-Pro methods are similar in one aspect: i) both require bacterial DNA extraction and PCR amplification before sequencing. However, analysis for targeted metagenomics is more complicated and the IS-Pro method is more expensive. The targeted metagenomics analysis requires QC analysis followed by clustering of sequences into OTUs and assigning taxonomy to the OTUs. This analysis requires the use of software, such as QIIME2 that is Linux-dependent and requires training to use correctly. The IS-Pro method uses proprietary software that only requires the upload of the sequencing data and the program performs the analysis, thereby requiring no prior knowledge or training.

5.4 Discussion

This study compared the targeted metagenomics and IS-Pro methods for their ability to determine the microbial composition of the lung microbiome in COPD patients. A single bacterial DNA extraction was performed for targeted metagenomics and IS-Pro methods to reduce bias. A comparison of targeted metagenomics and IS-Pro methods showed an increased

relative abundance of *Proteobacteria* for the IS-Pro method and a difference in alpha diversity and beta diversity between the two methods. This increased abundance could be attributed to bacteria, such as *Burkholderia*. Additionally, there was a log2fold difference between targeted metagenomics and IS-Pro methods in the abundance of several *Firmicutes* including *Veillonella*, which may indicate that the IS-Pro method is not optimised to detect *Firmicutes.*

A comparison of the alpha diversity analysis between the two technologies showed a statistically significant difference with the Shannon diversity measure, however, no statistically significant differences were detected using the Simpson diversity measure. The Shannon diversity measure is more sensitive to the number of species i.e. OTUs (richness) than the Simpson diversity measure [51]. The IS-Pro method had fewer OTUs than targeted metagenomics in this study and as such, this difference in alpha diversity between the Shannon (was statistically significant) and Simpson diversity (was not statistically significant) measures is not unexpected. In this study, the targeted metagenomics method had more OTUs than the IS-Pro method (the Shannon diversity was therefore statistically significant) whereas the relative abundance (evenness) with the two methods was similar (the Simpson index was therefore not statistically significant). In this study, the beta diversity analysis using PCoA plots showed two distinct clusters (of the same samples) that were associated with the two different technologies. With beta diversity analysis and particularly, cluster-based methods, such as PCoA, the more similar isolates are to each other the closer these isolates will cluster [52]. The results of this study can thus be interpreted as follows: i) the bacterial community structures in targeted metagenomics and IS-Pro methods are distinct i.e. using the same sample, the two methods showed differences between the microbiomes and ii) with the IS-Pro method, the community structure of samples were more similar to each other (in contrast, targeted metagenomics method showed samples that were more divergent from each other), i.e. targeted metagenomics showed a more diverse microbiome than the IS-Pro method. The alpha diversity and beta diversity results could not be compared to the literature at the time of publication, since there were limited microbiome studies that had performed a direct comparison between the targeted metagenomics and IS-Pro methods and none of these studies have reported diversity metrics; to determine if there is a difference in the alpha diversity and beta diversity, direct comparisons are needed [31, 53].

When the relative abundance profiles of the two technologies were compared, the IS-Pro method showed an increased abundance of the phylum *Proteobacteria* (16.1% for targeted

metagenomics and 38% for the IS-Pro method). There was only one other published study [by de Meij *et al.* (2016)] that used both targeted metagenomics and the IS-Pro method; however, this study did not observe an increase in *Proteobacteria*. However, this study was conducted using faecal samples of healthy children (n=61) and a different sequencing platform (454 sequencing) [31]. The phylum *Proteobacteria* is more commonly associated with disease and inflammation [32, 54]. The increased abundance of the *Proteobacteria* according to the IS-Pro method in this study could be attributed to the use of a master mix that contains primers that select specifically for members of the *Proteobacteria* phylum (PROTEO master mix; part of the IS-Pro kit), which may provide a selective advantage to this phylum [25]. This selective advantage of the master mix was observed for *Fusobacteria* as well (3% increase using the IS-Pro method). This observation was further highlight at a sample level; sample 29 (Figure 5.4) showed a higher relative abundance of *Proteobacteria* and *Fusobacteria* with the IS-Pro method.

At a genus level, the IS-Pro method showed a lower relative abundance for several genera, including *Streptococcus* (15% decrease), *Actinomyces* (5% decrease) and *Veillonella* (5% decrease) and an increased relative abundance of *Fusobacterium* (6% increase) and *Lactobacillus* (2.5% increase). Most of the genera that showed an increased relative abundance belonged to the *Proteobacteria* phylum, whereas the genera that showed a decreased relative abundance belonged mostly to the *Firmicutes* and *Actinobacteria* phyla. Members of the *Proteobacteria* phylum, which had log2fold increased abundance included *Burkholderia*, *Pseudomonas* and *Serratia*. These bacteria are known lung pathogens, although *Burkholderia* is more commonly found in cystic fibrosis (CF) patients than COPD patients [55-60]. Of the genera that showed a decreased relative abundance, three phyla were not detected by the IS-Pro method including *Granulicatella* (*Firmicutes*), *Leptotrichia* (*Fusobacteria*) and *Veillonella* (*Firmicutes*). Analysis of the current literature on targeted metagenomics and IS-Pro methods showed that for the same disease (such as irritable bowel disease), targeted metagenomics consistently detected *Veillonella* while the IS-Pro method only detected *Veillonella* in low numbers (or not at all) [26, 28-34, 37, 38, 40, 41, 43, 44, 61-69]. This limited detection of *Veillonella* with the IS-Pro method in these studies was surprising as most of the studies were conducted on faecal samples (i.e. the gastrointestinal tract) and this genus is a known coloniser of the gastrointestinal tract (as well as the lungs and oral cavity) and has been known to act as an opportunistic pathogen [70, 71]. Based on this analysis, it appears that the IS-Pro method has difficulty in detecting *Veillonella*, which may be due to primer design, the DNA target

region or analysis pipeline. A study by Mukherjee *et al.* (2018) provided a possible explanation for this by suggesting that *Veillonella* have multiple different intergenic spacer regions (these bacteria have different ribosomal operons that have different intergenic spacer regions), which may not be easily identifiable by the IS-Pro method analysis software and could be missed [72].

The IS-Pro method was able to identify more OTUs to a species level than targeted metagenomics, however, it showed a higher relative abundance (35%) of unclassified genera (i.e. OTUs that could not be assigned to a genus) than targeted metagenomics (5%). Most of the unclassified genera generated by targeted metagenomics could be classified to either a family or order level, however, the unclassified OTUs generated by the IS-Pro method could only be classified to a phylum level. As the current analysis strategy for the IS-Pro method does not include any QC steps, these unclassified OTUs may be low quality (short) sequences, chimeras or PCR artefacts [73]. It has been shown that the choice of the polymerase, the region sequenced, the number of PCR rounds, the platform used and even data analysis can affect the error rates with sequencing, however, these factors may affect the IS-Pro method as well even though the IS-Pro method uses capillary electrophoresis [73, 74]. The more errors introduced, the poorer the quality of the data is which affects the downstream analysis and could influence the bacteria identified [75]. A more detailed comparison between these two methods could not be achieved due to the nature of the outputs from the two methods and the proprietary nature of the IS-Pro method.

When comparing the time and ease of use of the two technologies, the IS-Pro method performed better than targeted metagenomics; the IS-Pro method was much easier to use (did not require the user to be familiar with Linux, i.e. requires a higher level of expertise) and had a faster turnaround time (7 days compared to 9 days for targeted metagenomics) (see Table 5.2). Essentially, targeted metagenomics needs a trained microbiologist or bioinformatician to analyse the data, whereas with the IS-Pro method any person can perform the analysis. The only disadvantage of the IS-Pro method was the operational cost was slightly more expensive than targeted metagenomics [$117.73 (R 1 937.85) compared to $87.57 (R 1 441.28) per sample].

Although this study had a small sample size and only studied a single disease, it provided a detailed comparison of targeted metagenomics and IS-Pro methods. Additionally, this was the first study to perform a direct comparison between targeted metagenomics and IS-Pro methods

on sputum specimens. The targeted metagenomics was able to detect more OTUs than the IS-Pro method and as a result, showed a more diverse microbiome population; however, these results could not be compared with other literature as there have been no studies that have performed a direct comparison between targeted metagenomics and IS-Pro methods. The targeted metagenomics and IS-Pro methods showed distinct communities for the same sample. Additionally, the IS-Pro method showed an overabundance of phyla, such as *Proteobacteria* and an underabundance of phyla, such as *Actinobacteria* and missed several genera that were identified using targeted metagenomics. These differing abundances were postulated to be the result of the IS-Pro kit design (primers that offered a selective advantage) and analysis software (lack of QC). However, while targeted metagenomics performed better than the IS-Pro method for the identification of the lung microbiome in this study [and gastrointestinal microbiome in other studies (based on indirect comparisons)] and was less costly, the IS-Pro method was easy to perform and analyse (using the propriety software) without any extensive training and had a shorter turnaround time. Based on the fact the IS-Pro method can miss relevant species, such as *Veillonella* and had more OTUs that could not be classified at a family level, a new IS-Pro kit with additional primers (for the amplification of *Veillonella*) and updated analysis software (with QC steps included), could result in an improved kit. The authors suggest that targeted metagenomics be used for research (as it had less bias towards certain phyla and genera) and the IS-Pro method be used as a diagnostic tool in clinical laboratories as it was able to identify most of the important clinical pathogens (especially those found in the lung), such as *Pseudomonas* and is easy to perform (the test can be conducted by any technician/technologist). However, due to the current pricing, the authors suggest the kit only be used in complicated cases or a reference laboratory. Future studies that compare targeted metagenomics and IS-Pro methods should include: i) different microbiomes, e.g. oral microbiome and skin microbiome, ii) different primer sets for the amplification of the 16S rRNA gene (to compare targeted metagenomics to the IS-Pro method), e.g. use primers that target the V4 region and iii) include a larger study population, preferably including different diseases.

5.5 Conclusions

The targeted metagenomics and the IS-Pro methods showed differences in their abilities to identify and characterise OTUs as well as in the diversity and microbial composition of the lung microbiome. The IS-Pro method might miss relevant species and could over-inflate the abundance of members of the *Proteobacteria.*, However, the IS-Pro kit was able to identify most of the important lung pathogens, such as *Burkholderia* and *Pseudomonas* and may work

well in a more diagnostics-oriented setting. Both methods were comparable in terms of time; however, the IS-Pro method was easier to use.

References

1. Foster KR, Schluter J, Coyte KZ, Rakoff-Nahoum S: **The evolution of the host microbiome as an ecosystem on a leash**. *Nature* 2017, **548**(7665):43-51.

2. Cani PD: **Human gut microbiome: hopes, threats and promises**. *Gut* 2018, **67**(9):1716-1725.

3. Flemming HC, Wuertz S: **Bacteria and archaea on Earth and their abundance in biofilms**. *Nat Rev Microbiol* 2019, **17**(4):247-260.

4. Lynch SV, Pedersen O: **The Human intestinal microbiome in health and disease**. *N Engl J Med* 2016, **375**(24):2369-2379.

5. Knight R, Callewaert C, Marotz C, Hyde ER, Debelius JW, McDonald D, Sogin ML: **The microbiome and human biology**. *Annu Rev Genomics Hum Genet* 2017, **18**:65-86.

6. Mammen MJ, Sethi S: **COPD and the microbiome**. *Respirology* 2016, **21**(4):590-599.

7. Su C, Lei L, Duan Y, Zhang KQ, Yang J: **Culture-independent methods for studying environmental microorganisms: methods, application, and perspective**. *Appl Microbiol Biotechnol* 2012, **93**(3):993-1003.

8. Hermann-Bank ML, Skovgaard K, Stockmarr A, Larsen N, Molbak L: **The Gut Microbiotassay: a high-throughput qPCR approach combinable with next generation sequencing to study gut microbial diversity**. *BMC Genomics* 2013, **14**:788.

9. Hiergeist A, Glasner J, Reischl U, Gessner A: **Analyses of intestinal microbiota: culture versus sequencing**. *ILAR J* 2015, **56**(2):228-240.

10. Hill C, Ross RP, Stanton C, O'Toole PW: **The human microbiome in health and disease**. In: *Host - Pathogen Interaction.* Edited by Unden G, Thines E, Schüffler A: John Wiley & Sons; 2016: 57-76.

11. Huang YJ, Erb-Downward JR, Dickson RP, Curtis JL, Huffnagle GB, Han MK: **Understanding the role of the microbiome in chronic obstructive pulmonary disease: principles, challenges, and future directions**. *Transl Res* 2017, **179**:71-83.

12. Benn A, Heng N, Broadbent JM, Thomson WM: **Studying the human oral microbiome: challenges and the evolution of solutions**. *Aust Dent J* 2018, **63**(1):14-24.

13. Kembel SW, Wu M, Eisen JA, Green JL: **Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance**. *PLoS Comput Biol* 2012, **8**(10):e1002743.

14. Martin C, Burgel PR, Lepage P, Andrejak C, de Blic J, Bourdin A, Brouard J, Chanez P, Dalphin JC, Deslee G *et al*: **Host-microbe interactions in distal airways: relevance to chronic airway diseases**. *Eur Respir Rev* 2015, **24**(135):78-91.

15. Woese CR, Fox GE: **Phylogenetic structure of the prokaryotic domain: the primary kingdoms**. *Proc Natl Acad Sci U S A* 1977, **74**(11):5088-5090.

16. Gürtler V, Subrahmanyam G, Shekar M, Maiti B, Karunasagar I: **Chapter 12- bacterial typing and identification by genomic analysis of 16S–23S rRNA intergenic transcribed spacer (ITS) sequences**. In: *Methods in Microbiology.* Edited by Michael Goodfellow, Iain Sutcliffe, Chun J, vol. 41: Academic Press; 2014: 253-274.

17. Madigan MT, Bender KS, Buckley DH, Sattley WM, Stahl DA: **Brock biology of microorganisms, global edition**. In., 15th eds. Harlow, United Kingdom: Pearson Education Limited; 2017.

18. Hao Y, Pei Z, Brown SM: **Bioinformatics in microbiome analysis**. In: *The Human Microbiome.* 2017: 1-18.

19. Yurovsky A, Amin MR, Gardin J, Chen Y, Skiena S, Futcher B: **Prokaryotic coding regions have little if any specific depletion of Shine-Dalgarno motifs**. *PLoS One* 2018, **13**(8):e0202768.

20. Yang B, Wang Y, Qian PY: **Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis**. *BMC Bioinformatics* 2016, **17**:135.

21. Osman MA, Neoh HM, Ab Mutalib NS, Chin SF, Jamal R: **16S rRNA Gene sequencing for deciphering the colorectal cancer gut microbiome: current protocols and workflows**. *Front Microbiol* 2018, **9**:767.

22. Mekuto L, Ntwampe SKO, Mudumbi JBN, Akinpelu EA, Mewa-Ngongang M: **Metagenomic data of free cyanide and thiocyanate degrading bacterial communities**. *Data Brief* 2017, **13**:738-741.

23. Kim M, Chun J: **16S rRNA gene-based identification of bacteria and archaea using the EzTaxon server**. In: *New Approaches to Prokaryotic Systematics.* 2014: 61-74.

24. Wang Z, Liu H, Wang F, Yang Y, Wang X, Chen B, Stampfli MR, Zhou H, Shu W, Brightling CE *et al*: **A refined view of airway microbiome in chronic obstructive pulmonary disease at species and strain-levels**. *Frontiers in Microbiology* 2020, **11**.

25. Budding AE, Grasman ME, Lin F, Bogaards JA, Soeltan-Kaersenhout DJ, Vandenbroucke-Grauls CM, van Bodegraven AA, Savelkoul PH: **IS-Pro: high-throughput molecular fingerprinting of the intestinal microbiota**. *FASEB J* 2010, **24**(11):4556-4564.

26. Budding AE, Hoogewerf M, Vandenbroucke-Grauls CM, Savelkoul PH: **Automated broad-range molecular detection of bacteria in clinical samples**. *J Clin Microbiol* 2016, **54**(4):934-943.

27. Budding AE, Grasman ME, Eck A, Bogaards JA, Vandenbroucke-Grauls CM, van Bodegraven AA, Savelkoul PH: **Rectal swabs for analysis of the intestinal microbiota**. *PLoS One* 2014, **9**(7):e101344.

28. Daniels L, Budding AE, de Korte N, Eck A, Bogaards JA, Stockmann HB, Consten EC, Savelkoul PH, Boermeester MA: **Fecal microbiome analysis as a diagnostic test for diverticulitis**. *Eur J Clin Microbiol Infect Dis* 2014, **33**(11):1927-1936.

29. Rutten NB, Gorissen DM, Eck A, Niers LE, Vlieger AM, Besseling-van der Vaart I, Budding AE, Savelkoul PH, van der Ent CK, Rijkers GT: **Long term development of gut microbiota composition in atopic children: Impact of probiotics**. *PLoS One* 2015, **10**(9):e0137681.

30. Aguirre M, Eck A, Koenen ME, Savelkoul PH, Budding AE, Venema K: **Diet drives quick changes in the metabolic activity and composition of human gut microbiota in a validated in vitro gut model**. *Res Microbiol* 2016, **167**(2):114-125.

31. de Meij TG, Budding AE, de Groot EF, Jansen FM, Frank Kneepkens CM, Benninga MA, Penders J, van Bodegraven AA, Savelkoul PH: **Composition and stability of intestinal microbiota of healthy children within a Dutch population**. *FASEB J* 2016, **30**(4):1512-1522.

32. de Meij TG, de Groot EF, Eck A, Budding AE, Kneepkens CM, Benninga MA, van Bodegraven AA, Savelkoul PH: **Characterization of microbiota in children with chronic functional constipation**. *PLoS One* 2016, **11**(10):e0164731.

33. Janssens PL, Penders J, Hursel R, Budding AE, Savelkoul PH, Westerterp-Plantenga MS: **Long-term green tea supplementation does not change the human gut microbiota**. *PLoS One* 2016, **11**(4):e0153134.

34. Eck A, de Groot EFJ, de Meij TGJ, Welling M, Savelkoul PHM, Budding AE: **Robust microbiota-based diagnostics for inflammatory bowel disease**. *J Clin Microbiol* 2017, **55**(6):1720-1732.

35. Meij T, Mulder C, Budding A, Vermont C, Meijer L, Beurden Y: **Serial microbiota analysis after fecal microbiota transplantation in a child with Down's Syndrome**. *Journal of Pediatric Infectious Diseases* 2017.

36. Muller PH, de Meij TGJ, Westedt M, de Groot EFJ, Allaart CF, Brinkman DMC, Schonenberg-Meinema D, van den Berg M, van Suijlekom-Smit LWA, van Rossum M *et al*:

Disturbance of microbial core species in new-onset juvenile idiopathic arthritis. *J Pediat Infect Dis* 2017, **12**(2):131-135.

37. Berkhout DJC, Niemarkt HJ, Benninga MA, Budding AE, van Kaam AH, Kramer BW, Pantophlet CM, van Weissenbruch MM, de Boer NKH, de Meij TGJ: **Development of severe bronchopulmonary dysplasia is associated with alterations in fecal volatile organic compounds**. *Pediatr Res* 2018, **83**(2):412-419.

38. de Meij TGJ, de Groot EFJ, Peeters CFW, de Boer NKH, Kneepkens CMF, Eck A, Benninga MA, Savelkoul PHM, van Bodegraven AA, Budding AE: **Variability of core microbiota in newly diagnosed treatment-naive paediatric inflammatory bowel disease patients**. *PLoS One* 2018, **13**(8):e0197649.

39. Koedooder R, Singer M, Schoenmakers S, Savelkoul PHM, Morre SA, de Jonge JD, Poort L, Cuypers WSS, Budding AE, Laven JSE *et al*: **The ReceptIVFity cohort study protocol to validate the urogenital microbiome as predictor for IVF or IVF/ICSI outcome**. *Reprod Health* 2018, **15**(1):202.

40. Calon TGA, Trobos M, Johansson ML, van Tongeren J, van der Lugt-Degen M, Janssen AML, Savelkoul PHM, Stokroos RJ, Budding AE: **Microbiome on the bone-anchored hearing system: A prospective study**. *Front Microbiol* 2019, **10**:799.

41. El Manouni El Hassani S, de Boer NKH, Jansen FM, Benninga MA, Budding AE, de Meij TGJ: **Effect of daily intake of *Lactobacillus casei* on microbial diversity and dynamics in a healthy pediatric population**. *Curr Microbiol* 2019, **76**(9):1020-1027.

42. Koedooder R, Singer M, Schoenmakers S, Savelkoul PHM, Morre SA, de Jonge JD, Poort L, Cuypers W, Beckers NGM, Broekmans FJM *et al*: **The vaginal microbiome as a predictor for outcome of in vitro fertilization with or without intracytoplasmic sperm injection: a prospective study**. *Hum Reprod* 2019, **34**(6):1042-1054.

43. Budding A, Sieswerda E, Wintermans B, Bos M: **An age dependent pharyngeal microbiota signature associated with SARS-CoV-2 infection**. *SSRN Electronic Journal* 2020.

44. Eck A, Rutten N, Singendonk MMJ, Rijkers GT, Savelkoul PHM, Meijssen CB, Crijns CE, Oudshoorn JH, Budding AE, Vlieger AM: **Neonatal microbiota development and the effect of early life antibiotics are determined by two distinct settler types**. *PLoS One* 2020, **15**(2):e0228133.

45. Hamid Q, Kelly MM, Linden M, Louis R, Pizzichini MMM, Pizzichini E, Ronchi C, Van Overveld F, Djukanovic R: **Methods of sputum processing for cell counts, immunocytochemistry and *in situ* hybridisation**. *Eur Respir J* 2002, **20**(Supplement 37):19S-23s.

46. Allen V, Nicol MP, Ah Tow L: **Sputum processing prior to *Mycobacterium tuberculosis* detection by culture or nucleic acid amplification testing: A narrative review**. 2016, **5**(1):96-109.

47. Terranova L, Oriano M, Teri A, Ruggiero L, Tafuro C, Marchisio P, Gramegna A, Contarini M, Franceschi E, Sottotetti S *et al*: **How to process sputum samples and extract bacterial DNA for microbiota analysis**. *Int J Mol Sci* 2018, **19**(10):3256-3568.

48. Stokell JR, Khan A, Steck TR: **Mechanical homogenization increases bacterial homogeneity in sputum**. *J Clin Microbiol* 2014, **52**(7):2340-2345.

49. Oluseyi Osunmakinde C, Selvarajan R, Mamba BB, Msagati TAM: **Profiling bacterial diversity and potential pathogens in wastewater treatment plants using high-throughput sequencing analysis**. *Microorganisms* 2019, **7**(11):506-524.

50. Mohsen A, Park J, Chen YA, Kawashima H, Mizuguchi K: **Impact of quality trimming on the efficiency of reads joining and diversity analysis of Illumina paired-end reads in the context of QIIME1 and QIIME2 microbiome analysis frameworks**. *BMC Bioinformatics* 2019, **20**(1):581.

51. Johnson KV, Burnet PW: **Microbiome: Should we diversify from diversity?** *Gut Microbes* 2016, **7**(6):455-458.

52. Borcard D, Gillet F, Legendre P: **Numerical ecology with R**: Springer International Publishing; 2018.

53. Singer M, Koedooder R, Bos M, Poort L, Savelkoul PHM, Laven J, Morré SA, Budding A: **The profiling of microbiota in vaginal swab samples using two different techniques**. In: *29th ECCMID Amsterdam, The Netherlands*. 2019.

54. Rizzatti G, Lopetuso LR, Gibiino G, Binda C, Gasbarrini A: ***Proteobacteria*: A common factor in human diseases**. *Biomed Res Int* 2017, **2017**:9351507.

55. Singh S, Sharma A, Nag VL: **Bacterial pathogens from lower respiratory tract infections: A study from Western Rajasthan**. *J Family Med Prim Care* 2020, **9**(3):1407-1412.

56. Jones AM: **Which pathogens should we worry about?** *Paediatr Respir Rev* 2019, **31**:15-17.

57. Henao-Martínez AF, Montoya JG: **Infections in heart, lung, and heart-lung transplantation**. In: *Principles and Practice of Transplant Infectious Diseases.* 2019: 21-39.

58. Fenker DE, McDaniel CT, Panmanee W, Panos RJ, Sorscher EJ, Sabusap C, Clancy JP, Hassett DJ: **A comparison between two pathophysiologically different yet microbiologically similar lung diseases: Cystic fibrosis and chronic obstructive pulmonary disease**. *Int J Respir Pulm Med* 2018, **5**(2).

59. de Vrankrijker AM, Wolfs TF, van der Ent CK: **Challenging and emerging pathogens in cystic fibrosis**. *Paediatr Respir Rev* 2010, **11**(4):246-254.

60. Sze MA, Dimitriu PA, Hayashi S, Elliott WM, McDonough JE, Gosselink JV, Cooper J, Sin DD, Mohn WW, Hogg JC: **The lung tissue microbiome in chronic obstructive pulmonary disease**. *Am J Respir Crit Care Med* 2012, **185**(10):1073-1080.

61. Dols JA, Molenaar D, van der Helm JJ, Caspers MP, de Kat Angelino-Bart A, Schuren FH, Speksnijder AG, Westerhoff HV, Richardus JH, Boon ME *et al*: **Molecular assessment of bacterial vaginosis by *Lactobacillus* abundance and species diversity**. *BMC Infect Dis* 2016, **16**:180.

62. Peeters T, Penders J, Smeekens SP, Galazzo G, Houben B, Netea MG, Savelkoul PH, Gyssens IC: **The fecal and mucosal microbiome in acute appendicitis patients: an observational study**. *Future Microbiol* 2019, **14**:111-127.

63. Shaw KA, Bertha M, Hofmekler T, Chopra P, Vatanen T, Srivatsa A, Prince J, Kumar A, Sauer C, Zwick ME *et al*: **Dysbiosis, inflammation, and response to treatment: a longitudinal study of pediatric subjects with newly diagnosed inflammatory bowel disease**. *Genome Med* 2016, **8**(1):75.

64. Santoru ML, Piras C, Murgia A, Palmas V, Camboni T, Liggi S, Ibba I, Lai MA, Orru S, Blois S *et al*: **Cross sectional evaluation of the gut-microbiome metabolome axis in an Italian cohort of IBD patients**. *Sci Rep* 2017, **7**(1):9523.

65. Lo Presti A, Zorzi F, Del Chierico F, Altomare A, Cocca S, Avola A, De Biasio F, Russo A, Cella E, Reddel S *et al*: **Fecal and mucosal microbiota profiling in irritable bowel syndrome and inflammatory bowel disease**. *Front Microbiol* 2019, **10**:1655.

66. Altomare A, Putignani L, Del Chierico F, Cocca S, Angeletti S, Ciccozzi M, Tripiciano C, Dalla Piccola B, Cicala M, Guarino MPL: **Gut mucosal-associated microbiota better discloses inflammatory bowel disease differential patterns than faecal microbiota**. *Dig Liver Dis* 2019, **51**(5):648-656.

67. Koliani-Pace JL, Siegel CA: **Prognosticating the course of inflammatory bowel sisease**. *Gastrointest Endosc Clin N Am* 2019, **29**(3):395-404.

68. Olbjorn C, Cvancarova Smastuen M, Thiis-Evensen E, Nakstad B, Vatn MH, Jahnsen J, Ricanek P, Vatn S, Moen AEF, Tannaes TM *et al*: **Fecal microbiota profiles in treatment-naive pediatric inflammatory bowel disease - associations with disease phenotype, treatment, and outcome**. *Clin Exp Gastroenterol* 2019, **12**:37-49.

69. Pittayanon R, Lau JT, Leontiadis GI, Tse F, Yuan Y, Surette M, Moayyedi P: **Differences in gut microbiota in patients with vs without inflammatory bowel diseases: A systematic review**. *Gastroenterology* 2020, **158**(4):930-946 e931.

70. Jorgensen JH, Pfaller MA: **Manual of clinical microbiology**. Washington, USA: ASM Press; 2015.

71. Chen YC, Ko PH, Yang CJ, Chen YC, Lay CJ, Tsai CC, Hsieh MH: **Epidural abscess caused by *Veillonella parvula*: Case report and review of the literature**. *J Microbiol Immunol Infect* 2016, **49**(5):804-808.

72. Mukherjee C, Beall CJ, Griffen AL, Leys EJ: **High-resolution ISR amplicon sequencing reveals personalized oral microbiome**. *Microbiome* 2018, **6**(1):153.

73. Sze MA, Schloss PD: **The impact of DNA polymerase and number of rounds of amplification in PCR on 16S rRNA gene sequence data**. *mSphere* 2019, **4**(3).

74. Faner R, Sibila O, Agusti A, Bernasconi E, Chalmers JD, Huffnagle GB, Manichanh C, Molyneaux PL, Paredes R, Perez Brocal V *et al*: **The microbiome in respiratory medicine: current challenges and future perspectives**. *Eur Respir J* 2017, **49**(4).

75. Lo CC, Chain PS: **Rapid evaluation and quality control of next generation sequencing data with FaQCs**. *BMC Bioinformatics* 2014, **15**:366.

CHAPTER 6

CONCLUDING REMARKS

6.1 Conclusions

Chronic obstructive pulmonary disease (COPD) is a lung disease characterised by airflow limitations and persistent respiratory symptoms (Global Initiative for Chronic Obstructive Lung Disease, 2020). This disease is one of the leading causes of morbidity and mortality worldwide (ranked fourth) and was estimated to be the world's third leading cause of death by 2020 (Abdool-Gaffar *et al.*, 2019). Factors that may contribute to COPD include smoking, exposure to biomass fumes (indoor air pollution), exposure to occupational dust (e.g. working in a mine), genetic factors and *Mycobacterium tuberculosis* infection (TB) (Doring *et al.*, 2011; van Gemert *et al.*, 2011; Salvi, 2015; Lalloo *et al.*, 2016; Abdool-Gaffar *et al.*, 2019). Human immunodeficiency virus (HIV) infection has been implicated as a risk factor for COPD as well (Lalloo *et al.*, 2016; Bigna *et al.*, 2018). A key feature of COPD is the inflammation of the airways, which results in an impaired response, allowing the lung to be colonised by microorganisms, such as bacteria and viruses (Molyneaux *et al.*, 2013; Cullen and McClean, 2015; Fan *et al.*, 2016). As COPD progresses, patients will experience states of worsened symptoms (both respiratory and non-respiratory) that are referred to as exacerbations (Miravitlles and Anzueto, 2015; Pavord *et al.*, 2016). Most of these exacerbations have been linked to infection by either bacterial (50%) or viral (30%) agents (Shimizu *et al.*, 2015). However, bacteria and viruses have been detected during the stable state of disease and their role in COPD is unclear (Doring *et al.*, 2011; D'Anna *et al.*, 2016).

A study conducted in Cameroon in 2012 and 2013 showed no difference in the prevalence of COPD in HIV-positive and HIV-negative individuals (Pefura-Yone *et al.*, 2015; Ho *et al.*, 2019). A study conducted on a HIV-positive population in South Africa, suggests that the lung function decline (in HIV-positive individuals) is more likely associated with TB infection than COPD (Varkila *et al.*, 2019). A study conducted in Uganda that sought to determine if there were associations between HIV, TB and COPD, could not draw any clear conclusions (North *et al.*, 2017). While a study conducted on older patients (over 50 years of age) admitted to a hospital in the North West Province, South Africa, showed a lower prevalence of COPD in the HIV-positive population (Naidoo *et al.*, 2020). These studies suggest that even though HIV has

been implicated as a risk factor for COPD, in South Africa, there appears to be no clear association between HIV and COPD.

Based on the above data the prevalence of COPD in the HIV-positive population and the general population should be similar, however, in this study we were only able to recruit one HIV-positive participant with COPD. Recruitment occurred at three different study sites including a tertiary academic hospital and a private hospital in an urban setting. The population attending these study sites may be younger (the national HIV prevalence in the 50 years and older category is 12.5%) or that these populations have a low viral load as most of the population is on antiretroviral therapy (ARTs); Bigna *et al.* (2018) observed an association between viral load and COPD in a meta-analysis that include over 30 studies that suggests higher HIV viral loads are associated with COPD.

As only one HIV participant could be recruited, this study could not compare the effects of HIV infection on the COPD lung microbiome. As previous studies, such as the study by Twigg *et al.* (2017), showed that HIV infection affected the healthy lung microbiome, further studies into the COPD lung microbiome with more HIV-positive individuals are still needed. The different states in COPD of the bacterial microbiome could be compared using next-generation sequencing. This study used the V1-V3 region of the 16S rRNA gene (i.e. targeted metagenomics) and the MiSeq platform to determine the lung microbiome in both disease states. A total of 24 participants (including one HIV-positive participant) were recruited in this study. At the time of collection, 18 participants were in the stable state of disease and six participants were in the exacerbation state. Chapter 4 discussed the methodology and results of the bacterial microbiome and virome of sputum specimens of stable and exacerbation COPD in detail. While individual samples showed variation in the alpha diversity and beta diversity, no specific differences were noted between the stable and exacerbation states of COPD in the current study. However, the relative abundance of the key phyla, i.e. *Firmicutes*, *Bacteroidetes*, *Fusobacteria*, *Proteobacteria* and *Actinobacteria* differed between the disease states; *Firmicutes* had a higher relative abundance during the exacerbation state and the other phyla had a lower relative abundance. This higher relative abundance *of Firmicutes* has been observed in previous studies in Europe and the USA, however, the implications of this are unclear but maybe be due to increased oxidative stress in the lungs caused during COPD progression (McGuinness and Sapey, 2017; Hufnagl *et al.*, 2020). The higher prevalence of *Firmicute*s could not be attributed to a single genus in this study; however, previous studies could attribute

these changes in the prevalence of *Firmicutes* to a single genus, such as *Lactobacillus* and other members of the *Lactobacillales* order, such as *Streptococcus* (Pragman *et al.*, 2012; Sze *et al.*, 2012; Kim *et al.*, 2017; Jubinville *et al.*, 2018; Leiten *et al.*, 2020). In this study, two *Firmicutes* genera showed a higher prevalence, *Granulicatella* (*Lactobacillales*) and *Veillonella* (*Veillonellales*). The difference in abundance between the previous studies and this study could be attributed to the gastrointestinal tract and lung cross-talk (due to the movement of various bacteria from the gastrointestinal tract to the lung) and the diet of the different populations (Tennert *et al.*, 2020). Additionally, these bacteria have been associated with gastrointestinal reflux disease (GERD) and may act as indicators of COPD exacerbations and could be indicated as a potential treatment point, i.e. treatment of GERD may improve COPD exacerbations (Lee and Goldstein, 2015; Park *et al.*, 2020; Sanchez *et al.*, 2020). This observation showed that differences in the abundances of *Firmicutes* at genera level could have an impact on the treatment program and clinical outcomes of COPD patients.

In order to reduce the potential for bias between the results of targeted metagenomics and IS-Pro methods, a single bacterial DNA extraction was performed. The extracted DNA for each sample was aliquoted for the two analyses, one for targeted metagenomics and one for the IS-Pro method. Analysis of the microbiome using the IS-Pro method showed similar results to targeted metagenomics. Both methods showed diverse bacterial communities in the lung (based on the alpha diversity analysis); however, the two communities were distinct (based on the beta diversity analysis). This clear distinction, between the two microbial communities detected, highlights the impact that the different region of the ribosomal RNA (the hypervariable regions of the 16S rRNA gene with targeted metagenomics and the intergenic spacer region between the 16S rRNA and 23S rRNA genes for IS-Pro method) and analysis methodology can have on the microbiome of the same sample as shown in Chapter 5. To summarise, the IS-Pro method detected higher relative abundances of *Proteobacteria* and *Fusobacteria* and lower relative abundance of *Actinobacteria*. These differences in relative abundance could be attributed to the design of the assay (i.e. the IS-Pro method), particularly the primer design and the composition of the master mixes. One key genus that was not detected with the IS-Pro method was *Veillonella* which could be attributed to either primer design of the IS-Pro method (which may have had a selective advantage for other genera) or to the intergenic spacer (IS) region (since this genus contains multiple copies of the IS region that may be polymorphic). Additionally, while the IS-Pro method identified more OTUs to a species level, it showed a higher frequency of unclassified genera than targeted metagenomics. These sequences could be only

characterised at a phylum level; the IS-Pro method does not detect OTUs but directly detects the abundance of the different bacteria at a species level. As a result, all bacteria that could not be classified at a certain taxonomic level are grouped, i.e. the abundance of unclassified sequences are not attributable to a single OTU but could belong to multiple OTUs. These OTUs could be false OTUs (due to the lack of classification), that were generated during the PCR process of the IS-Pro method and occurred due to a lack of quality control measures (in the IS-Pro method analysis) to remove chimeras and PCR artefacts (Caporaso *et al.*, 2011; Auer *et al.*, 2017; Sze and Schloss, 2019). The IS-Pro kit could be improved by reducing the number of PCR cycles, by using an improved algorithm to detect sequences that are shorter than the average and by constantly updating the database of the IS-Pro database to include rare and novel species. When comparing targeted metagenomics and IS-Pro methods, the following conclusions could be made: i) the IS-Pro method needs to be improved before it can be used as a research tool for investigating the lung microbiome as this method shows an overabundance of the *Proteobacteria* phylum, may miss genera and had several OTUs that could not be resolved at a family level and ii) the IS-Pro method could work well in a diagnostic-orientated setting for the determination of causative bacteria in polymicrobial infections as this method can detect clinically relevant bacteria, such as *Fusobacterium* (that was present in 21/23 samples and showed a higher prevalence with the IS-Pro method).

Next-generation sequencing to determine the virome (using shotgun metagenomics) was conducted on a subset of six samples (three stable and three exacerbation samples). Due to the small sample size, the virome of the different disease states (stable and exacerbation states of COPD) could not be compared. The results showed that the virome of COPD participants was dominated by i) the *Poxviridae* family (present in the highest frequencies) and ii) bacteriophages families, such as *Siphoviridae*. The high *Poxviridae* frequency was attributed to being due to the detection of the BeAn 58058 virus. The details of this virus are discussed in section 4.4. The source of this virus is unknown, it could be from the environment (and may have been introduced as laboratory contamination) or it could be present in the human genome. Based on the analysis of the virome in this study (with only selected samples) and previous literature on the COPD lung virome, most of the viruses found in the COPD lung have no known pathogenicity (Garcia-Nunez *et al.*, 2018; van Rijn *et al.*, 2019). These findings suggest that the virome does not have a direct impact on the pathogenesis in the COPD lung and as such future studies should only focus on the virome to determine the effect of specific viral pathogens, e.g. of influenza or respiratory syncytial viruses.

At the beginning of this study there were several gaps in the knowledge of the microbiome of COPD patients including i) no studies had been conducted comparing the microbiome between HIV-positive and HIV-negative individuals with COPD, ii) limited studies had been conducted on the virome of COPD patients and iii) none of the studies had been conducted in Africa. This study was one of the first studies conducted in Africa and one of the first studies to observe the virome of COPD patients. However, due to challenges in patient recruitment, this study was unable to compare HIV-positive and HIV-negative individuals and had a limited study population size. This study highlighted that the microbiome of COPD patients in Africa is similar to the microbiome of COPD patients in Europe and America, with minor differences in the frequencies of key phyla and genera and provided an overview of the virome in COPD patients. Additionally, the study identified several new findings, such as i) the bacteria that were detected in higher abundance during exacerbation have been previously associated with GERD; these bacteria could potentially be used as predictors of diseases and may have identified a potential treatment area for COPD patients and ii) the high prevalence of the BeAn 58058 virus that was found in all six samples. This study would have been further strengthened if paired samples could have been obtained for the stable and exacerbation states of disease and if a longitudinal study had been conducted. Additionally, this study highlighted several key areas for future research including studying the COPD microbiome in conjunction with its comorbidities and further studies of the BeAn 58058 virus and its clinical relevance. This study also highlighted the need to study the virome in chronic respiratory diseases as most studies to have focused on cystic fibrosis, with other diseases, such as asthma and sarcoidosis have been neglected. Seasonal variation with viruses, such as Influenza A (at its potential impact on the microbiome and virome) as well as the effects on the global pandemic caused by SARS-CoV-2 on chronic respiratory disease needs to be evaluated.

6.2 Future Research

The microbiome of the lung has been extensively studied in COPD, cystic fibrosis and other disease as well as in the healthy lung (Fabbrizzi *et al.*, 2019). However, while the microbiome i.e. microbial composition of the lung has been well characterised, most of these studies have only been conducted using targeted metagenomics (Fabbrizzi *et al.*, 2019). The use of targeted metagenomics is limited as it only provides the bacterial composition and it does not provide information of the role of these bacteria in health and disease (Charalampous *et al.*, 2019; Fabbrizzi *et al.*, 2019). Shotgun metagenomics can provide information about the functional

capacity of these bacteria and other microorganism, such as viruses (Fabbrizzi *et al.*, 2019; Sun *et al.*, 2020). However, shotgun metagenomics is mostly performed on DNA, as a result, live and dead cells cannot be differentiated; RNA viruses cannot be detected and all data generated is conjecture only (Emerson *et al.*, 2017; Quince *et al.*, 2017). Metatranscriptomics can use RNA and can differentiate between living and dead microorganisms (Emerson *et al.*, 2017). Additionally, metatranscriptomics or rather RNA-seq (RNA sequencing) can detect RNA viruses (Shi *et al.*, 2018; Noell and Kolls, 2019). Metabolomics provides information about microbial-derived metabolites or host-derived metabolites that have been modified by microorganisms (Ditz *et al.*, 2020). Previous studies have used this method to differentiate between different COPD types and to differentiate lung cancer from COPD (Deja *et al.*, 2014; Nobakht *et al.*, 2015). However, more metabolomic, metatranscriptomic and shotgun metagenomic studies are needed to study COPD (Millares *et al.*, 2015; Lee *et al.*, 2016).

One of the aspects of COPD research that warrants further study is a comparison of the different COPD phenotypes; shotgun metagenomics, metabolomics or metatranscriptomics studies would provide more information on these phenotypes and identify potential biomarkers and areas for therapy. Chronic obstructive pulmonary disease is a complex disease that has been grouped together based on characteristic airflow limitations (Sin, 2018). It has been postulated that the different phenotypes of COPD have different microbial influences; by using shotgun metagenomics combined with metatranscriptomics the genes that are more active in each phenotype could be determined, whereas metabolomics could provide us with biomarkers indicative of the different phenotypes which would have a beneficial effect on patient treatment and outcome. Another area of research would be to compare different lung diseases, such as asthma with COPD; by using metabolomics, specific biomarkers could be identified that differentiate between the different disease and may improve patient outcome. Additionally, by testing the “healthy” smoker population as well as an ageing population over a long period using either metabolomics or metatranscriptomics, biomarkers for the determination of COPD onset may be determined. These methods could be used to evaluate the effectiveness of antibiotics and other treatments over long periods.

However, before these studies can be conducted standardisation of the methods used to study the lung microbiome is needed. Both biological (e.g. diet, disease and body site) and methodological factors (e.g. sequencing platform and bioinformatics analysis) can affect the microbiome and can make comparisons between studies difficult (Faner *et al.*, 2017; Rogers,

2017; Ditz *et al.*, 2020). In lung microbiome studies, sampling (and the choice of body site) is one of the factors that has the greatest impact on the microbiome, specifically the type of specimen used. Lung microbiome studies have previously used bronchial alveolar lavage (BAL), exhaled breath condensate, bronchoscopy, lung tissue biopsies or explants, oropharyngeal swabs or sputum specimens (Faner *et al.*, 2017; Moffatt and Cookson, 2017; Ditz *et al.*, 2020). However, studies have shown that these specimens have distinct microbiomes (Hogan *et al.*, 2016; Chang *et al.*, 2020). The choice of specimen is affected by factors, such as how representative is the sample of the lower airways and on the invasiveness of the collection of the specimen (Carney *et al.*, 2020; Ditz *et al.*, 2020; Sulaiman *et al.*, 2020). Most of the specimens that are representative of the lower airways are invasive e.g. BAL, however, sputum specimen is non-invasive and has a component of the lower airways (even though it contains components of the upper airways) (Carney *et al.*, 2020; Ditz *et al.*, 2020; Sulaiman *et al.*, 2020). Sampling, processing and approach to bioinformatical analysis can each impact the microbiome generated and therefore standardised protocols for each of these steps is important (Faner *et al.*, 2017; Rogers, 2017; Ditz *et al.*, 2020).

Furthermore, it is important to characterise the microbiome (and other "omics") in different geographical regions (The Lancet Respiratory, 2019). In the African continent, few microbiome studies have been conducted (Ameur *et al.*, 2014; Segal *et al.*, 2017; Kaambo *et al.*, 2018; Masekela *et al.*, 2018; Roodt *et al.*, 2018). Most of these studies have focused on HIV and there are no studies on the healthy microbiome. It is important to study the microbiome of healthy individuals in different geographic regions, to determine the impact the microbial composition may have on disease, i.e. if the shift in a particular phylum is attributed to factors, such as diet or disease (Rinninella *et al.*, 2019).

However, as highlighted by Cox *et al.* (2019), one of the challenges of studying the microbiome is that the microbiome is not only composed of bacteria but also fungi and viruses. There have been several studies conducted on the lung virome, however, the fungal microbiome i.e. the mycobiome remains widely unstudied (Cui *et al.*, 2015; Su *et al.*, 2015; Tipton *et al.*, 2017; Ali *et al.*, 2019; Weaver *et al.*, 2019; Ditz *et al.*, 2020). One of the biggest challenges of mycobiome research is DNA extraction, to break the cell walls of fungi, mechanical methods are often required which will also release human DNA and will shear the DNA (Tipton *et al.*, 2017). If enzymatic methods are used instead, there may be a bias towards yeasts (Weaver *et al.*, 2019).

While, the lung microbiome has been extensively studied in COPD, in other disease states and the healthy individuals, comparative longitudinal studies are still lacking in certain areas. Some of the key areas of research that could benefit from longitudinal studies (in larger cohorts over extended periods of time e.g. 10 years or greater) are: i) the effect of smoking on the respiratory microbiome in healthy individuals, ii) the effect of ageing on the respiratory microbiome in healthy individuals, iii) the effect of treatment (including antimicrobials and corticosteroids) on the respiratory microbiome in COPD patients, iv) the effect on TB infection on the respiratory microbiome in HIV-positive individuals, COPD patients and on its own and v) the effect of HIV infection on the microbiome of individuals with COPD and without COPD, with and without TB infection, etc. (Faner *et al.*, 2017; Sulaiman *et al.*, 2020). Additionally, functional studies (using RNA-seq and/or metagenomics) need to be conducted to determine the antimicrobial resistance and virulence genes in the microbiome.

Although progress has been made in the understanding of the microbiome of the human lung in disease states, such as COPD, further research is still required.The addition of methods, such as metabolomics and transcriptomics, to studies, would allow the role of these microorganisms to be more fully elucidated and may improve our understanding of some of the disease mechanisms and microbial interactions in the future.

References

Aaron, SD (2014). Management and prevention of exacerbations of COPD. *BMJ*, **349:** g5237.

Abdool-Gaffar, MS, Calligaro, G, Wong, ML, Smith, C, Lalloo, UG, Koegelenberg, CFN, Dheda, K, Allwood, BW, Goolam-Mahomed, A & Van Zyl-Smit, RN (2019). Management of chronic obstructive pulmonary disease-a position statement of the South African thoracic society: 2019 update. *J Thorac Dis*, **11:** 4408-4427.

Ali, N, Mac Aogain, M, Morales, RF, Tiew, PY & Chotirmall, SH (2019). Optimisation and benchmarking of targeted amplicon sequencing for mycobiome analysis of respiratory specimens. *Int J Mol Sci*, **20**.

Ameur, A, Meiring, TL, Bunikis, I, Haggqvist, S, Lindau, C, Lindberg, JH, Gustavsson, I, Mbulawa, ZZ, Williamson, AL & Gyllensten, U (2014). Comprehensive profiling of the vaginal

microbiome in HIV positive women using massive parallel semiconductor sequencing. *Sci Rep,* **4:** 4398.

Auer, L, Mariadassou, M, O'Donohue, M, Klopp, C & Hernandez-Raquet, G (2017). Analysis of large 16S rRNA Illumina data sets: Impact of singleton read filtering on microbial community description. *Mol Ecol Resour,* **17:** e122-e132.

Bigna, JJ, Kenne, AM, Asangbeh, SL & Sibetcheu, AT (2018). Prevalence of chronic obstructive pulmonary disease in the global population with HIV: A systematic review and meta-analysis. *Lancet Glob Health,* **6:** e193-e202.

Caporaso, JG, Lauber, CL, Walters, WA, Berg-Lyons, D, Lozupone, CA, Turnbaugh, PJ, Fierer, N & Knight, R (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A,* **108 Suppl 1:** 4516-4522.

Carney, SM, Clemente, JC, Cox, MJ, Dickson, RP, Huang, YJ, Kitsios, GD, Kloepfer, KM, Leung, JM, Levan, TD, Molyneaux, PL, Moore, BB, O'Dwyer, DN, Segal, LN & Garantziotis, S (2020). Methods in lung microbiome research. *Am J Respir Cell Mol Biol,* **62:** 283-299.

Chang, Dela Cruz, CS & Sharma, L (2020). Challenges in understanding lung microbiome: It is not like the gut microbiome. *Respirology,* **25:** 244-245.

Charalampous, T, Kay, GL & O'Grady, J (2019). Applying clinical metagenomics for the detection and characterisation of respiratory infections. *In:* Cox, M. J., Ege, M. J. & Von Mutius, E. (eds.) The lung microbiome. Sheffield: European Respiratory Society.

Cox, MJ, Ege, MJ & Von Mutius, E (2019). The lung microbiome. Sheffield: European Respiratory Society.

Cui, L, Lucht, L, Tipton, L, Rogers, MB, Fitch, A, Kessinger, C, Camp, D, Kingsley, L, Leo, N, Greenblatt, RM, Fong, S, Stone, S, Dermand, JC, Kleerup, EC, Huang, L, Morris, A & Ghedin, E (2015). Topographic diversity of the respiratory tract mycobiome and alteration in HIV and lung disease. *Am J Respir Crit Care Med,* **191:** 932-942.

Cullen, L & Mcclean, S (2015). Bacterial adaptation during chronic respiratory infections. *Pathogens,* **4:** 66-89.

D'anna, SE, Balbi, B, Cappello, F, Carone, M & Di Stefano, A (2016). Bacterial-viral load and the immune response in stable and exacerbated COPD: Significance and therapeutic prospects. *Int J Chron Obstruct Pulmon Dis*, **11:** 445-453.

Deja, S, Porebska, I, Kowal, A, Zabek, A, Barg, W, Pawelczyk, K, Stanimirova, I, Daszykowski, M, Korzeniewska, A, Jankowska, R & Mlynarz, P (2014). Metabolomics provide new insights on lung cancer staging and discrimination from chronic obstructive pulmonary disease. *J Pharm Biomed Anal*, **100:** 369-380.

Ditz, B, Christenson, S, Rossen, J, Brightling, C, Kerstjens, HAM, Van Den Berge, M & Faiz, A (2020). Sputum microbiome profiling in COPD: Beyond singular pathogen detection. *Thorax*, **75:** 338-344.

Doring, G, Parameswaran, IG & Murphy, TF (2011). Differential adaptation of microbial pathogens to airways of patients with cystic fibrosis and chronic obstructive pulmonary disease. *FEMS Microbiol Rev*, **35:** 124-146.

Emerson, JB, Adams, RI, Roman, CMB, Brooks, B, Coil, DA, Dahlhausen, K, Ganz, HH, Hartmann, EM, Hsu, T, Justice, NB, Paulino-Lima, IG, Luongo, JC, Lymperopoulou, DS, Gomez-Silvan, C, Rothschild-Mancinelli, B, Balk, M, Huttenhower, C, Nocker, A, Vaishampayan, P & Rothschild, LJ (2017). Schrodinger's microbes: Tools for distinguishing the living from the dead in microbial ecosystems. *Microbiome*, **5:** 86.

Fabbrizzi, A, Amedei, A, Lavorini, F, Renda, T & Fontana, G (2019). The lung microbiome: Clinical and therapeutic implications. *Intern Emerg Med*, **14:** 1241-1250.

Fan, VS, Gharib, SA, Martin, TR & Wurfel, MM (2016). COPD disease severity and innate immune response to pathogen-associated molecular patterns. *Int J Chron Obstruct Pulmon Dis*, **11:** 467-477.

Faner, R, Sibila, O, Agusti, A, Bernasconi, E, Chalmers, JD, Huffnagle, GB, Manichanh, C, Molyneaux, PL, Paredes, R, Perez Brocal, V, Ponomarenko, J, Sethi, S, Dorca, J & Monso, E (2017). The microbiome in respiratory medicine: Current challenges and future perspectives. *Eur Respir J*, **49**.

Garcia-Nunez, M, Gallego, M, Monton, C, Millares, L, Pomares, X, Monso, E, Capilla, S, Espasa, M, Ferrari, R, Moya, A & Perez-Brocal, V (2018). The respiratory virome in chronic obstructive pulmonary disease. *Future Virol*, **13:** 457-466.

Global Initiative for Chronic Obstructive Lung Disease (2020). Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease (2020 report).

Ho, T, Cusack, RP, Chaudhary, N, Satia, I & Kurmi, OP (2019). Under- and over-diagnosis of COPD: A global perspective. *Breathe (Sheff)*, **15:** 24-35.

Hogan, DA, Willger, SD, Dolben, EL, Hampton, TH, Stanton, BA, Morrison, HG, Sogin, ML, Czum, J & Ashare, A (2016). Analysis of lung microbiota in bronchoalveolar lavage, protected brush and sputum samples from subjects with mild-to-moderate cystic fibrosis lung disease. *PLoS One*, **11:** e0149998.

Hufnagl, K, Pali-Scholl, I, Roth-Walter, F & Jensen-Jarolim, E (2020). Dysbiosis of the gut and lung microbiome has a role in asthma. *Semin Immunopathol*, **42:** 75-93.

Jubinville, E, Veillette, M, Milot, J, Maltais, F, Comeau, AM, Levesque, RC & Duchaine, C (2018). Exacerbation induces a microbiota shift in sputa of COPD patients. *PLoS One*, **13:** e0194355.

Kaambo, E, Africa, C, Chambuso, R & Passmore, JS (2018). Vaginal microbiomes associated with aerobic vaginitis and bacterial vaginosis. *Front Public Health*, **6:** 78.

Kim, HJ, Kim, YS, Kim, KH, Choi, JP, Kim, YK, Yun, S, Sharma, L, Dela Cruz, CS, Lee, JS, Oh, YM, Lee, SD & Lee, SW (2017). The microbiome of the lung and its extracellular vesicles in nonsmokers, healthy smokers and COPD patients. *Exp Mol Med*, **49:** e316.

Lalloo, UG, Pillay, S, Mngqibisa, R, Abdool-Gaffar, S & Ambaram, A (2016). HIV and COPD: A conspiracy of risk factors. *Respirology*, **21:** 1166-1172.

Lee, AL & Goldstein, RS (2015). Gastroesophageal reflux disease in COPD: Links and risks. *Int J Chron Obstruct Pulmon Dis*, **10:** 1935-1949.

Lee, SW, Kuan, CS, Wu, LS & Weng, JT (2016). Metagenome and metatranscriptome profiling of moderate and severe COPD sputum in Taiwanese Han males. *PLoS One*, **11:** e0159066.

Leiten, EO, Nielsen, R, Wiker, HG, Bakke, PS, Martinsen, EMH, Drengenes, C, Tangedal, S, Husebo, GR & Eagan, TML (2020). The airway microbiota and exacerbations of COPD. *ERJ Open Res*, **6**.

Masekela, R, Vosloo, S, Venter, SN, De Beer, WZ & Green, RJ (2018). The lung microbiome in children with HIV-bronchiectasis: A cross-sectional pilot study. *BMC Pulm Med*, **18:** 87.

McGuinness, AJ & Sapey, E (2017). Oxidative stress in COPD: Sources, markers, and potential mechanisms. *J Clin Med*, **6:** 21-39.

Millares, L, Perez-Brocal, V, Ferrari, R, Gallego, M, Pomares, X, Garcia-Nunez, M, Monton, C, Capilla, S, Monso, E & Moya, A (2015). Functional metagenomics of the bronchial microbiome in COPD. *PLoS One*, **10:** e0144448.

Miravitlles, M & Anzueto, A (2015). Antibiotic prophylaxis in COPD: Why, when, and for whom? *Pulm Pharmacol Ther*, **32:** 119-123.

Moffatt, MF & Cookson, WO (2017). The lung microbiome in health and disease. *Clin Med (Lond)*, **17:** 525-529.

Molyneaux, PL, Mallia, P, Cox, MJ, Footitt, J, Willis-Owen, SA, Homola, D, Trujillo-Torralbo, MB, Elkin, S, Kon, OM, Cookson, WO, Moffatt, MF & Johnston, SL (2013). Outgrowth of the bacterial airway microbiome after rhinovirus exacerbation of chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*, **188:** 1224-1231.

Naidoo, VA, Martinson, NA, Moodley, P, Joyimbana, W, Mothlaoleng, K, Abraham, P, Otwombe, K & Variava, E (2020). HIV prevalence and morbidity in older inpatients in a high HIV prevalence setting. *AIDS Res Hum Retroviruses*, **36:** 186-192.

Nobakht, MGBF, Aliannejad, R, Rezaei-Tavirani, M, Taheri, S & Oskouie, AA (2015). The metabolomics of airway diseases, including COPD, asthma and cystic fibrosis. *Biomarkers*, **20:** 5-16.

Noell, K & Kolls, JK (2019). Further defining the human virome using NGS: Identification of *Redondoviridae*. *Cell Host Microbe*, **25:** 634-635.

North, CM, Valeri, L, Hunt, PW, Mocello, AR, Martin, JN, Boum, Y, 2nd, Haberer, JE, Bangsberg, DR, Christiani, DC & Siedner, MJ (2017). Cooking fuel and respiratory symptoms among people living with HIV in rural Uganda. *ERJ Open Res,* **3**.

Park, CH, Seo, SI, Kim, JS, Kang, SH, Kim, BJ, Choi, YJ, Byun, HJ, Yoon, JH & Lee, SK (2020). Treatment of non-erosive reflux disease and dynamics of the esophageal microbiome: A prospective multicenter study. *Sci Rep,* **10:** 15154.

Pavord, ID, Jones, PW, Burgel, PR & Rabe, KF (2016). Exacerbations of COPD. *Int J Chron Obstruct Pulmon Dis,* **11 Spec Iss:** 21-30.

Pefura-Yone, EW, Fodjeu, G, Kengne, AP, Roche, N & Kuaban, C (2015). Prevalence and determinants of chronic obstructive pulmonary disease in HIV infected patients in an African country with low level of tobacco smoking. *Respir Med,* **109:** 247-254.

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

Quince, C, Walker, AW, Simpson, JT, Loman, NJ & Segata, N (2017). Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol,* **35:** 833-844.

Rinninella, E, Raoul, P, Cintoni, M, Franceschi, F, Miggiano, GAD, Gasbarrini, A & Mele, MC (2019). What is the healthy gut microbiota composition? A changing ecosystem across age, environment, diet, and diseases. *Microorganisms,* **7**.

Rogers, GB (2017). The lung microbiome. *Emerging Topics in Life Sciences,* **1:** 313-324.

Roodt, AP, Naude, Y, Stoltz, A & Rohwer, E (2018). Human skin volatiles: Passive sampling and GCxGC-ToFMS analysis as a tool to investigate the skin microbiome and interactions with anthropophilic mosquito disease vectors. *J Chromatogr B Analyt Technol Biomed Life Sci,* **1097-1098:** 83-93.

Salvi, S (2015). The silent epidemic of COPD in Africa. *The Lancet Global Health,* **3:** e6-e7.

Sanchez, J, Schumann, DM, Karakioulaki, M, Papakonstantinou, E, Rassouli, F, Frasnelli, M, Brutsche, M, Tamm, M & Stolz, D (2020). Laryngopharyngeal reflux in chronic obstructive pulmonary disease - a multi-centre study. *Respir Res,* **21:** 220.

Segal, LN, Clemente, JC, Li, Y, Ruan, C, Cao, J, Danckers, M, Morris, A, Tapyrik, S, Wu, BG, Diaz, P, Calligaro, G, Dawson, R, Van Zyl-Smit, RN, Dheda, K, Rom, WN & Weiden, MD (2017). Anaerobic bacterial fermentation products increase tuberculosis risk in antiretroviral-drug-treated HIV patients. *Cell Host Microbe,* **21:** 530-537 e534.

Shi, M, Zhang, YZ & Holmes, EC (2018). Meta-transcriptomics and the evolutionary biology of RNA viruses. *Virus Res,* **243:** 83-90.

Shimizu, K, Yoshii, Y, Morozumi, M, Chiba, N, Ubukata, K, Uruga, H, Hanada, S, Saito, N, Kadota, T, Ito, S, Wakui, H, Takasaka, N, Minagawa, S, Kojima, J, Hara, H, Numata, T, Kawaishi, M, Saito, K, Araya, J, Kaneko, Y, Nakayama, K, Kishi, K & Kuwano, K (2015). Pathogens in COPD exacerbations identified by comprehensive real-time PCR plus older methods. *Int J Chron Obstruct Pulmon Dis,* **10:** 2009-2016.

Sin, DD (2018). Precision health in COPD: Now is the time. *CTSJ* **2:** 128-132.

Su, J, Liu, HY, Tan, XL, Ji, Y, Jiang, YX, Prabhakar, M, Rong, ZH, Zhou, HW & Zhang, GX (2015). Sputum bacterial and fungal dynamics during exacerbations of severe COPD. *PLoS One,* **10:** e0130736.

Sulaiman, I, Schuster, S & Segal, LN (2020). Perspectives in lung microbiome research. *Curr Opin Microbiol,* **56:** 24-29.

Sun, S, Jones, RB & Fodor, AA (2020). Inference-based accuracy of metagenome prediction tools varies across sample types and functional categories. *Microbiome,* **8:** 46.

Sze, MA, Dimitriu, PA, Hayashi, S, Elliott, WM, Mcdonough, JE, Gosselink, JV, Cooper, J, Sin, DD, Mohn, WW & Hogg, JC (2012). The lung tissue microbiome in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med,* **185:** 1073-1080.

Sze, MA & Schloss, PD (2019). The impact of DNA polymerase and number of rounds of amplification in PCR on 16S rRNA gene sequence data. *mSphere,* **4**.

Tennert, C, Reinmuth, AC, Bremer, K, Al-Ahmad, A, Karygianni, L, Hellwig, E, Vach, K, Ratka-Kruger, P, Wittmer, A & Woelber, JP (2020). An oral health optimized diet reduces the load of potential cariogenic and periodontal bacterial species in the supragingival oral plaque: A randomized controlled pilot study. *Microbiologyopen,* **9:** e1056.

The Lancet Respiratory, M (2019). Harnessing the microbiome for lung health. *The Lancet Respiratory Medicine*, **7:** 827.

Tipton, L, Ghedin, E & Morris, A (2017). The lung mycobiome in the next-generation sequencing era. *Virulence*, **8:** 334-341.

Twigg, HL, 3rd, Weinstock, GM & Knox, KS (2017). Lung microbiome in human immunodeficiency virus infection. *Transl Res*, **179:** 97-107.

van Gemert, F, Van Der Molen, T, Jones, R & Chavannes, N (2011). The impact of asthma and COPD in sub-Saharan Africa. *Prim Care Respir J*, **20:** 240-248.

van Rijn, AL, Van Boheemen, S, Sidorov, I, Carbo, EC, Pappas, N, Mei, H, Feltkamp, M, Aanerud, M, Bakke, P, Claas, ECJ, Eagan, TM, Hiemstra, PS, Kroes, ACM & De Vries, JJC (2019). The respiratory virome and exacerbations in patients with chronic obstructive pulmonary disease. *PLoS One*, **14:** e0223952.

Varkila, MRJ, Vos, AG, Barth, RE, Tempelman, HA, Deville, WLJ, Coutinho, RA, Grobbee, DE & Klipstein-Grobusch, K (2019). The association between HIV infection and pulmonary function in a rural African population. *PLoS One*, **14:** e0210573.

Weaver, D, Gago, S, Bromley, M & Bowyer, P (2019). The human lung mycobiome in chronic respiratory disease: Limitations of methods and our current understanding. *Current Fungal Infection Reports*, **13:** 109-119.

APPENDIX A

REAGENTS, BUFFERS AND GELS USED IN EXPERIMENTAL PROCEDURES

1. 0.1% Dithiothreitol (DTT)

DTT powder (Roche, Switzerland;10197777001)	270	mg
Phosphate buffered saline (PBS) (pH 7.4) (ThermoFisher, USA)	270	mL

Dissolve 10 mg DTT in 10 mL PBS. Sterilise using 0.45 µM filter and syringe.

2. Ethylene diamine tetra-acetate (EDTA) (0.5 M; pH 8.0) (Green and Sambrook, 2012)

EDTA, disodium salt (Sigma-Aldrich, USA)	93.05	g
Ultrapure water	400	mL
Sodium hydroxide (NaOH) pellets (Merckmillipore, USA)		

Dissolve 93.05 g EDTA in 400 mL ultrapure water, adding the NaOH pellets until the solution becomes clear. Bring the volume to 500 mL and autoclave at 121°C for 15 min.

3. Tris (1 M; pH 8.0) ("Tris-HCl", 2006)

Tris-base (Sigma-Aldrich, USA)	60.55	g
Hydrochloric acid (HCl) (Merckmillipore, USA)	21	mL
Ultrapure water	400	mL

Dissolve 60.55 g Tris-base in ultrapure water. Add the 21 mL of HCl and mix the solution. Bring the volume to 500 mL and autoclave at 121°C for 15 min.

4. TE buffer (10 mM Tris: 1 mM EDTA; pH 8.0) ("Tris-EDTA buffer", 2009)

1 M Tris (pH 8.0)	1	mL
0.5 M EDTA	0.2	mL
Ultrapure water	0.8	mL

Dissolve 1 mL of Tris and 0.2 mL of EDTA in 0.8 mL of ultrapure water. Adjust the pH and bring the volume to 100 mL. Autoclave at 121°C for 15 min.

5. Lysostaphin (3 U/µL)

Lysostaphin (Sigma-Aldrich, USA; L9043)	5	mg
Nuclease-free water	5	mL

Dissolve 5 mg lysostaphin in 5 mL nuclease-free water on ice. Once completely dissolved, aliquot into working solution and freeze at -20°C.

6. Lysozyme (10 mg/mL stock solution)

Lysozyme (Sigma-Aldrich, USA; L6876-5g)	10	mg
Nuclease-free water	1	mL

Dissolve 10 mg lysozyme in 1 mL nuclease-free water on ice. Once completely dissolved, aliquot into working solution and freeze at -20°C.

7. Mutanolysin (10 U/µL)

Mutanolysin (Sigma-Aldrich, USA; M9901-1KU)	1000	U
Nuclease-free water	100	µL

Add 100 µL nuclease-free water to mutanolysin (1000 U) on ice. Once completely dissolved, aliquot into working solution and freeze at -20°C.

8. Tris-boric EDTA (TBE) buffer, 1X (Green and Sambrook, 2012)

10X Tris-boric EDTA (TBE) buffer (Thermofisher, USA)	100	mL
Ultrapure water	900	mL

Add 900 mL ultrapure water to 100 mL of 10X TBE.

9. Hard-to-lyse buffer (20 mM Tris, 2 mM EDTA, 1% Triton X-100)

1M Tris (pH 8.0)	1	mL
Triton X-100 (Amresco, USA)	0.5	mL
0.5M EDTA (pH 8.0)	0.2	mL

Add 0.2 mL Triton X-100 and 0.2 mL EDTA to 1 mL Tris. Make up solution to 50 mL. Autoclave at 121°C for 15 min.

10. Agarose gel (1.5%)

SeaKem LE agarose powder (Lonza, USA)	1.5	g
1X TBE buffer	100	mL
Ethidium bromide [10 mg/ml (Sigma-Aldrich, USA)]	5	µL

Add 1.5 g of SeaKem LE agarose powder to 100 mL 1X TBE buffer. Microwave the solution on medium heat for 2 to 3 min, stopping to swirl the solution at intervals. Allow to cool down to 50°C and add 5 µL ethidium bromide. Pour into clean casting tray, add a comb, and allow to set for 30 min.

References

Green, MR & Sambrook, J (2012). Molecular cloning: A laboratory manual, Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory Press.

Tris-EDTA buffer. (2009). *Cold Spring Harbor Protocols*, **2009:** pdb.rec11661.

Tris-HCl. (2006). *Cold Spring Harbor Protocols*, **2006:** pdb.rec8747.

APPENDIX B

EXPERIMENTAL PROCEDURES

1. Pre-processing of sputum specimens

1. Add an equal volume of 0.1% dithiothreitol (DTT) (Roche, Switzerland) to the sputum sample i.e. the volume of DTT should be the same as the volume of the sputum sample (Hamid *et al.*, 2002; Allen *et al.*, 2016; Park *et al.*, 2018; Terranova *et al.*, 2018).
2. Vortex (Vortex-Genie® 2, Scientific Industries Inc., USA) the samples for 30 seconds (sec) and leave at room temperature (+/- 25°C) for 15 min.
3. Divide the samples into aliquots for bacterial DNA extraction, viral DNA extraction and RNA extraction.

2. Bacterial deoxyribonucleic acid (DNA) extraction

1. Centrifuge (Spectrafuge™ 24D, Labnet International Inc., USA) at 4 000 x *g* for 30 min (Bioline II Genomic DNA Manual; section 9.2 sample preparation). Discard the supernatant.
2. Resuspend in 193.95 µL hard to lyse buffer (20 mM Tris (Sigma-Aldrich, USA) pH 8.0; 1% Triton X-100 (Amresco, USA); 2 mM EDTA(Sigma-Aldrich, USA)).
3. Add 22.5 µL lysozyme (10 mg/ml, Sigma-Aldrich), 6.75 µL mutanolysin (25 KU/ml, Sigma-Aldrich), and 1.8 µL lysostaphin (4000 U/mL, Sigma-Aldrich) to a 500 µL aliquot of the cell suspension (Yuan *et al.*, 2012).
4. Vortex (Vortex-Genie® 2, Scientific Industries Inc., USA) briefly and incubate for 1 hour at 37°C (AccuBlock™ Digital Dry Bath, Labnet International Inc., USA).
5. Add 25 µL Proteinase K (included in the kit), vortex (Vortex-Genie® 2, Scientific Industries Inc., USA) briefly and incubate at 56°C (AccuBlock™ Digital Dry Bath, Labnet International Inc., USA) for 2 hours.
6. Vortex (Vortex-Genie® 2, Scientific Industries Inc., USA) each sample briefly (30 sec) and add 200 µL of Lysis buffer G3.
7. Vortex (Vortex-Genie® 2; Scientific Industries Inc., USA) and incubate (AccuBlock™ Digital Dry Bath, Labnet International Inc., USA) at 70°C for 10 min.
8. Vortex (Vortex-Genie® 2, Scientific Industries Inc., USA) and add 210 µL ethanol (96-100%). Vortex (Vortex-Genie® 2, Scientific Industries Inc., USA) again.
9. Place Spin column in a collection tube.

10. Add all sample to the column and centrifuge (Spectrafuge™ 24D, Labnet International Inc., USA) at 11 000 x *g* for 1 min.
11. Discard flow-through (but keep collection tube).
12. Add 500 µL Wash buffer GW1.
13. Centrifuge (Spectrafuge™ 24D, Labnet International Inc., USA) at 11 000 x *g* for 1 min.
14. Discard flow-through (but keep collection tube).
15. Add 600 µL Wash buffer GW2.
16. Centrifuge (Spectrafuge™ 24D, Labnet International Inc., USA) at 11 000 x *g* for 1 min.
17. Discard flow-through (but keep collection tube).
18. Centrifuge (Spectrafuge™ 24D, Labnet International Inc., USA) at 11 000 x *g* for 1 min.
19. Place Spin column in a microcentrifuge tube (Axygen, Corning, Germany).
20. Add 100 µL Elution buffer G (70°C) directly onto silica membrane.
21. Incubate at room temperature (+/- 25°C) for 1 min.
22. Centrifuge (Spectrafuge™ 24D, Labnet International Inc., USA) at 11 000 x *g* for 1 min.

3. Viral deoxyribonucleic acid (DNA) extraction

1. Treat the viral DNA aliquot with 10 U/mL TURBO™ DNase (Ambion, USA) at 37°C for 30 min (AccuBlock™ Digital Dry Bath, Labnet International Inc., USA).
2. Inactive the DNase with 15 mM ethylenediaminetetraacetic acid (EDTA) (Sigma-Aldrich, USA) at 75°C for 10 min (AccuBlock™ Digital Dry Bath, Labnet International Inc., USA) according to manufacturer's instructions (de la Cruz Pena *et al.*, 2018).
3. Centrifuge (Spectrafuge™ 24D, Labnet International Inc., USA) at 4 000 x *g* for 30 min (Bioline II Genomic DNA Manual; section 9.13 sample preparation). Discard the supernatant.
4. Transfer 200 µL to a new microcentrifuge tube (Axygen, Corning, Germany).
5. Add 180 µL of lysis buffer GL and 25 µL of Proteinase K (provided in the kit) to the solution, vortex (Vortex-Genie® 2, Scientific Industries Inc., USA) briefly and incubate at 56°C (AccuBlock™ Digital Dry Bath, Labnet International Inc., USA) for 2 hours.
6. Vortex (Vortex-Genie® 2, Scientific Industries Inc., USA) each sample briefly (30 sec) and add 200 µL Lysis buffer G3.

7. Vortex (Vortex-Genie® 2, Scientific Industries Inc., USA) and incubate (AccuBlock™ Digital Dry Bath, Labnet International Inc., USA) at 70°C for 10 min.
8. Vortex (Vortex-Genie® 2, Scientific Industries Inc., USA) and add 210 µL ethanol (96-100%). Vortex (Vortex-Genie® 2, Scientific Industries Inc., USA) again.
9. Place Spin column in a collection tube.
10. Add all sample to the column and centrifuge (Spectrafuge™ 24D, Labnet International Inc., USA) at 11 000 x *g* for 1 min.
11. Discard flow-through (but keep collection tube).
12. Add 500 µL Wash buffer GW1.
13. Centrifuge (Spectrafuge™ 24D, Labnet International Inc., USA) at 11 000 x *g* for 1 min.
14. Discard flow-through (but keep collection tube).
15. Add 600 µL Wash buffer GW2.
16. Centrifuge (Spectrafuge™ 24D, Labnet International Inc., USA) at 11 000 x *g* for 1 min.
17. Discard flow-through (but keep collection tube).
18. Centrifuge (Spectrafuge™ 24D, Labnet International Inc., USA) at 11 000 x *g* for 1 min.
19. Place Spin column in a microcentrifuge tube (Axygen, Corning, Germany).
20. Add 100 µL Elution buffer G (70°C) directly onto silica membrane.
21. Incubate at room temperature (+/- 25°C) for 1 min.
22. Centrifuge (Spectrafuge™ 24D, Labnet International Inc., USA) at 11 000 x *g* for 1 min.

4. Viral ribonucleic acid (RNA) extraction

1. The viral RNA aliquot was treated with 10 U/mL TURBO™ DNase (Ambion, USA) at 37°C for 30 min (AccuBlock™ Digital Dry Bath, Labnet International Inc., USA).
2. This was followed by inactivation with 15 mM ethylenediaminetetraacetic acid (EDTA) (Sigma-Aldrich, USA) at 75°C for 10 min (AccuBlock™ Digital Dry Bath, Labnet International Inc., USA) according to manufacturer's instructions (de la Cruz Pena *et al.*, 2018).
3. Add 560 µL Buffer AVL (containing carrier RNA) to 1.5 mL microcentrifuge tube.
4. Add 140 µL of sample to the tube and vortex (Vortex-Genie® 2, Scientific Industries Inc., USA) for 15 sec.

5. Incubate at room temperature (+/- 25°C) for 10 min.
6. Centrifuge (Spectrafuge™ 24D, Labnet International Inc., USA) the tube at 4 000 x *g* for 30 sec.
7. Add 560 µL of ethanol (96% to100%) and vortex (Vortex-Genie® 2, Scientific Industries Inc., USA) for 15 sec.
8. Centrifuge (Spectrafuge™ 24D, Labnet International Inc., USA) the tube at 4 000 x *g* for 30 sec.
9. Add 630 µL of the solution to the column (in a collection tube). Do not wet the rim.
10. Centrifuge (Spectrafuge™ 24D, Labnet International Inc., USA) at 6 000 x *g* for 1 min.
11. Place column in a new collection tube and add 500 µL Buffer AW1.
12. Centrifuge (Spectrafuge™ 24D, Labnet International Inc., USA) at 6 000 x *g* for 1 min.
13. Place column in a new collection tube and add 500 µL Buffer AW2.
14. Centrifuge (Spectrafuge™ 24D, Labnet International Inc., USA) at 16 000 x *g* (full speed) for 3 min.
15. Place column in a new microcentrifuge tube (Axygen, Corning, Germany) and add 60 µL of Buffer AVE [at room temperature(+/- 25°C)].
16. Incubate at room temperature (+/- 25°C) for 1 min.
17. Centrifuge (Spectrafuge™ 24D, Labnet International Inc., USA) at 6 000 x *g* for 1 min.

5. cDNA synthesis

1. For each reaction (i.e. each sample):

Component	X1
dNTPs (10 mM mix)	1 µL
random hexamers primers (50 ng/ µL)	1 µL
nuclease-free water	2 µL
RNA	6 µL

2. Incubate (Bio-rad T100™ Thermal cycle, Bio-rad Laboratories Inc., USA) at 65°C for 5 min.
3. Incubate on ice for 1 min.
4. In a different (new) tube, add the following components:

Component	X1
10X RT buffer	2 µL
25 mM magnesium chloride ($MgCl_2$),	4 µL
0.1 M DTT	2 µL
RNaseOUT™ (40 U/µL)	1 µL

5. Add 9 µL to this mixture to each RNA/primer mix from step 3 and incubate at room temperature (+/- 25°C) for 2 min.
6. Add 1 µL of SuperScript™ II RT to each tube (except for the RT control; to the RT control, add 1 µL nuclease-free water) and incubate at room temperature (+/- 25°C) for 10 min.
7. Incubate (Bio-rad T100™ Thermal cycle, Bio-rad Laboratories Inc., USA) at 42°C for 50 min.
8. Terminate the reaction at 70°C (Bio-rad T100™ Thermal cycle, Bio-rad Laboratories Inc., USA) for 15 min. Chill on ice.
9. Add 1 µL of RNase H to each tube and incubate at 37°C (Bio-rad T100™ Thermal cycle, Bio-rad Laboratories Inc., USA) for 20 min.
10. For second strand synthesis: in a different (new) tube, add the following components (Kufner *et al.*, 2019; Wu *et al.*, 2019):

Component	X1
Klenow fragment (5 U)	1 µL
cDNA	20 µL
Buffer	4 µL

11. Add to each sample and incubate at 37°C (Bio-rad T100™ Thermal cycle, Bio-rad Laboratories Inc., USA) for 1 hour. Incubate at 72°C for 10 min.

6. Amplification of cDNA and ssDNA

1. Combine the following components in a tube:

Component	X1
KAPA HiFi	12.5 µL
DNA/cDNA	2.5 µL
FR20RV primer (40 pmol)	1 µL
Nuclease-free water	9 µ

2. Run PCR amplification (Bio-rad T100™ Thermal cycle, Bio-rad Laboratories Inc., USA) as follows: initial denaturation at 95°C for 5 min, 40 cycles of 98°C for 1 min, 65°C for 1 min, and 72°C for 2 min, followed by a final extension step at 72°C for 1 min.

7. Polymerase chain reaction (PCR) for the IS-Pro method

1. Add 12 µL of PROTEO master mix in a well (in a microtiter plate) for each sample as well as for the positive and negative controls.

2. Add 12 µL of FIRBAC master mix in separate wells (in a microtiter plate) for each sample as well as for the positive and negative controls.
3. Add 8 µL of extracted bacterial DNA to each PROTEO and each FIRBAC well.
4. Run PCR amplification (AB GeneAmp PCR 9700, Thermofisher, USA) as follows: 35 cycles of 94°C for 30 s, 56°C for 45 s, and 72°C for 1 min, followed by a final extension step at 72°C for 10 min.
5. After amplification, add 16 µL of the eMix to each well (for the number of samples and controls) in a new microtiter plate and add 4 µL of each amplicon to a well, followed by denaturation at 94°C for 3 min.
6. Analyse the samples on the 3730xL genetic analyser (Thermofisher, USA) at the central analytical facility (CAF) at Stellenbosch University, Cape Town, South Africa.

References

Allen, V, Nicol, MP & Ah Tow, L (2016). Sputum processing prior to *Mycobacterium tuberculosis* detection by culture or nucleic acid amplification testing: A narrative review. **5:** 96-109.

de La Cruz Pena, MJ, Martinez-Hernandez, F, Garcia-Heredia, I, Lluesma Gomez, M, Fornas, O & Martinez-Garcia, M (2018). Deciphering the human virome with single-virus genomics and metagenomics. *Viruses*, **10**.

Hamid, Q, Kelly, MM, Linden, M, Louis, R, Pizzichini, MMM, Pizzichini, E, Ronchi, C, Van Overveld, F & Djukanovic, R (2002). Methods of sputum processing for cell counts, immunocytochemistry and *in situ* hybridisation. *Eur Respir J*, **20:** 19S-23s.

Kufner, V, Plate, A, Schmutz, S, Braun, DL, Gunthard, HF, Capaul, R, Zbinden, A, Mueller, NJ, Trkola, A & Huber, M (2019). Two years of viral metagenomics in a tertiary diagnostics unit: Evaluation of the first 105 cases. *Genes (Basel)*, **10**.

Park, HJ, Woo, A, Cha, JM, Lee, KS & Lee, MY (2018). Closed-type pre-treatment device for point-of-care testing of sputum. *Sci Rep*, **8:** 16508.

Terranova, L, Oriano, M, Teri, A, Ruggiero, L, Tafuro, C, Marchisio, P, Gramegna, A, Contarini, M, Franceschi, E, Sottotetti, S, Cariani, L, Bevivino, A, Chalmers, JD, Aliberti, S & Blasi, F (2018). How to process sputum samples and extract bacterial DNA for microbiota analysis. *Int J Mol Sci*, **19:** 3256-3568.

Wu, Q, Peng, D, Liu, Q, Shabbir, MAB, Sajid, A, Liu, Z, Wang, Y & Yuan, Z (2019). A novel microbiological method in microtiter plates for screening seven kinds of widely used antibiotics residues in milk, chicken egg and honey. *Frontiers in Microbiology*, **10**.

Yuan, S, Cohen, DB, Ravel, J, Abdo, Z & Forney, LJ (2012). Evaluation of methods for the extraction and purification of DNA from the human microbiome. *PLoS One*, **7:** e33865.

APPENDIX C

JOURNAL GUIDELINES AND REQUIREMENTS

1. MICROBIOME JOURNAL (BMC JOURNAL; PART OF SPRINGER NATURE GROUP)

A. Criteria

AVAILABILITY OF DATA, METADATA AND ANALYTICAL SCRIPTS

At *Microbiome* we are striving to make reproducibility a priority. Data availability at time of submission is a key aspect to this process as it allows reviewers to fully evaluate your work.
Microbiome follows a strict data release policy (Research Data Policy Type 4). We require that all datasets on which the conclusions of the paper rely should be available to the reviewers and readers. We ask that authors make sure their datasets are either deposited in publicly available repositories (where available and appropriate) or presented in the main manuscript or additional supporting files whenever possible. Accompanying metadata must be available in the repository or as supporting files to the manuscript. Metadata should be formatted according to the MixS (Minimum Information about any (x) Sequence) standards developed by the Genome Standards Consortium (GSC). Template can be found here: http://gensc.org/mixs/). The sample identifiers in the repository must refer to the same sample identifiers used in the manuscript. Please see Springer Nature's information on recommended repositories.
We are also requiring that authors make the code/scripts used for their analysis available as knitr files, iPython Notebooks, or any other formats they might find suitable. Again, this effort encourages transparency and complete reproducibility of your study. A good example is a paper published in *Microbiome* by Meadow et al.
Please include the live accession number, or similar, in a section entitled "Availability of data and materials".

USE OF EXPERIMENTAL CONTROLS

As with reproducibility, at *Microbiome* we are striving to publish high quality study, and we believe that the use of experimental controls is critical to guarantee quality and credibility. We expect that studies include controls, especially when analyzing samples believed to carry a very low bacterial/fungal/viral biomass. Materials and reagent, experimental sampling and processing can introduce contamination (DNA or cells) that if not controlled would jeopardize the integrity and quality of a study. Thus, we expect that studies submitted to *Microbiome* include sampling controls, extraction controls, PCR amplification controls as negative controls, but also positive controls (mock communities or others). These controls should be sequenced, and the sequence data reported in the paper and made available along with the sample sequence data in a public repository.

NOMENCLATURE OF ORGANISMS

Bacterial names should be written according to the guidelines of the American Society for Microbiology and the Journal of Bacteriology. Essentially, the names of all microbial taxa (kingdom, phyla, class, order, family, genus, species, and subspecies) should be italicized in the manuscript and the figures. Do not italicize strain designations or numbers.

TERMINOLOGY TO DESCRIBE MICROBIOME STUDIES

At *Microbiome* we have decided to follow the recommendations of Marchesi et al. with regards to vocabulary used to describe different aspects of microbial communities and their environments.
A common example is the use of the term 16S,16S rDNA, 16S rDNA gene, 16S gene which are not appropriate. These should be replaced with 16S rRNA gene.
Please make sure that you comply with all these criteria.

Preparing your Manuscript

The information below details the section headings that you should include in your manuscript and what information should be within each section.
Please note that your manuscript must include a 'Declarations' section including all of the subheadings (please see below for more information).

Title page

The title page should:

- present a title that includes, if appropriate, the study design
- list the full names, institutional addresses and email addresses for all authors

○ if a collaboration group should be listed as an author, please list the Group name as an author. If you would like the names of the individual members of the Group to be searchable through their individual PubMed records, please include this information in the "Acknowledgements" section in accordance with the instructions below

- indicate the corresponding author

Abstract

The Abstract should not exceed 350 words. Please minimize the use of abbreviations and do not cite references in the abstract. The abstract must include the following separate sections:

- **Background:** the context and purpose of the study
- **Results:** the main findings
- **Conclusions:** a brief summary and potential implications

Keywords

Three to ten keywords representing the main content of the article.

Background

The Background section should explain the background to the study, its aims, a summary of the existing literature and why this study was necessary.

Methods

The methods section should include:

- the aim, design and setting of the study
- the characteristics of participants or description of materials
- a clear description of all processes, interventions and comparisons. Generic names should generally be used. When proprietary brands are used in research, include the brand names in parentheses
- the type of statistical analysis used, including a power calculation if appropriate

Results

This should include the findings of the study including, if appropriate, results of statistical analysis which must be included either in the text or as tables and figures.

Discussion

For research articles this section should discuss the implications of the findings in context of existing research and highlight limitations of the study. For study protocols and methodology manuscripts this section should include a discussion of any practical or operational issues involved in performing the study and any issues not covered in other sections.

Conclusions

This should state clearly the main conclusions and provide an explanation of the importance and relevance of the study to the field.

List of abbreviations

If abbreviations are used in the text they should be defined in the text at first use, and a list of abbreviations can be provided.

Declarations

All manuscripts must contain the following sections under the heading 'Declarations':

- Ethics approval and consent to participate
- Consent for publication
- Availability of data and material
- Competing interests
- Funding
- Authors' contributions
- Acknowledgements
- Authors' information (optional)

Please see below for details on the information to be included in these sections.

If any of the sections are not relevant to your manuscript, please include the heading and write 'Not applicable' for that section.

Ethics approval and consent to participate

Manuscripts reporting studies involving human participants, human data or human tissue must:

- include a statement on ethics approval and consent (even where the need for approval was waived)
- include the name of the ethics committee that approved the study and the committee's reference number if appropriate

Studies involving animals must include a statement on ethics approval.
See our editorial policies for more information.
If your manuscript does not report on or involve the use of any animal or human data or tissue, please state "Not applicable" in this section.

Consent for publication

If your manuscript contains any individual person's data in any form (including any individual details, images or videos), consent for publication must be obtained from that person, or in the case of children, their parent or legal guardian. All presentations of case reports must have consent for publication.
You can use your institutional consent form or our consent form if you prefer. You should not send the form to us on submission, but we may request to see a copy at any stage (including after publication).
See our editorial policies for more information on consent for publication.
If your manuscript does not contain data from any individual person, please state "Not applicable" in this section.

Availability of data and materials

All manuscripts must include an 'Availability of data and materials' statement. Data availability statements should include information on where data supporting the results reported in the article can be found including, where applicable, hyperlinks to publicly archived datasets analysed or generated during the study. By data we mean the minimal dataset that would be necessary to interpret, replicate and build upon the findings reported in the article. We recognise it is not always possible to share research data publicly, for instance when individual privacy could be compromised, and in such instances data availability should still be stated in the manuscript along with any conditions for access.
Data availability statements can take one of the following forms (or a combination of more than one if required for multiple datasets):

- The datasets generated and/or analysed during the current study are available in the [NAME] repository, [PERSISTENT WEB LINK TO DATASETS]
- All data generated or analysed during this study are included in this published article [and its supplementary information files].
- Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.
- Not applicable. If your manuscript does not contain any data, please state 'Not applicable' in this section.

More examples of template data availability statements, which include examples of openly available and restricted access datasets, are available here.
BioMed Central also requires that authors cite any publicly available data on which the conclusions of the paper rely in the manuscript. Data citations should include a persistent identifier (such as a DOI) and should ideally be included in the reference list. Citations of datasets, when they appear in the reference list, should include the minimum information recommended by DataCite and follow journal style. Dataset identifiers including DOIs should be expressed as full URLs. For example:

Hao Z, AghaKouchak A, Nakhjiri N, Farahmand A. Global integrated drought monitoring and prediction system (GIDMaPS) data sets. figshare. 2014. http://dx.doi.org/10.6084/m9.figshare.853801
With the corresponding text in the Availability of data and materials statement:
The datasets generated during and/or analysed during the current study are available in the [NAME] repository, [PERSISTENT WEB LINK TO DATASETS].[Reference number]

Competing interests

All financial and non-financial competing interests must be declared in this section.
See our editorial policies for a full explanation of competing interests. If you are unsure whether you or any of your co-authors have a competing interest please contact the editorial office.
Please use the authors initials to refer to each authors' competing interests int his section.

If you do not have any competing interests, please state "The authors declare that they have no competing interests" in this section.

Funding

All sources of funding for the research reported should be declared. The role of the funding body in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript should be declared.

Authors' contributions

The individual contributions of authors to the manuscript should be specified in this section. Guidance and criteria for authorship can be found in our editorial policies.
Please use initials to refer to each author's contribution in this section, for example: "FC analyzed and interpreted the patient data regarding the hematological disease and the transplant. RH performed the histological examination of the kidney, and was a major contributor in writing the manuscript. All authors read and approved the final manuscript."

Acknowledgements

Please acknowledge anyone who contributed towards the article who does not meet the criteria for authorship including anyone who provided professional writing services or materials.
Authors should obtain permission to acknowledge from all those mentioned in the Acknowledgements section.
See our editorial policies for a full explanation of acknowledgements and authorship criteria.
If you do not have anyone to acknowledge, please write "Not applicable" in this section.
Group authorship (for manuscripts involving a collaboration group): if you would like the names of the individual members of a collaboration Group to be searchable through their individual PubMed records, please ensure that the title of the collaboration Group is included on the title page and in the submission system and also include collaborating author names as the last paragraph of the “Acknowledgements” section. Please add authors in the format First Name, Middle initial(s) (optional), Last Name. You can add institution or country information for each author if you wish, but this should be consistent across all authors.
Please note that individual names may not be present in the PubMed record at the time a published article is initially included in PubMed as it takes PubMed additional time to code this information.

Authors' information

This section is optional.
You may choose to use this section to include any relevant information about the author(s) that may aid the reader's interpretation of the article, and understand the standpoint of the author(s). This may include details about the authors' qualifications, current positions they hold at institutions or societies, or any other relevant background information. Please refer to authors using their initials. Note this section should not be used to describe any competing interests.

Endnotes

Endnotes should be designated within the text using a superscript lowercase letter and all notes (along with their corresponding letter) should be included in the Endnotes section. Please format this section in a paragraph rather than a list.

References

All references, including URLs, must be numbered consecutively, in square brackets, in the order in which they are cited in the text, followed by any in tables or legends. The reference numbers must be finalized and the reference list fully formatted before submission.
Examples of the BioMed Central reference style are shown below. Please ensure that the reference style is followed precisely.
See our editorial policies for author guidance on good citation practice.
Web links and URLs: All web links and URLs, including links to the authors' own websites, should be given a reference number and included in the reference list rather than within the text of the manuscript. They should be provided in full, including both the title of the site and the URL, as well as the date the site was accessed, in the following format: The Mouse Tumor Biology Database. http://tumor.informatics.jax.org/mtbwi/index.do. Accessed 20 May 2013. If an author or group of authors can clearly be associated with a web link (e.g. for blogs) they should be included in the reference.

Example reference style:

Article within a journal
Smith JJ. The world of science. Am J Sci. 1999;36:234-5.
Article within a journal (no page numbers)
Rohrmann S, Overvad K, Bueno-de-Mesquita HB, Jakobsen MU, Egeberg R, Tjønneland A, et al. Meat consumption and mortality - results from the European Prospective Investigation into Cancer and Nutrition. BMC Med. 2013;11:63.
Article within a journal by DOI
Slifka MK, Whitton JL. Clinical implications of dysregulated cytokine production. Dig J Mol Med. 2000; doi:10.1007/s801090000086.
Article within a journal supplement
Frumin AM, Nussbaum J, Esposito M. Functional asplenia: demonstration of splenic activity by bone marrow scan. Blood 1979;59 Suppl 1:26-32.
Book chapter, or an article within a book
Wyllie AH, Kerr JFR, Currie AR. Cell death: the significance of apoptosis. In: Bourne GH, Danielli JF, Jeon KW, editors. International review of cytology. London: Academic; 1980. p. 251-306.
OnlineFirst chapter in a series (without a volume designation but with a DOI)
Saito Y, Hyuga H. Rate equation approaches to amplification of enantiomeric excess and chiral symmetry breaking. Top Curr Chem. 2007. doi:10.1007/128_2006_108.
Complete book, authored
Blenkinsopp A, Paxton P. Symptoms in the pharmacy: a guide to the management of common illness. 3rd ed. Oxford: Blackwell Science; 1998.
Online document
Doe J. Title of subordinate document. In: The dictionary of substances and their effects. Royal Society of Chemistry. 1999. http://www.rsc.org/dose/title of subordinate document. Accessed 15 Jan 1999.
Online database
Healthwise Knowledgebase. US Pharmacopeia, Rockville. 1998. http://www.healthwise.org. Accessed 21 Sept 1998.
Supplementary material/private homepage
Doe J. Title of supplementary material. 2000. http://www.privatehomepage.com. Accessed 22 Feb 2000.
University site
Doe, J: Title of preprint. http://www.uni-heidelberg.de/mydata.html (1999). Accessed 25 Dec 1999.
FTP site
Doe, J: Trivial HTTP, RFC2169. ftp://ftp.isi.edu/in-notes/rfc2169.txt (1999). Accessed 12 Nov 1999.
Organization site
ISSN International Centre: The ISSN register. http://www.issn.org (2006). Accessed 20 Feb 2007.
Dataset with persistent identifier
Zheng L-Y, Guo X-S, He B, Sun L-J, Peng Y, Dong S-S, et al. Genome data from sweet and grain sorghum (Sorghum bicolor). GigaScience Database. 2011. http://dx.doi.org/10.5524/100012.

Figures, tables additional files

See General formatting guidelines for information on how to format figures, tables and additional files.

APPENDIX D

SCRIPTS AND TOOLS USED FOR BIOINFORMATICS ANALYSIS

1. Scripts used in QIIME 2 (microbiome)

a. Importing data

- qiime tools import --type 'SampleData[PairedEndSequencesWithQuality]' --input-path /home/qiime2/Documents/COPD_Microbiome/wetransfer-d046bb --input-format CasavaOneEightSingleLanePerSampleDirFmt --output-path demux-paired-end.qza
- qiime demux summarize --i-data demux-paired-end.qza --o-visualization demux.qzv

b. Sequence quality control and feature table construction

- qiime quality-filter q-score --i-demux demux-paired-end.qza --o-filtered-sequences demux-filtered.qza --o-filter-stats demux-filter-stats.qza
- qiime deblur denoise-16S --i-demultiplexed-seqs demux-filtered.qza --p-trim-length 240 --o-representative-sequences rep-seq-deblur.qza --o-table table.deblur.qza --p-sample-stats --o-stats deblur-ststs.qza
- qiime feature-table summarize --i-table table.deblur.qza --o-visualization table.qzv --m-sample-metadata-file sample-metadata.tsv
- qiime feature-table tabulate-seqs --i-data rep-seq-deblur.qza --o-visualization rep.seq.qzv

c. Obtaining and importing reference data sets

- qiime tools import --type 'FeatureData[Sequence]' --input-path 99_otus.fasta --output-path 99_otus.qza
- qiime tools import --type 'FeatureData[Taxonomy] --input-format HeaderlessTSVTaxonomyFormat --input-path 99_otu_taxonomy.txt --ouput-path ref-taxonomy.qza

d. Extract reference reads

- qiime feature-classifier extract-reads --i-sequences 99_otus.qza --p-f-primer AGAGTTTGATCMTGGCTCAG --p-r-primer GTATTACCGCGGCTGCTGG --o-reads ref-seqs_1.qza

e. Train classifier

- qiime feature-classifier fit-classifier-naive-bayes --i-reference-reads ref-seqs_1.qza --i-reference-taxonomy ref-taxonomy.qza --o-classifier classifier_1.qza

f. Test the classifier

- qiime feature-classifier classify-sklearn –i-classifier classifier.qza --i-reads Tut-rep-seqs.qza --o-classification tut-taxonomy_1.qza
- qiime metadata tabulate --m-input-file tut-taxonomy_1.qza --o-visualization tut-taxonomy_1.qzv

g. Taxonomic analysis

- qiime feature-classifier classify-sklearn --i-classifier classifier_1.qza –i-reads rep-seq-deblur.qza --o-classification taxonomy_1.qza
- qiime metadata tabulate --m-input-file taxonomy.qza --o-visualization taxonomy_1.qzv
- qiime taxa barplot --i-table table.deblur.qza --i-taxonomy taxonomy_1.qza --m-metadata-file sample-metadata.tsv --o-visualization taxa-bar-plots_1.qzv

2. Tools used for Kraken 2 in Galaxy

a. Upload data

b. Use FASTQ joiner to join paired-end reads

c. Remove Human DNA using Bowtie2 (Reference genome: *Homo sapiens* hg38 Full)

d. Run Kraken 2

- Single or paired reads: Single
- Input sequences: FASTQ joiner file
- Print Scientific names instead of just taxids: No
- Confidence: 0.0
- Minimum Base Quality: 0
- Enable quick operation: No
- Split classified and unclassified outputs: No
- Create report:
  - Print a report with aggregate counts/clade to file: No
  - Format report output like Kraken 1's kraken-mpa-report: No
  - Report counts for ALL taxa, even if counts are zero: No
- Select a Kraken2 database: Standard or Viral genomes

3. Converting Kraken 2 files (virome) to .biom files (Must be done in a Linux environment)

```
kraken-biom s8.txt -o S8.biom
```

4. Importing QIIME2 files (microbiome) into phyloseq (in R)

```
library("phyloseq")
library("qiime2R")
Physeq=qza_to_phyloseq(features="C:/Users/Tanweer/Documents/FilesForR/table-deblur.qza", taxonomy
="C:/Users/Tanweer/Documents/FilesForR/taxonomy_1.qza" , metadata
='C:/Users/Tanweer/Documents/FilesForR/sample-metadata.tsv')
```

5. Importing .biom files (virome) files into phyloseq (in R)

```
library("phyloseq")
virseq=import_biom("F:/Virtual Machines/Shared/Kraken_Viraldb_Reports/virome_all.biom", parseFunction =
parse_taxonomy_greengenes)
S25=import_biom("F:/Virtual Machines/Shared/Kraken_Viraldb_Reports/S25.biom", parseFunction =
parse_taxonomy_greengenes)
sample_names(S25)="S25"
Virseq2=merge_phyloseq(virseq, S25)
```

6. Importing IS-Pro files into phyloseq (in R)

```
library("phyloseq")
OTU1=read.csv(file ="C:/Users/Tanweer/Documents/FilesForR/IS-Pro_OTUtable_6.csv" , header = TRUE, sep
= ";", row.names =1)
OTU2=as.matrix(OTU1)
OTU3=otu_table(OTU2, taxa_are_rows = TRUE)
TAX1=read.csv(file = 'C:/Users/Tanweer/Documents/FilesForR/Tax_table_22June.csv', header = TRUE, sep =
";", row.names =1)
TAX2=tax_table(as.matrix(TAX1))
ISPhyseq=merge_phyloseq(OTU3, TAX2)
```

7. Merging QIIME2 and IS-Pro files into a single phyloseq object (in R)

```
Merge_Physeq=merge_phyloseq(ISPhyseq, Physeq)
Meta1=read.delim(file ="C:/Users/Tanweer/Documents/FilesForR/Metadata_forComp_3.txt" , header = TRUE, sep = "\t", row.names = 1)
Meta2=sample_data(Meta1)
Merge_Physeq2=merge_phyloseq(Merge_Physeq, Meta2)
Merge_physeq3=subset_samples(Merge_Physeq2, sample_names(Merge_Physeq2)!="TG-M25_QIIME2")
Merge_Physeq4=subset_taxa(Merge_physeq3, Kingdom!="Archaea" & Family!="mitochondria" & Class!="Chloroplast")
```

8. Calculating alpha diversity (in R)

a. On QIIME2 data

i. Calculate Richness

```
richness=estimate_richness(physeq2)
write.table(richness, "C:/Users/Tanweer/Documents/FilesForR/AlphaDiveristy.txt", sep = "\t")
```

ii. Statistical analysis (done using Kruskal Wallis test)

```
alpha_stats=cbind(richness, sample_data(physeq2))
kt1=kruskal.test(Simpson~DiseaseState, data=alpha_stats)
kruskal.test(Chao1~DiseaseState, data=alpha_stats)
```

b. On IS-Pro data

i. Calculate Richness (all diversity measures)

```
richness=estimate_richness(Merge_Physeq2)
write.table(richness, "C:/Users/Tanweer/Documents/FilesForR/AlphaDiveristy_IS-Pro_2.txt", sep = "\t")
```

ii. Statistical analysis (done using Mann-Whitney test)

```
pairwise.wilcox.test(richness$Simpson, sample_data(Merge_Physeq2)$Method)
pairwise.wilcox.test(richness$Shannon, sample_data(Merge_Physeq2)$Method)
```

c. Creating alpha diversity box plot

```
library("ggplot2")
physeq2=subset_taxa(physeq, Kingdom!="Archaea" & Family!="mitochondria" & Class!="Chloroplast")
plot_richness(physeq2, x= "DiseaseState", color = "DiseaseState", measures = c("Chao1", "Simpson")) + geom_boxplot()
```

9. Calculating beta diversity (in R)

a. Using beta diversity measures

```
physeq.distUF=distance(physeq2, method="uunifrac")
physeq.distWUF=distance(physeq2, method="wunifrac")
physeq.distJac=distance(physeq2, method="jaccard", binary=TRUE)
physeq.distMH=distance(physeq2, method="horn")
```

b. PCoA analysis

```
physeq.distUF.ord=ordinate(physeq2, method = "PCoA", distance = physeq.distUF)
```

c. NMDS analysis

```
physeq.distUF.ord_2=ordinate(physeq2, method = "NMDS", distance = physeq.distUF)
```

d. Generate ordination plot

```
library("ggplot2")
```

```
plot_ordination(physeq2, physeq.distUF.ord, color="DiseaseState") + geom_point() + stat_ellipse()
```

e. Dendrogram/Hierarchical clustering

```
physeq.hclust.distUF=hclust(physeq.distUF, method="average")
plot(as.phylo(physeq.hclust.distUF)
```

10. Creating abundance plots (in R)

a. Calculate relative abundance

```
physeq_RA=transform_sample_counts(physeq2, function(x) x/sum(x))
```

b. Generate plot

```
library("ggplot2")
plot_bar(physeq_RA, fill = "Phylum")+ geom_bar(aes(color=Phylum, fill=Phylum), stat = "identity", position = "stack")
```

11. Analysis using DESeq2 (in R)

a. Analyse using DESeq2

```
library("DESeq2")
Di_des=phyloseq_to_deseq2(physeq, ~DiseaseState)
Di_des_1=DESeq(Di_des)
resultsNames(Di_des_1)
resdf=as.data.frame(DESeq2::results(Di_des_1, format = "DataFrame", name="DiseaseState_Stable_vs_Exacerbation"))
resdf_2=results(Di_des_1, contrast = c("DiseaseState", "Stable", "Exacerbation"))
res=results(Di_des_1, cooksCutoff = FALSE)
alpha=0.1
sigtab=res[which(res$padj <0.01), ]
sigtab_2=cbind(as(sigtab, "data.frame"), as(tax_table(physeq)[rownames(sigtab), ], "matrix"))
head(sigtab_2)
dim(sigtab_2)
head(sigtab [order(sigtab$log2FoldChange ), ] )
```

b. Generate plot

```
library("ggplot2")
theme_set(theme_bw())
scale_fill_discrete <- function(palname= "Set1", ...) {scale_fill_brewer(palette = palname, ...)}
x=tapply(sigtab_2$log2FoldChange, sigtab_2$Phylum, function(x) max(x))
x=sort(x, TRUE)
x=tapply(sigtab_2$log2FoldChange, sigtab_2$Genus, function(x) max(x))
x=sort(x, TRUE)
sigtab_2$Genus=factor(as.character(sigtab_2$Genus), levels=names(x))
ggplot(sigtab_2, aes(x=Genus, y=log2FoldChange, color=Phylum)) +geom_point(size=6)+ theme(axis.text.x = element_text(angle = -90, hjust = 0, vjust = 0.5))
```

APPENDIX E

METADATA

Table 1: Metadata table

Sample ID	Disease State	Age	Gender	Year	Month	Season	HIV status	Hospital	Smoking status	Years smoked	Weather affects cough	phlegm without a cold	phlegm in morning	wheezing	allergies	Previous TB diagnosis	flu vaccine this year	worked in mine	the area where the participant lives
1	S	59	F	2017	Oct	Spring	-	A	Yes	23	Yes	Yes	Yes	Often	Yes	No	No	No	Bronkhorstspruit
2	S	60	M	2017	Oct	Spring	-	A	Yes	38	NA	Yes	Yes	Often	Yes	No	Yes	No	Pretoria
3	S	59	M	2017	Nov	Spring	-	A	Yes	42	Yes	Yes	Yes	Never	No	No	Yes	Yes	Boksburg
4	S	67	M	2018	Jan	Summer	-	A	Stopped	30	Yes	Yes	No	Often	Yes	No	No	No	Doornpoort
7	E	70	F	2018	May	Autumn	-	A	Yes	45	Yes	Yes	Yes	Often	No	No	Yes	No	Wonderboom South
8	S	55	M	2018	May	Autumn	-	A	Yes	15	Yes	Yes	Yes	Sometimes	No	Yes	No	Yes	Mamelodi
9	S	57	M	2018	May	Autumn	-	A	No	NA	Yes	No	Yes	Often	Yes	No	No	Yes	Nelspruit
10	S	74	F	2018	June	Winter	-	A	Yes	30	No	Yes	Yes	Often	No	No	No	No	Jan Niemand Park
11	E	62	F	2018	June	Winter	-	A	No	NA	Maybe	No	No	Sometimes	No	No	No	No	Hammanskraal
13	E	74	F	2018	June	Winter	-	A	Stopped	20	Maybe	Yes	No	Never	No	No	No	No	Eersterust
14	E	56	M	2018	June	Winter	-	B	Yes	20	Yes	Yes	Maybe	Often	No	No	No	No	Pretoria Central
15	S	62	M	2018	July	Winter	-	A	Stopped	25	Yes	Maybe	Maybe	Sometimes	No	No	No	No	Pretoria North
16	S	60	F	2018	July	Winter	-	A	No	NA	Yes	No	No	Often	No	No	Yes	No	Kammeldrift East
17	S	58	M	2018	July	Winter	-	A	Stopped	33	Yes	Maybe	Maybe	Sometimes	No	No	No	No	Hatfield
18	S	70	F	2018	July	Winter	-	A	Stopped	20	Yes	Yes	Yes	Often	No	No	No	No	Rietfontein
20	S	50	M	2018	July	Winter	+	C	Stopped	32	No	Yes	No	Sometimes	No	Yes	No	No	Soshanguve
22	S	60	F	2018	Aug	Winter	-	A	Stopped	20	Yes	Maybe	Yes	Often	No	No	No	No	Bronkhorstspruit

Table 1: Metadata table (continued)

Sample ID	Disease State	Age	Gender	Year	Month	Season	HIV status	Hospital	Smoking status	Years smoked	Weather affects cough	phlegm without a cold	phlegm in morning	wheezing	allergies	Previous TB diagnosis	flu vaccine this year	worked in mine	the area where the participant lives
23	S	60	M	2018	Aug	Winter	-	C	Yes	40	No	Yes	Yes	Never	No	No	No	No	Hammanskraal
24	E	60	M	2018	Oct	Spring	-	C	Stopped	20	Yes	Yes	Yes	Often	No	Yes	No	No	Centurion
25	E	68	F	2018	Oct	Spring	-	A	Stopped	40	No	No	No	Often	No	No	Yes	Yes	Doornpoort
26	S	82	M	2018	Oct	Spring	-	C	Stopped	10	Yes	Yes	Yes	Often	Yes	No	No	No	Mamelodi
27	S	56	M	2018	Nov	Spring	-	C	No	NA	Yes	Yes	Yes	Never	No	No	No	Yes	Mamelodi
28	S	54	M	2018	Nov	Spring	-	C	Yes	40	Yes	Yes	Yes	Often	No	No	No	No	Pretoria West
29	S	59	M	2019	Feb	Summer	-	C	Stopped	13	No	Yes	Yes	Sometimes	No	No	Yes	No	Nellmapius

Table 2: Weight of sputum specimens (frozen) and the volume of 0.1% DTT added

Sample	Date of collection	Date of processing	Weight (g)	Amount of DTT added (mL)
COPD 1		16/05/2019	0.63	0.63
COPD 2		16/05/2019	3.76	3.76
COPD 3		16/05/2019	0.81	0.81
COPD 4		16/05/2019	1.76	1.76
COPD 5	NO specimen			
COPD 6	NO specimen			
COPD 7		16/05/2019	0.2	0.2
COPD 8		16/05/2019	1.24	1.24
COPD 9		16/05/2019	13.6	13.6
COPD 10		16/05/2019	1.12	1.12
COPD 11		16/05/2019	0.68	0.68
COPD 12	NO specimen			
COPD 13		16/05/2019	1.78	1.78
COPD 14		16/05/2019	1.03	1.03
COPD 15		16/05/2019	3.34	3.34
COPD 16		16/05/2019	0.9	0.9
COPD 17		16/05/2019	0.73	0.73
COPD 18		16/05/2019	0.68	0.68
COPD 19	NO specimen			
COPD 20		16/05/2019	1.48	1.48
COPD 21	NO specimen			
COPD 22		16/05/2019	2.45	2.45
COPD 23		16/05/2019	0.66	0.66
COPD 24		16/05/2019	2.43	2.43
COPD 25		16/05/2019	1.19	1.19
COPD 26		16/05/2019	13.52	13.52
COPD 27		16/05/2019	1.77	1.77
COPD 28		16/05/2019	1.58	1.58
COPD 29		16/05/2019	4.2	4.2

Table 3: DNA quality and quantity (Absorbance of DNA)

	Bacterial DNA extraction		Viral DNA extraction	
Sample	**260/280 ratio**	**ng/µL**	**260/280 ratio**	**ng/µL**
COPD 1	1.72	80.06	1.66	36.63
COPD 2	1.45	9.96	1.52	4.59
COPD 3	1.62	12.72	0.62	0.136
COPD 4	1.76	227.95	1.14	10.38
COPD 7	1.75	235.3	17.71	94.2
COPD 8	1.71	150.89	1.76	21.27
COPD 9	1.69	25.77	1.74	132.28
COPD 10	1.63	30.73	1.65	18.72
COPD 11	1.77	135.94	1.02	2.161
COPD 13	1.71	54.89	1.15	4.56
COPD 14	1.71	150.48	1.69	116.16
COPD 15	1.68	48.4	1.52	21.68
COPD 16	1.68	29.28	1.62	11.25
COPD 17	1.14	3.085	1.39	8.127
COPD 18	1.5	5.46	No value	<0
COPD 20	1.7	32.2	1.71	61.28
COPD 22	1.6	9.599	No value	<0
COPD 23	1.69	11.78	0.688	0.31
COPD 24	1.76	79.36	No value	<0
COPD 25	1.7	363.05	1.6	18.76
COPD 26	1.73	55.83	No value	<0
COPD 27	1.68	37.24	1.66	27.68
COPD 28	1.71	92.19	1.37	10.95
COPD 29	1.67	150.89	No value	<0

Table 4: RNA quality and quantity (Absorbance of RNA)

	Viral RNA extraction	
Sample	**Absorbance**	**ng/µL**
COPD 1	0.019	15.05
COPD 2	0.021	16.44
COPD 3	0.069	55.22
COPD 4	0.101	80.97
COPD 7	0.034	27.59
COPD 8	0.189	150.9
COPD 9	0.839	671.5
COPD 10	0.041	32.77
COPD 11	0.195	155.9
COPD 13	0.067	53.99
COPD 14	0.771	616.9
COPD 15	0.052	41.93
COPD 16	0.062	49.34
COPD 17	0.132	105.3
COPD 18	0.068	54.33
COPD 20	0.059	46.97
COPD 22	0.070	55.72
COPD 23	0.048	30.22
COPD 24	0.103	82.77
COPD 25	0.184	147.1
COPD 26	0.089	71.33
COPD 27	0.030	24.12
COPD 28	0.041	32.58
COPD 29	0.124	99.21

APPENDIX F

APPROVAL DOCUMENTS

The Research Ethics Committee, Faculty Health Sciences, University of Pretoria complies with ICH-GCP guidelines and has US Federal wide Assurance.
- FWA 00002567, Approved dd 22 May 2002 and Expires 03/20/2022.
- IRB 0000 2235 IORG0001762 Approved dd 22/04/2014 and Expires 03/14/2020.

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Health Sciences Research Ethics Committee

29/06/2017

Approval Certificate
New Application

Ethics Reference No.: 237/2017

Title: Lung microbiome of chronic obstructive pulmonary disease patients with and without HIV infection in Pretoria, South Africa.

Dear Tanweer G Mahomed

The **New Application** as supported by documents specified in your cover letter dated 20/06/2017 for your research received on the 23/06/2017, was approved by the Faculty of Health Sciences Research Ethics Committee on its quorate meeting of 28/06/2017.

Please note the following about your ethics approval:
- Ethics Approval is valid for 3 years
- Please remember to use your protocol number (**237/2017**) on any documents or correspondence with the Research Ethics Committee regarding your research.
- Please note that the Research Ethics Committee may ask further questions, seek additional information, require further modification, or monitor the conduct of your research.

Ethics approval is subject to the following:
- The ethics approval is conditional on the receipt of **6 monthly written Progress Reports**, and
- The ethics approval is conditional on the research being conducted as stipulated by the details of all documents submitted to the Committee. In the event that a further need arises to change who the investigators are, the methods or any other aspect, such changes must be submitted as an Amendment for approval by the Committee.

Additional Conditions:
- Approval is conditional that the Research Ethics Committee receives an export permit before any materials are exported.

We wish you the best with your research.

Yours sincerely

Dr R Sommers; MBChB; MMed (Int); MPharMed,PhD
Deputy Chairperson of the Faculty of Health Sciences Research Ethics Committee, University of Pretoria

The Faculty of Health Sciences Research Ethics Committee complies with the SA National Act 61 of 2003 as it pertains to health research and the United States Code of Federal Regulations Title 45 and 46. This committee abides by the ethical norms and principles for research, established by the Declaration of Helsinki, the South African Medical Research Council Guidelines as well as the Guidelines for Ethical Research: Principles Structures and Processes, Second Edition 2015 (Department of Health).

☎ 012 356 3084 deepeka.behari@up.ac.za http://www.up.ac.za/healthethics
✉ Private Bag X323, Arcadia, 0007 - Tswelopele Building, Level 4, Room 60, Gezina, Pretoria

The Research Ethics Committee, Faculty Health Sciences, University of Pretoria complies with ICH-GCP guidelines and has US Federal wide Assurance.
- FWA 00002567, Approved dd 22 May 2002 and Expires 03/20/2022.
- IRB 0000 2235 IORG0001762 Approved dd 22/04/2014 and Expires 03/14/2020.

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Health Sciences Research Ethics Committee

1/02/2018

Approval Certificate
Amendment
(to be read in conjunction with the main approval certificate)

Ethics Reference No: 237/2017

Title: Lung microbiome of chronic obstructive pulmonary disease patients with and without HIV infection in Pretoria, South Africa

Dear Miss Tanweer Goolam Mahomed

The **Amendment** as described in your documents specified in your cover letter dated 11/12/2017 received on 11/12/2017 was approved by the Faculty of Health Sciences Research Ethics Committee on its quorate meeting of 31/01/2018.

Please note the following about your ethics amendment:
- Please remember to use your protocol number (**237/2017**) on any documents or correspondence with the Research Ethics Committee regarding your research.
- Please note that the Research Ethics Committe may ask further questions, seek additional information, require further modification, or monitor the conduct of your research.

Ethics amendment is subject to the following:
- The ethics approval is conditional on the receipt of **6 monthly written Progress Reports**, and
- The ethics approval is conditional on the research being conducted as stipulated by the details of all documents submitted to the Committee. In the event that a further need arises to change who the investigators are, the methods or any other aspect, such changes must be submitted as an Amendment for approval by the Committee.

We wish you the best with your research.

Yours sincerely

Dr R Sommers, MBChB; MMed (Int); MPharMed; PhD
Deputy Chairperson of the Faculty of Health Sciences Research Ethics Committee, University of Pretoria

The Faculty of Health Sciences Research Ethics Committee complies with the SA National Act 61 of 2003 as it pertains to health research and the United States Code of Federal Regulations Title 45 and 46. This committee abides by the ethical norms and principles for research, established by the Declaration of Helsinki, the South African Medical Research Council Guidelines as well as the Guidelines for Ethical Research: Principles Structures and Processes, Second Edition 2015 (Department of Health).

☎ 012 356 3084 deepeka.behari@up.ac.za / fhsethics@up.ac.za http://www.up.ac.za/healthethics
✉ Private Bag X323, Arcadia, 0007 - Tswelopele Building, Level 4, Room 60 / 61, 31 Bophelo Road, Gezina, Pretoria

The Research Ethics Committee, Faculty Health Sciences, University of Pretoria complies with ICH-GCP guidelines and has US Federal wide Assurance.
- FWA 00002567, Approved dd 22 May 2002 and Expires 03/20/2022.
- IRB 0000 2235 IORG0001762 Approved dd 22/04/2014 and Expires 03/14/2020.

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Health Sciences Research Ethics Committee

4/06/2018

Approval Certificate
Amendment
(to be read in conjunction with the main approval certificate)

Ethics Reference No: 237/2017

Title: Lung microbiome of chronic obstructive pulmonary disease patients with and without HIV infection in Pretoria, South Africa.

Dear Miss Tanweer Goolam Mahomed

The **Amendment** as described in your documents specified in your cover letter dated 4/06/2018 received on 4/06/2018 was approved by the Faculty of Health Sciences Research Ethics Committee on its quorate meeting of 11/06/2018.

Please note the following about your ethics amendment:
- Please remember to use your protocol number (**237/2017**) on any documents or correspondence with the Research Ethics Committee regarding your research.
- Please note that the Research Ethics Committe may ask further questions, seek additional information, require further modification, or monitor the conduct of your research.

Ethics amendment is subject to the following:
- The ethics approval is conditional on the receipt of **6 monthly written Progress Reports**, and
- The ethics approval is conditional on the research being conducted as stipulated by the details of all documents submitted to the Committee. In the event that a further need arises to change who the investigators are, the methods or any other aspect, such changes must be submitted as an Amendment for approval by the Committee.

We wish you the best with your research.

Yours sincerely

Dr R Sommers; MBChB; MMed (Int); MPharMed; PhD
Deputy Chairperson of the Faculty of Health Sciences Research Ethics Committee, University of Pretoria

The Faculty of Health Sciences Research Ethics Committee complies with the SA National Act 61 of 2003 as it pertains to health research and the United States Code of Federal Regulations Title 45 and 46. This committee abides by the ethical norms and principles for research, established by the Declaration of Helsinki, the South African Medical Research Council Guidelines as well as the Guidelines for Ethical Research: Principles Structures and Processes, Second Edition 2015 (Department of Health).

☎ 012 356 3084 deepeka.behari@up.ac.za / fhsethics@up.ac.za http://www.up.ac.za/healthethics
✉ Private Bag X323, Arcadia, 0007 - Tswelopele Building, Level 4, Room 60 / 61, 31 Bophelo Road, Gezina, Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Health Sciences

The Research Ethics Committee, Faculty Health Sciences, University of Pretoria complies with ICH-GCP guidelines and has US Federal wide Assurance.

- FWA 00002567, Approved dd 22 May 2002 and Expires 03/20/2022.
- IRB 0000 2235 IORG0001762 Approved dd 22/04/2014 and Expires 03/14/2020.

1 February 2019

Approval Certificate
Annual Renewal

Ethics Reference No.: 237/2017
Title: Lung microbiome of chronic obstructive pulmonary disease patients with and without HIV infection in Pretoria, South Africa

Dear Miss T Goolam Mahomed

The **Annual Renewal** as supported by documents received between 2019-01-07 and 2019-01-30 for your research, was approved by the Faculty of Health Sciences Research Ethics Committee on its quorate meeting of 2019-01-30.

Please note the following about your ethics approval:

- Renewal of ethics approval is valid for 1 year, subsequent annual renewal will become due on 2020-02-01.
- Please remember to use your protocol number (237/2017) on any documents or correspondence with the Research Ethics Committee regarding your research.
- Please note that the Research Ethics Committee may ask further questions, seek additional information, require further modification, monitor the conduct of your research, or suspend or withdraw ethics approval.

Ethics approval is subject to the following:

- The ethics approval is conditional on the research being conducted as stipulated by the details of all documents submitted to the Committee. In the event that a further need arises to change who the investigators are, the methods or any other aspect, such changes must be submitted as an Amendment for approval by the Committee.

We wish you the best with your research.

Yours sincerely

Dr R Sommers
MBChB MMed (Int) MPharmMed PhD
Deputy Chairperson of the Faculty of Health Sciences Research Ethics Committee, University of Pretoria

The Faculty of Health Sciences Research Ethics Committee complies with the SA National Act 61 of 2003 as it pertains to health research and the United States Code of Federal Regulations Title 45 and 46. This committee abides by the ethical norms and principles for research, established by the Declaration of Helsinki, the South African Medical Research Council Guidelines as well as the Guidelines for Ethical Research: Principles Structures and Processes, Second Edition 2015 (Department of Health)

Research Ethics Committee
Room 4-60, Level 4, Tswelopele Building
University of Pretoria, Private Bag X323
Arcadia 0007, South Africa
Tel +27 (0)12 356 3084
Email deepeka.behari@up.ac.za
www.up.ac.za

Fakulteit Gesondheidswetenskappe
Lefapha la Disaense tša Maphelo

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Health Sciences

Institution: The Research Ethics Committee, Faculty Health Sciences, University of Pretoria complies with ICH-GCP guidelines and has US Federal wide Assurance.

- FWA 00002567, Approved dd 22 May 2002 and Expires 03/20/2022.
- IORG #: IORG0001762 OMB No. 0990-0279 Approved for use through February 28, 2022 and Expires: 03/04/2023.

11 March 2020

Approval Certificate
Annual Renewal

Ethics Reference No.: 237/2017
Title: Lung microbiome of chronic obstructive pulmonary disease patients with and without HIV infection in Pretoria, South Africa

Dear Miss T Goolam Mahomed

The **Annual Renewal** as supported by documents received between 2020-02-19 and 2020-03-11 for your research, was approved by the Faculty of Health Sciences Research Ethics Committee on its quorate meeting of 2020-03-11.

Please note the following about your ethics approval:

- Renewal of ethics approval is valid for 1 year, subsequent annual renewal will become due on 2021-03-11.
- Please remember to use your protocol number (237/2017) on any documents or correspondence with the Research Ethics Committee regarding your research.
- Please note that the Research Ethics Committee may ask further questions, seek additional information, require further modification, monitor the conduct of your research, or suspend or withdraw ethics approval.

Ethics approval is subject to the following:

- The ethics approval is conditional on the research being conducted as stipulated by the details of all documents submitted to the Committee. In the event that a further need arises to change who the investigators are, the methods or any other aspect, such changes must be submitted as an Amendment for approval by the Committee.

We wish you the best with your research.

Yours sincerely

Dr R Sommers
MBChB MMed (Int) MPharmMed PhD
Deputy Chairperson of the Faculty of Health Sciences Research Ethics Committee, University of Pretoria

The Faculty of Health Sciences Research Ethics Committee complies with the SA National Act 61 of 2003 as it pertains to health research and the United States Code of Federal Regulations Title 45 and 46. This committee abides by the ethical norms and principles for research, established by the Declaration of Helsinki, the South African Medical Research Council Guidelines as well as the Guidelines for Ethical Research: Principles Structures and Processes, Second Edition 2015 (Department of Health)

Research Ethics Committee
Room 4-60, Level 4, Tswelopele Building
University of Pretoria, Private Bag x323
Gezina 0031, South Africa
Tel +27 (0)12 356 3084
Email: deepeka.behari@up.ac.za
www.up.ac.za

Fakulteit Gesondheidswetenskappe
Lefapha la Disaense tša Maphelo

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Health Sciences

Institution: The Research Ethics Committee, Faculty Health Sciences, University of Pretoria complies with ICH-GCP guidelines and has US Federal wide Assurance.

- FWA 00002567, Approved dd 22 May 2002 and Expires 03/20/2022.
- IORG #: IORG0001762 OMB No. 0990-0279 Approved for use through February 28, 2022 and Expires: 03/04/2023.

22 January 2021

Acknowledgement Certificate
Research Completed

Ethics Reference No.: 237/2017
Title: Lung microbiome of chronic obstructive pulmonary disease patients with and without HIV infection in Pretoria, South Africa

Dear Miss T Goolam Mahomed

The **Research Completed Report** as supported by documents received between 2021-01-06 and 2021-01-20 for your research, was acknowledged by the Faculty of Health Sciences Research Ethics Committee on 2021-01-20 as resolved by its quorate meeting.

Yours sincerely

Dr R Sommers
MBChB MMed (Int) MPharmMed PhD
Deputy Chairperson of the Faculty of Health Sciences Research Ethics Committee, University of Pretoria

The Faculty of Health Sciences Research Ethics Committee complies with the SA National Act 61 of 2003 as it pertains to health research and the United States Code of Federal Regulations Title 45 and 46. This committee abides by the ethical norms and principles for research, established by the Declaration of Helsinki, the South African Medical Research Council Guidelines as well as the Guidelines for Ethical Research: Principles Structures and Processes, Second Edition 2015 (Department of Health)

Research Ethics Committee
Room 4-60, Level 4, Tswelopele Building
University of Pretoria, Private Bag x323
Gezina 0031, South Africa
Tel +27 (0)12 356 3084
Email: deepeka.behari@up.ac.za
www.up.ac.za

Fakulteit Gesondheidswetenskappe
Lefapha la Disaense tša Maphelo

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R14/49 Prof Rajen Morar et al

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M1902104

NAME: (Principal Investigator)	Prof Rajen Morar et al
DEPARTMENT:	Internal Medicine Charlotte Maxeke Johannesburg Academic Hospital
PROJECT TITLE:	Lung microbiome of chronic obstructive pulmonary disease patients with and without HIV infection in Pretoria, South Africa
DATE CONSIDERED:	Ad hoc
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	
APPROVED BY:	CBPenny Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	05/04/2019

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on the Third Floor, Faculty of Health Sciences, Phillip Tobias Building, 29 Princess of Wales Terrace, Parktown, 2193, University of the Witwatersrand. I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in **February** and will therefore be due in the month of **February** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

R Morar R Morar — 5 April 2019

Principal Investigator Signature — Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Faculty of Health Sciences

20 June 2017

Prof MM Ehlers

Department of Medical Microbiology

Pathology Building

UNIVERSITY OF PRETORIA

STUDENT: GOOLAM MAHOMED T (PhD Medical Microbiology)

"Lung microbiome of chronic obstructive pulmonary disease patients with and without HIV infection in Pretoria, South Africa"

Abovementioned student's resubmission has been approved by the committee meeting held on the 13th of June 2017.

Kind regards

PROF BG LINDEQUE
DEPUTY DEAN: SCHOOL OF MEDICINE

HW Snyman North Building, Room 5-16.1 Tel: +27 12 319 2100
University of Pretoria
PRETORIA 0002
Republic of South Africa

emmie.vandermerwe@up.ac.za
www.up.ac.za

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Health Sciences

13 June 2018

Prof MM Ehlers
Dept of Medical Microbiology
Faculty of Health Sciences

Dear Prof Ehlers

STUDENT : GOOLAM MAHOMED T (PhD MEDICAL MICROBIOLOGY)

Lung microbiome of chronic obstructive pulmonary disease patients with and without HIV infection in Pretoria, South Africa

The amendment by the ethics committee re. the inclusion of more clinics in the study, has been approved by the PhD committee on the 5th of June 2018.

We wish you all the best with your study.

Kind regards

PROF V STEENKAMP
CHAIR: PhD COMMITTEE

Pharmacology Dept., BMS Building
University of Pretoria, Private Bag X323
Arcadia 0007, South Africa
Tel +27 (0)12 319 2254
Email: vanessa.steenkamp@up.ac.za

Fakulteit Gesondheidswetenskappe
Lefapha la Disaense tša Maphelo