

---

# Investigating the role of respiratory co-infections and the nasopharyngeal microbiome in children with suspected pulmonary tuberculosis

---

By Bianca Leigh Hamman



Thesis presented in fulfilment of the requirements for the degree of Master of Science in the Faculty of  
Medicine and Health Sciences at Stellenbosch University

Supervisors: Dr Mae Newton-Foot &  
Dr Marieke Van der Zalm  
Division of Medical Microbiology, Department of Pathology

December 2020

UNIVERSITY OF STELLENBOSCH

## Declaration

“By submitting this thesis/dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.”

Date: December 2020

Copyright © 2020 Stellenbosch University

All rights reserved

## Abstract

**Introduction:** Tuberculosis (TB) is a global health problem, causing morbidity, mortality and devastating social and economic impacts. Pediatric TB is particularly challenging due to difficulties in diagnosis. Children are particularly susceptible to respiratory infections and this may be influenced by the microbial colonization of the respiratory tract, which may play a role in the clinical presentation and pathogenesis of TB. The nasopharyngeal microbiome is critical for respiratory health and may impact on the development, presentation and diagnosis of TB disease. Antibiotics contribute to microbial dysbiosis which may lead to the development, progression or exacerbation of other diseases. However, there is limited data describing the nasopharyngeal microbiota of children with and without TB, or the effect of TB treatment on the nasopharyngeal microbiome.

**Methods:** Respiratory samples were obtained from pediatric patients with suspected pulmonary TB (PTB) at baseline and follow up visits (2 and 6 months). Participants were classified as having bacteriologically confirmed PTB, clinically diagnosed PTB or unlikely PTB (well-defined ill controls). Respiratory pathogens were detected in all baseline respiratory samples using the Seegene Allplex™ Respiratory Panel 4 and a *Pneumocystis jirovecii* real-time PCR assay.

The nasopharyngeal microbiome of 26 participants was characterized and the effect of TB treatment determined by 16S rRNA sequencing, using the Illumina Miseq platform.

**Results:** Seventy children were included; 27.1% were categorized with bacteriologically confirmed PTB, 32.9% with clinically diagnosed PTB and 40% with unlikely PTB. The most frequently detected bacterial pathogens were *Haemophilus influenzae* (52/70, 74.2%) and *Streptococcus pneumoniae* (42/70, 60%). There was no association between the presence of bacterial pathogens/pathogens and TB disease.

Due to poor sequence quality resulting from load shedding during sequencing, the reverse reads were excluded from microbiome analysis. The most commonly detected phyla in all samples were *Proteobacteria*, *Fusobacteria*, *Firmicutes* and *Bacteroidetes*. Common familia included *Streptococcaceae*, *Pasteurellaceae*, *Moraxellaceae*, *Prevotellaceae*, *Veillonellaceae* and *Neisseriaceae*. There were no significant differences in the microbiome profile or alpha and beta diversity between TB cases and controls at baseline. However, differential abundance testing showed 4-5 fold differences in abundance of *Pasteurellaceae* and *Prevotellaceae* between the TB cases and ill controls. There was also no significant difference in microbiota profile or alpha diversity at 2 or 6 months in TB cases, who received TB treatment. However, differential abundance testing identified a reduction in the abundance of *Veillonellaceae*, *Staphylococcaceae*, *Prevotellaceae*, *Neisseriaceae*, *Enterobacteriaceae* and *Aerococcaceae* in TB cases after treatment.

**Conclusion:** This study observed no significant differences between the respiratory pathogens in children with and without PTB. Similarly, no differences in alpha or beta diversity were observed between the respiratory microbiota of TB cases and controls, or after TB treatment.

However, differences in abundance of some families, between TB cases and controls at baseline, and before and after TB treatment, suggest that further research on this topic is warranted, considering the numerous limitations which may have impacted the findings of this study. This study contributed to the data available regarding respiratory microbiota in children with suspected PTB in a TB endemic setting and highlighted the challenges of conducting microbiome research in resource limited settings.

## Opsomming

**Inleiding:** Tuberkulose (TB) is 'n wêreldwye gesondheidsprobleem wat morbiditeit en mortaliteit veroorsaak en verrykende sosiale en ekonomiese impakte het. Pediatriese TB is veral uitdagend weens diagnose moeilikhede. Kinders is veral vatbaar vir respiratoriese infeksies en dit kan beïnvloed word deur die mikrobiële kolonisasie van die respiratoriese kanaal wat 'n rol kan speel in die kliniese aanbieding en patogenese van TB. Die nasofaringeale mikrobiom is krities vir respiratoriese gesondheid en kan die ontwikkeling, aanbieding en diagnose van TB beïnvloed. Antibiotika kan bydra tot disbiose van die nasofaringeale mikrobiota wat kan lei tot die ontwikkeling, bevordering of verergering van ander siektes. Alhoewel, daar is beperkte data wat die nasofaringeale mikrobiota van kinders met en sonder TB, of die effek van TB behandeling op die nasofaringeale mikrobiom beskryf.

**Metodes:** Respiratoriese monsters is geneem vanaf pediatriese pasiënte met moontlike pulmonêre TB by basislyn- en opvolgbesoeke (2 en 6 maande). Deelnemers is geklassifiseer as volg: met bakteriologies-bevestigde PTB, klinies gediagnoseerde PTB of onwaarskynlike PTB (goed beskryfde siek kontrole).

Respiratoriese patogene is in alle basislyn respiratoriese monsters opgespoor, deur die gebruik van Seegene Allplex™ Respiratory Panel 4 en 'n *Pneumocystis jirovecii* PCR toets.

Die nasofaringeale mikrobiom van 26 deelnemers was karakteriseer en die effek van TB behandeling bepaal, deur 16S rRNA volgordebepaling, geteiken met Illumina Miseq tegnologie.

**Resultate:** Sewentig kinders is ingesluit; 27.1% met bakteriologies-bevestigde PTB geklassifiseer is, 32.9% met klinies gediagnoseerde PTB en 40% met onwaarskynlike PTB. Die mees algemeen opgespoorde bakteriële patogene was *Haemophilus influenzae* (52/70, 74.2%) en *Streptococcus pneumoniae* (42/70, 60%). Daar was geen verwantskap tussen die teenwoordigheid van sekere bakteriële “pathobionts”/patogene en TB siekte nie.

As gevolg van die slegte kwaliteit van die volgordes as gevolg van beurtkrag, was die “reverse reads” nie ingesluit in die mikrobiom analise nie. *Proteobacteria*, *Fusobacteria*, *Firmicutes* en *Bacteroidetes* was die mees algemeen opgespoorde filums. Algemene gesinne het onder andere *Streptococcaceae*, *Pasteurellaceae*, *Moraxellaceae*, *Prevotellaceae*, *Veillonellaceae* en *Neisseriaceae* ingesluit. Daar was geen merkwaardige verskille in die mikrobiota profiele of alfa en beta diversiteit tussen TB gevalle en kontrole by basislyn nie. Alhoewel, differensiaal oorfloed toetse wys dat daar 4-5 vou verskillende in die oorfloed van *Pasteurellaceae* en *Prevotellaceae* tussen die TB gevalle en kontrole groep. Daar was boonop geen merkwaardige verskille in die mikrobiota profiele of alfa diversiteite in TB gevalle wat TB behandeling ontvang het by 2 of 6 maande nie. Alhoewel, differensiaal oorfloed toetse het 'n vermindering in die oorfloed van *Veillonellaceae*, *Staphylococcaceae*, *Prevotellaceae*, *Neisseriaceae*, *Enterobacteriaceae* and *Aerococcaceae* identifiseer in die TB gevalle na TB behandeling.

### **Gevolgtrekking:**

Hierdie ondersoek het geen merkwaardige verskille tussen die respiratoriese patogene in kinders met en sonder PTB waargeneem nie. Ingelyks, geen verskille in alfa en beta diversiteit tussen die respiratoriese mikrobiota van TB gevalle en kontrole, of na TB behandeling was waargeneem nie. Alhoewel verskille in die oorfloed van sommige gesinne, tussen TB gevalle en kontrole by basislyn, en voor en na TB behandeling voorstel dat verder navorsing gedoen moet word op hierdie onderwerp aagesien dat daar baie beperkings was wat die bevinding kon beïnvloed. Hierdie studie het bygedra tot die tekort aan beskikbare data met betrekking tot die respiratoriese mikrobiota in kinders met moontlike PTB in 'n TB endemiese omgewing en het die uitdagings van die uitvoer van mikrobiom navorsing in hulpbron-beperkte omgewings uitgelig.

## Acknowledgements

I would like to thank God for seeing me through and giving me the strength and ability to complete this postgraduate journey.

I would like to thank my supervisors (Dr Mae Newton-Foot and Dr Marieke van der Zalm) for their patience and constant words of encouragement during the course of my MSc degree. I would like to thank Dr Kirby for her support during the laboratory section of the work and allowing me to use their facilities under her supervision and guidance to conduct my laboratory work at the Institute for Microbial Biotechnology and Metagenomics, Department of Biotechnology, University of the Western Cape. As well as Dr Rubin Rhode for his mentorship and guidance.

I would like to acknowledge and thank Kristien and Jacques Nel van Zyl for helping me understand and for guiding me through the use of the server that was required for the bioinformatic analysis of the project. In addition, I would like to thank my friends in- and the students in the Medical Microbiology Division for always checking up on me to make sure I was okay and their constant words of encouragement and prayers.

I would like to acknowledge the National Research Fund (NRF) for funding my postgraduate studies for the duration of my MSc degree, I would not have been able to do my postgraduate studies without the financial support. As well as the National Health Laboratory Services Research Trust grant that allowed us to conduct this research.

Lastly, I would like to thank my parents, siblings and the rest of the family for their love, support and prayers during this time. I hope that by completing this degree it shows that through prayer, perseverance, love and support that any goal in life can be achieved!

## Table of Contents

Declaration.....	i
Abstract .....	ii
Opsomming .....	iv
Acknowledgements.....	vi
Table of Contents.....	vii
List of abbreviations .....	x
Terminology .....	xii
List of Tables .....	xiii
List of Figures .....	xiv
CHAPTER 1: Literature Review .....	16
1.1 Etiology of Tuberculosis.....	16
1.2 Clinical presentation .....	17
1.3 Pediatric TB.....	18
1.4 TB Diagnosis .....	20
1.4.1 Bacteriological diagnosis.....	20
1.4.2 Diagnosis of pediatric TB .....	22
1.5 TB Treatment .....	23
1.6 TB and the Microbiome.....	26
1.6.1 Nasopharyngeal microbiome.....	26
1.6.2 The respiratory microbiota and clinical presentation of PTB.....	28
1.6.3 The microbiome and diagnosis of PTB.....	28
1.6.4 TB treatment and the microbiome .....	29
1.7 Problem statement .....	31
1.8 Aims and objectives.....	31
1.9 Study Population .....	32
1.10 Sample collection .....	34
CHAPTER 2: Detection of “other” respiratory pathogens in children suspected of pulmonary tuberculosis.....	35
2.1 Introduction.....	35
2.2 Materials and Methods .....	37
2.2.1 Sample collection.....	37
2.2.2 DNA extraction.....	37
2.2.3 Seegene Allplex Respiratory Panel 4 PCR .....	38
2.2.4 Pneumocystis jirovecii real time-PCR .....	40
2.2.5 Clinical data collection .....	41
2.2.6 Statistical Analysis .....	41
2.3 Results .....	42
2.3.1 Pathogen detection.....	44

2.3.2	Evaluation of risk factors for the presence of respiratory pathogens .....	47
2.4	Discussion.....	50
2.5	Conclusion.....	57
CHAPTER 3: The description of the respiratory microbiome using 16S rRNA gene sequencing...		58
3.1	Introduction.....	58
3.2	Materials and Methods:.....	60
3.2.1	Sample selection .....	61
3.2.2	DNA Extraction .....	62
3.2.3	16S rRNA library preparation .....	64
3.2.4	Generation of <i>fastq</i> files .....	71
3.2.5	FAST Quality Control (QC) .....	71
3.2.6	Data analysis Pipeline.....	72
3.3	Results .....	73
3.3.1	DNA quality assessment of samples.....	73
3.3.2	16S rRNA library preparation .....	73
3.3.3	Illumina Miseq sequencing.....	75
3.3.4	Sequencing quality: FASTQ quality control (QC) assessment.....	75
3.3.5	Taxonomic classification .....	77
3.4	Discussion .....	84
3.5	Conclusion.....	90
CHAPTER 4: Tuberculosis and the Nasopharyngeal microbiome (microbiota) .....		91
4.1	Introduction.....	91
4.2	Material and Methods .....	93
4.3	Results .....	96
4.3.1	The respiratory microbiota in TB cases and ill controls .....	96
4.3.2	The respiratory microbiota during TB treatment .....	104
4.3.3	Baseline TB cases compared to month 6 ill controls as a proxy for healthy microbiome (microbiota).....	111
4.4	Discussion .....	114
4.5	Conclusion.....	121
CHAPTER 5: Concluding remarks .....		122
References .....		130
Addendum 1 .....		144
Addendum 2 .....		145
Addendum 3 .....		188
Addendum 4 .....		189
Addendum 5 .....		190
Addendum 6 .....		191
Addendum 7 .....		192

Addendum 8 .....	210
Addendum 9 .....	211
Addendum 10 .....	212

## List of abbreviations

ACP	Annealing control primer
AFB	Acid fast bacilli
AMK	Amikacin
ATCC	American Type Culture Collection
bp	Base pair
CAP	Community acquired pneumonia
CSF	Cerebrospinal fluid
C/S	Caesarean section
DNA	Deoxyribonucleic acid
DPO	Dual priming oligonucleotides
EB	Elution buffer
EMB	Ethambutol
FLD	First line drugs
GA	Gastric aspirates
HS	High sensitivity
HT1	Hybridization buffer (Illumina)
IGRA	Interferon gamma release assay
INH	Isoniazid
IS	Induced sputum
KAN	Kanamycin
LPA	Line probe assay
MDR-TB	Multi drug resistant TB
<i>M.tb</i>	<i>Mycobacterium tuberculosis</i>
MuDT	Multiple detection temperature
NEC	Negative extraction control
NPA	Nasopharyngeal aspirates
NPO	Nil per os
NGS	Next generation sequencing

NTC	Non-template control
NVD	Natural vaginal delivery
OTU	Operational taxonomic unit
PC	Positive control
PTB	Pulmonary TB
PAS	Para-aminosalicylic
PCOA	Principal coordinate analysis
PCP	<i>Pneumocystis jirovecii</i>
PCR	Polymerase chain reaction
PR2	Incorporation buffer (Illumina)
PZA	Pyrazinamide
ReAD	Real amplicon detection
RIF	Rifampicin
RP4 PC	Respiratory Panel 4 positive control
RP-B IC	Respiratory Panel bacteria internal control
SLD	Second line drug
TB	Tuberculosis
TST	Tuberculin skin test
TOCE	Tagging oligonucleotide cleavage extension
VTM	Viral transport media
WHO	World Health Organization
WRR	Within run repeat
XDR-TB	Extensively drug resistant tuberculosis

## Terminology

<b>16S ribosomal RNA gene</b>	Encoded the 16S ribosomal RNA, a component of the 30S small subunit of prokaryotic ribosomes. Used to reconstruct phylogenies owing to the extremely slow rate of evolution of this gene and the presence of both variable and conserved regions allowing amplification and sequence comparison.
<b>Amplicon</b>	PCR amplified DNA product.
<b>Diversity</b>	The number and distribution of distinct OTUs in a sample or in the originating population.
<b>Dysbiosis</b>	Alteration of microbial composition linked to perturbation of local ecological conditions, generally associated with impaired host-microbe interactions.
<b>Evenness</b>	Measure of the similarity of the relative abundances of the different OTUs in the population.
<b>Microbiome</b>	This term refers to the microbial community with its genetic information and inferred physio-chemical properties of the gene products of the microbiota.
<b>Microbiota</b>	All microorganisms including bacteria, viruses, fungi and archaea.
<b>Operational Taxonomic Unit (OTU)</b>	A cluster of microorganisms grouped by DNA sequence similarity of a specific taxonomic marker gene (e.g. 16S rRNA). OTUs are used as representative for microbial “species” at different taxonomic levels: phylum, class, order, family, genus and species.
<b>Relative abundance</b>	How common or rare an OTU is relative to other OTUs in a community, measured as a percentage of the total number of OTUs in the population.
<b>Sequencing read</b>	The primary output of DNA sequencing, consisting of a short stretch of DNA sequence that is produced from sequencing a region of a single DNA fragment
<b>Throughput</b>	Number of samples that can be run on a sequencing platform simultaneously and at a reasonable cost.

## List of Tables

Table 1.1: The frequency of symptoms and signs of PTB according to age..	19
Table 1.2: First line and second line anti-TB drugs recommended for children.....	25
Table 2.1: Thermal profile setup on CFX-96™ Real Time PCR machine. ....	39
Table 2.2: Fluorophores used for the detection of analytes. ....	39
Table 2.3: MSG Heminested primer sequences and product sizes. ....	40
Table 2.4: Participant information and risk factors.....	43
Table 2.5: The number of bacteria detected in the TB and unlikely TB groups.....	47
Table 2.6: Risk factors for the presence of bacterial pathogens. ....	48
Table 3.1: Sequencing controls.....	63
Table 3.2: 16S rRNA V4 PCR primers. ....	65
Table 3.3: 16S rRNA V4 touchdown Amplicon PCR cycling conditions.....	65
Table 4.1: Alpha diversity measures. ....	94
Table 4.2: Common familia present in the baseline TB cases and month 6 ill controls.....	111

## List of Figures

Figure 1.1: The diagnosis of TB based on the detection of <i>M. tuberculosis</i> .....	21
Figure 1.2: Host and environmental factors contributing to changes in the respiratory microbiome (microbiota).....	27
Figure 1.3: Overview of diagnostic categories assigned to participants and the characterization of those participants as either TB cases or well-defined ill controls.....	34
Figure 2:1: Basic workflow of DNA extraction using the Qiagen QIAamp DNA extraction Kit. ....	37
Figure 2:2: Representative <i>P. jirovecii</i> melt curve. ....	44
Figure 2:3: The percentage of samples (n=70) in which pathogens were detected .....	45
Figure 2:4: Pathogens detected in defined participant categories .....	46
Figure 3.1: Wet and dry laboratory workflow for 16S library preparation, sequencing and analysis. ....	60
Figure 3.2: The breakdown of preselected samples for the microbiome analysis. ....	61
Figure 3.3: Basic workflow for PCR Clean-up using the Agencourt AMPure XP PCR purification system. ....	66
Figure 3.4: Index 1 and 2 appended to either end of amplified 16S rRNA V4 sequence. ....	67
Figure 3.5: Miseq Workflow.....	70
Figure 3.6: Representative gel of amplified amplicon and indexed PCR products from a mock control sample:.....	74
Figure 3.7: A graphical representation of the Fast Quality Control (QC) report for the forward (A) and reverse reads (B) obtained from a mock control.....	766
Figure 3.8: The relative abundance of phyla observed across samples and sequencing controls..	78
Figure 3.9: The relative abundance of phyla observed in (A) Mock control 1-equal volume, (B) Mock control 2- equal concentrations and (C) the PCR positive control ( <i>E. coli</i> ).....	80
Figure 3.10: The relative abundance of genera observed in (A) Mock control 1 - equal volume, (B) Mock control 2- equal concentrations and (C) Positive control ( <i>E. coli</i> ). ....	81
Figure 3.11: The relative abundance of family observed in (A) Mock control 1-equal volume, (B) Mock control 2- equal concentrations and (C) Positive control ( <i>E. coli</i> ). ....	82
Figure 3.12: The relative abundance of genera observed in other sequencing controls. ....	83
Figure 4:1: Brief analysis workflow summary for alpha- and beta- diversity analyses.....	93
Figure 4:2: Taxonomic categorical levels, using the classification of <i>Escherichia coli</i> as an example. ....	95
Figure 4:3: The relative abundance of familia observed in the TB case group, categorized into (A) Bacteriologically confirmed TB and (B) Clinically diagnosed TB at baseline.....	98
Figure 4:4: The relative abundance of familia observed in the unlikely TB control group (well-defined ill controls) at baseline.....	98

Figure 4:5: Alpha Diversity between the TB case and ill control groups at baseline using (A) Shannon's Index and (B) Simpson's index as measures of diversity. ....	99
Figure 4:6: Bray Curtis dissimilarity metric heatmap between the samples at the family level. ...	101
Figure 4:7: Principal coordinate analysis (Bray-Curtis) showing the difference between respiratory samples obtained from the TB case and ill-control groups at all time points. ....	102
Figure 4:8: Principal coordinate analysis (Bray-Curtis) showing the difference between respiratory samples obtained from the TB case (bacteriologically confirmed TB or clinically diagnosed) and ill-control groups at baseline. ....	102
Figure 4:9: Differentially abundant familia in TB cases compared to ill controls at baseline. ....	103
Figure 4:10: The relative abundance of familia observed in samples obtained at month 2 from the TB case group, categorized into (A) Bacteriologically confirmed TB and (B) Clinically diagnosed TB. ....	106
Figure 4:11: The relative abundance of familia observed in samples obtained at month 6 from the TB case group, categorized into (A) Bacteriologically confirmed TB and (B) clinically diagnosed TB. ....	106
Figure 4:12: The relative abundance of familia observed in the unlikely TB control group (well-defined ill controls) at month 2. ....	108
Figure 4:13: The relative abundance of familia observed in the unlikely TB control group (well-defined ill controls) at month 6. ....	108
Figure 4:14: Alpha Diversity measures in TB case and ill-control groups during treatment based on the (A) Shannon and (B) Simpson Indices. ....	109
Figure 4:15: Principal coordinate analysis (Bray-Curtis) showing no differences between respiratory samples obtained from the TB case and ill-control group during TB treatment. ....	110
Figure 4.16: Differentially abundant taxa in month 6 TB cases. ....	110
Figure 4:17: Alpha Diversity comparison between baseline TB cases and month 6 ill controls (A) Shannon and (B) Simpson Index. ....	113

## CHAPTER 1: Literature Review

Tuberculosis (TB) is a renowned disease that contributes significantly to morbidity and mortality worldwide (Lewinsohn, Gennaro and Scholvinck, 2004), and was one of the first infectious diseases declared a global health emergency by the World Health Organization (WHO). TB is recognized as the leading cause of death from a single infectious agent, ranked above HIV/AIDS for the last 5 years (World Health Organization, 2019). Each year approximately 10 million people become ill and according to the WHO, approximately 1.2 million TB deaths were estimated among HIV negative people and an additional 251 000 deaths among HIV positive people in 2018 (World Health Organization, 2019). In children (< 15 years of age), 1.1 million incident cases are reported globally, indicating that approximately 11% of TB occurs in children (World Health Organization, 2019). In 2018, the global estimates for TB mortality in HIV negative and positive children (<15 years) in Africa were 60000 and 30000, respectively (World Health Organization, 2019). Although a relative small proportion of TB cases are reported in children, they represent ongoing transmission of TB from adults to children in communities (Tsai *et al.*, 2013). South Africa is classified as a high burden TB country, with one of the world's worst TB epidemics driven by HIV (Churchyard *et al.*, 2014).

### 1.1 Etiology of Tuberculosis

Robert Koch identified the etiological agent responsible for TB, the “tubercle bacillus”, in 1882 (Gradmann, 2006); it later became known as *Mycobacterium tuberculosis (M.tb)*. *Mycobacteria* are classified within the Mycobacteriaceae family and can be described as aerobic, non-motile, acid-fast bacilli (AFB) that are either straight or slightly curved (0.2-0.6 mm wide and 1-10 mm long). They can be classified according to the measurement of growth (slow or rapid growing) and the ability to produce pigment (photochromogens, scotochromogens and non-chromogens) (Saleem and Azher, 2013). For instance, *M.tb* can be described as an obligate aerobic, large non-motile, acid-fast bacillus, non-spore forming, catalase positive and oxidase negative bacterium and classified as a non-chromogenic (non-pigmented) slow growing bacterium with a generation time of 15-20 hours (Lawn and Zumla, 2011; Saleem and Azher, 2013; Dunn, Starke and Revell, 2016), which in practice contributes to diagnostic delays.

## 1.2 Clinical presentation

TB can manifest in many ways depending on the immune system of an individual. In some instances, an individual may become infected with TB without becoming diseased; that is often referred to as latent TB (Piccini *et al.*, 2014). During latent TB an infected individual does not present with symptoms of TB disease. In this case the immune system controls further spread by enclosing bacteria within a calcified shell, known as a granuloma. These granulomas protect the lungs from further damage and if these bacteria are contained, an individual will not present with symptoms or be contagious and therefore cannot spread the disease (Kim *et al.*, 2010). When an individual's immune system is compromised, latent TB may progress to active TB.

Although the granulomas contain the bacteria and act as a host defense mechanism, these granulomas provide an environment for the persistence of *M.tb* (Shaler *et al.*, 2013). The progression to active disease occurs when these granulomas rupture and *M.tb* is no longer contained, resulting in *M.tb* inhabiting the lungs, damaging surrounding tissue (Kim *et al.*, 2010) and reaching the part of the lungs that connects to the airway where bacilli can be expectorated (Ehlers and Schaible, 2012).

Without treatment, 5-10% of infected individuals will develop TB disease in their lifetime; approximately half of those who develop active TB will do so within 2 years of being infected (CDC, 2011). However, the risk of developing TB is greater in the presence of predisposing factors, especially HIV, which increases the risk 16-27 times ('WHO | Tuberculosis and HIV', 2019). Others at increased risk of progression to active TB disease are those affected by other conditions affecting the immune system like malnutrition, diabetes mellitus, smoking, sepsis, renal failure, chemotherapy, organ transplantation, chronic alcohol consumption and long term use of corticosteroids (Knechel, 2009; World Health Organization, 2018).

The clinical manifestation of TB is dependent on where in the body *M.tb* proliferates. Pulmonary TB (PTB) occurs in about 85% of TB patients and is therefore the most prominent form of TB disease (Fogel, 2015). The usual clinical signs and symptoms associated with PTB include chronic cough, night sweats, blood tinged sputum, weight loss, shortness of breath, fever chest pain and pleurisy (inflammation of the pleura membrane surrounding the lungs) (Fogel, 2015). However, these signs and symptoms are not always evident in all PTB cases and may vary between age groups. TB can also occur outside of the lungs (extrapulmonary TB); in the spine, hips, gastrointestinal tract or other areas.

### 1.3 Pediatric TB

Pediatric TB has not always been a priority of global TB control as much of the focus has been on the identification and effective management of the most infectious cases of TB to reduce the transmission of infection with *M.tb*. This group is usually represented by adults or adolescents with sputum smear positive PTB. Furthermore, the true burden of TB disease in children is obscured due to diagnostic difficulties involving both atypical clinical presentation and low confirmation rates due to paucibacillary disease.

Generally, children infected with TB have a higher risk of progressing towards disease within the first year after exposure or infection (Marais and Schaaf, 2010). Additionally, children are at a higher risk of developing more severe, disseminated forms of TB disease such as TB meningitis and miliary TB (Donald, Marais and Barry, 2010). Progression from infection to disease is determined by factors such as age at the time of exposure, nutritional and immune status, genetic factors, virulence of the organism and the magnitude of the initial infection. Children can develop TB within two to 12 months after initial TB infection (Cruz and Starke, 2007). The greatest risk of disease progression after infection is seen in young children <5 years of age (especially in those <2 years), where children between 5 and 10 years' experience a lower risk of disease progression, followed by an increase in risk in the adolescent group (>10 years) (Seddon *et al.*, 2018). It is recognized that TB primarily affects the lungs and it has previously been reported that pulmonary TB accounts for 60-80% of cases in children (Cruz and Starke, 2007).

Children with PTB present with symptoms such as a chronic, unremitting cough that does not improve and is present for more than two weeks, a fever of more than 38°C for at least two weeks with other common causes having been excluded, and weight loss or failure to thrive (Adams and Starke, 2019). These symptoms are however nonspecific, especially in the youngest children, and are similar to those of common childhood diseases, including pneumonia, generalized bacterial and viral infections, malnutrition and HIV infection (Tsai *et al.*, 2013). Schaaf *et al* (1995) found no differences with respect to weight loss, chronic cough and duration of symptoms between children with culture confirmed TB and other lung diseases. The only differences between the two groups were history of contact with an infectious TB case and a positive tuberculin skin test (TST) as a marker of TB infection.

The evidence of lung disease is not always clear, especially in children between the age of 5-10 years where they present with radiographically apparent clinically silent disease (Table 1.1). Infants and adolescents are more likely to be symptomatic and show physical signs of lung

disease (Cruz and Starke, 2007). HIV infected individuals can also present with nonspecific signs and symptoms, with clinical and radiographic features which overlap with other lung diseases, such as pneumonias or chronic lung diseases (Swaminathan, 2004). The nonspecific and overlapping respiratory features observed in pediatric TB contribute to the difficulty in diagnosing TB in children.

Table 1.1: The frequency of symptoms and signs of PTB according to age. (Cruz and Starke, 2007)

<b>Clinical feature</b>	<b>Infants (&lt;2 years)</b>	<b>Children (5-10 years)</b>	<b>Adolescents (&gt;10 years)</b>
<b>Symptom</b>			
Fever	common	uncommon	common
Night sweats	rare	rare	uncommon
Cough	common	common	common
Productive cough	rare	rare	common
Haemoptysis	never	rare	rare
Dyspnoea	common	rare	rare
<b>Sign</b>			
Rales	common	uncommon	rare
Wheezing	common	uncommon	uncommon
Fremitus	rare	rare	uncommon
Dullness to percussion	rare	rare	uncommon
Decreased breath sounds	common	rare	uncommon

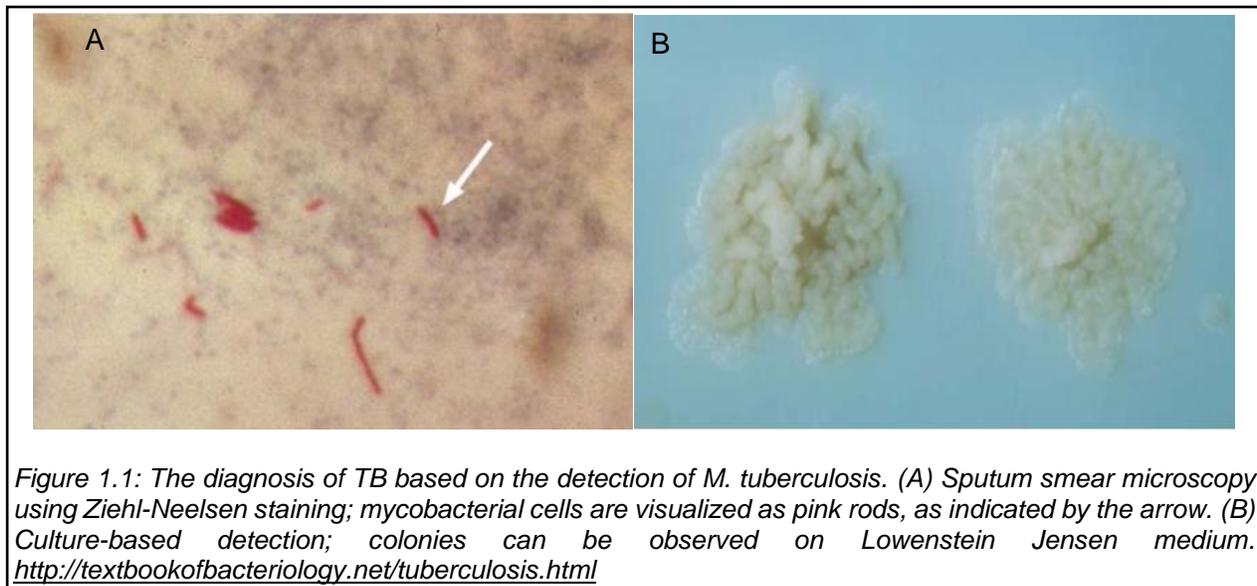
## 1.4 TB Diagnosis

Most TB deaths can be averted with early diagnosis and appropriate treatment. An estimated 58 million people/individuals were successfully treated for TB between 2000 and 2018, but there is still a gap between the number of people being notified and treated for TB and people becoming ill from TB (World Health Organization, 2019). The rapid diagnosis of PTB is difficult and the early detection of PTB continues to be challenging for clinicians, especially in young children (Ryu, 2015), since there is not a high quality diagnostic test available for paucibacillary disease.

Diagnosis of TB in children is based on bacterial confirmation or clinical presentation alone. *M.tb* detection is required for bacteriological confirmation of TB; this includes sputum smear microscopy, culture-based techniques and rapid molecular tests, like GeneXpert MTB/RIF (Ultra). Pediatric TB is typically paucibacillary; as fewer organisms are involved in the disease process. As a result only an estimated 40% of children with TB are bacteriologically confirmed ( $\pm 10\%$  smear positive) and the other 60% of cases are diagnosed based on a combination of clinical signs and symptoms, radiography, epidemiology and tests of infection (Chiang, Swanson and Starke, 2015; Dunn, Starke and Revell, 2016).

### 1.4.1 Bacteriological diagnosis

Sputum smear microscopy involves the examination of bacteria under a microscope after Ziehl-Neelsen (Figure 1.1) or Auramine O staining; which exploits the acid-fast nature of the mycobacterial cell envelope. Smear microscopy is a simple, yet rapid and inexpensive test for the diagnosis of pulmonary TB that offers good specificity, but has low sensitivity when it comes to detecting *M.tb* in patients with non-cavitary pulmonary disease or low bacillary load in sputum; this is particularly evident in children and HIV positive patients (Ryu, 2015; World Health Organization, 2017). Additionally, smear microscopy cannot distinguish between *M.tb* and other mycobacterial species.



*M.tb* culture is the gold standard for the diagnosis of *M.tb*, and allows for accurate speciation and phenotypic drug-susceptibility testing to be performed (Dunn, Starke and Revell, 2016). *M.tb* culture can be done on either solid (Lowenstein Jensen and Middlebrook 7H11) or liquid medium (Middlebrook 7H9) or on automated liquid culture systems such as the Bactec MGIT (BD) or BacT/Alert systems (Dunn, Starke and Revell, 2016). The automated system assists in reducing the detection time in comparison to solid medium culture which has a slow turnaround time of 4-6 weeks for culture positivity (Balajee and Dhana Rajan, 2011; World Health Organization, 2017).

Technological advancements have allowed the introduction of the GeneXpert MTB/Rif, a rapid molecular test that has become an important diagnostic measure for the detection of *M.tb* and drug resistance. This automated diagnostic technique detects *M.tb* DNA directly from sputum samples and is able to detect mutations associated with resistance to rifampicin (RIF) using nucleic acid PCR amplification. In 2010, the WHO endorsed the use of the GeneXpert MTB/Rif in TB endemic areas such as South Africa. In 2017, the GeneXpert MTB/Rif Ultra was launched and was found to be significantly more sensitive compared to the Xpert MTB/Rif for the detection of low bacillary loads. This technique is particularly useful in the case of smear negative TB patients, culture positive specimens, extrapulmonary specimens (CSF) and specimens obtained from children.

Drug susceptibility testing is important for accurate treatment. In addition to the GeneXpert, rapid drug susceptibility testing can be done using line probe assays (LPA) and sequencing technologies. LPAs test for resistance to RIF and isoniazid (INH) (first line LPAs) and resistance

to fluoroquinolones and injectable anti TB drugs (second-line LPAs). However, culture-based methods currently remain the reference standard for drug susceptibility testing (World Health Organization, 2017).

#### 1.4.2 Diagnosis of pediatric TB

Accurate diagnosis and confirmation of pediatric TB is particularly challenging and is exacerbated by two biological factors, namely the inability to expectorate sputum and the paucibacillary nature of childhood TB (Chiang, Swanson and Starke, 2015). To circumvent challenges in obtaining sputum from children, alternative specimen collection methods can be used, such as gastric aspiration, induced sputum, string test, nasopharyngeal aspiration, bronchoalveolar lavage, stool or urine (Marais and Schaaf, 2010). However, each method is not without benefit or limitation (Marais and Schaaf, 2010). Due to the paucibacillary nature of the disease test sensitivities are reduced, especially in acid fast smear microscopy and mycobacterial culture tests (Chiang, Swanson and Starke, 2015). Even with meticulous specimen collection, only 10-15% of sputum samples reveal acid fast bacilli (AFB) and approximately 30-40% of sputum cultures remain negative in probable pediatric TB cases (Marais *et al.*, 2006). As a result, bacteriological confirmation is achievable in less than 50% of children and 75% of infants (Adams and Starke, 2019).

In cases where bacteriological confirmation is not possible, PTB is diagnosed using clinical criteria such as signs and symptoms, tests for TB infection (tuberculin skin test; TST or interferon gamma release assay; IGRA), exposure history and radiographic findings. However, the TST and IGRA tests only determine infection status without information about recent or old infection. A positive test of infection in high burden countries and nonspecific clinical and radiographic findings can contribute to uncertainty in clinical diagnosis (Chiang, Swanson and Starke, 2015). Nevertheless, the diagnosis of TB in children is largely based on, (1) well defined symptoms (Marais *et al.*, 2004), 2) recent close contact with an infectious TB case, (3) a positive TST or IGRA result, and (4) suggestive findings on chest radiograph or physical examination (Adams and Starke, 2019).

## 1.5 TB Treatment

In 1940, it was discovered that streptomycin killed *M.tb*, however the success was short lived as it was realized that *M.tb* had the propensity to acquire/develop antibiotic resistance. This led to the establishment of the initial combined treatment therapy with the following antibiotics; streptomycin, para-aminosalicylic acid (PAS) and INH (Sotgiu *et al.*, 2015). The principle of combined treatment therapy is still in practice, aiming to eliminate the actively replicating and dormant/near dormant mycobacteria by using a combination of drugs with varying actions, while attempting to prevent the emergence of drug resistant organisms; and all being achieved with a minimum risk of toxicity (Graham, 2011).

Current treatment regimens involve a cocktail of drugs that is taken in two phases, an initial intensive phase and continuation phase. The cocktail of drugs is taken intensively for 2 months (intensive phase) to kill rapidly growing bacilli (bactericidal drugs) aiming to reduce the microbial load. This reduces the inflammation process, symptoms and clinical signs (clinical recovery) and terminates disease transmission (Sotgiu *et al.*, 2015). This is followed by a 4 month continuation phase to eradicate slower growing more persistent bacilli and those in acidic environments, to prevent relapse (sterilizing drugs) (Graham, 2011; Tsai *et al.*, 2013; Sotgiu *et al.*, 2015). In cases of drug susceptible TB, the two drugs INH and RIF are usually sufficient during the continuous phase, whereas in the initial phase pyrazinamide (PZA) with or without ethambutol (EMB) is added.

Generally, TB treatment drugs are divided into first line drugs (FLDs) and second line drugs (SLDs). FLDs include INH, RIF, PZA and EMB and are used in combination in cases of drug susceptible TB. SLDs are used for the treatment of multidrug resistant TB (MDR-TB), which is resistance to two of the most effective FLDs, RIF and INH (World Health Organization, 2013). SLDs include 6 classes of drugs, namely aminoglycosides, polypeptides, fluoroquinolones, thioamides, cycloserine and para-aminosalicylic acid (Saleem and Azher, 2013); of which fluoroquinolones and aminoglycosides (second line injectable agents, amikacin (AMK), capreomycin (CAP) or kanamycin (KAN)) are the main drugs used (Jnawali and Ryoo, 2013). New drug regimens including delamanid and bedaquiline have been introduced and bedaquiline has been approved for use in adolescents and adults in South Africa.

In more recent years extensively drug resistant TB (XDR-TB) has been reported; defined as resistance to RIF and INH with the additional resistance to at least one drug in each of the two most important classes of drugs in an MDR-TB regimen, the fluoroquinolones and second line

injectable agents (CDC, 2013). Drug resistance contributes to the persistence of TB which threatens global TB care and prevention as the remaining treatment options are limited, less effective, have more side effects and are expensive, especially in low income countries (Sotgiu *et al.*, 2015).

The principles of TB treatment are the same for adults and children (Tsai *et al.*, 2013), consisting of the intensive and continuation phases to rapidly kill and eradicate slower persistent bacilli, respectively. Due to the fact that laboratory confirmation is uncommon and often delayed in children, treatment is often guided by the culture and drug susceptibility results of the index case. According to the South African guidelines for the management of TB in children, during uncomplicated TB disease (e.g. low bacillary load such as PTB with minimal lung parenchyma involvement) the TB treatment regimen is comprised of 3 or 4 drugs (RIF, INH, PZA without or with EMB) for the first two months (intensive phase), followed by 2 drugs (RIF and INH) during the continuation phase lasting for 4 months (Table 1.2). The continuation phase could be extended to 7 months during complicated TB (e.g. high bacillary load PTB- smear positive, parenchymal involvement and cavities on chest X ray). On the other hand, the length of treatment for drug resistant TB will be 12 months or more depending on the extent of the disease. Also, an extended treatment regimen of 9 months may be considered for HIV infected and HIV uninfected children as a result of a slower treatment response rate. The duration of treatment and dosing of drugs are all based on adult data as pediatric data is lacking. Studies are underway to investigate treatment shorting in children with TB. Drug dosages used during treatment depends on the weight of the child and is adjusted accordingly throughout the course of treatment (South African National Department of Health, 2013).

Table 1.2: First line and second line anti-TB drugs recommended for children (South African National Department of Health, 2013); mechanisms of action, spectrum of activity and activity against other Gram positive and Gram-negative bacteria (Naidoo et al., 2019).

Antibiotic name	Antibiotic action	Mechanism of action	Spectrum of activity	Activity against	
				G+	G-
<b>Drug susceptible TB: First line drugs (FLDs)</b>					
Isoniazid	Bactericidal/static	Inhibits mycolic acid and acyl carrier protein reductase	Narrow ( <i>M.tb</i> , <i>M. kansasii</i> , <i>M. xenopi</i> )	N	N
Rifampicin	Bactericidal	Targets DNA dependent RNA polymerase	broad	Y	Y
Pyrazinamide	Bactericidal	Targets membrane energy metabolism	Narrow ( <i>M.tb</i> )	N	N
Ethambutol	Bacteriostatic	Targets arabinosyl transferase	Narrow ( <i>M.tb</i> , <i>M. avium</i> )	N	N
<b>Drug resistant TB: Second Line drugs (SLDs)</b>					
Ethionamide/ Prothionamide	weakly Bactericidal	Targets peptide synthesis	Narrow	Y	Y
<b>Fluoroquinolones</b>					
Levofloxacin	Bactericidal	Inhibits DNA replication	Broad	Y	Y
Moxifloxacin					
<b>Aminoglycosides</b>					
Kanamycin	Bactericidal	Inhibition of protein synthesis	Broad	Y	Y
Amikacin					
Capreomycin					
Terizidone/ Cycloserine	Bacteriostatic	Inhibition of cell wall synthesis	Broad	Y	Y
Para-aminosalicylic	Bacteriostatic	Inhibition of folic acid synthesis and cell wall synthesis	Narrow	N	N
<b>New drugs</b>					
Bedaquiline	Bactericidal	Targets ATP synthesis	Narrow	N	N
Delamanid	Bactericidal	Targets mycolic acid synthesis	Narrow	N	N

First and second-line agents used in pediatric TB treatment. G+: Gram positive bacteria, G-: Gram negative bacteria, Y: Yes, N: No

## 1.6 TB and the Microbiome

Respiratory health has been linked to the microbial colonization of both the upper and lower respiratory tract. Biesbroek *et al* (2014) provided insight into microbial succession in the respiratory tract in infancy and linked early life profiles to microbiota stability and respiratory health characteristics. Understanding this is important, as children are prone to the development of respiratory infections due to an immature immune system.

The susceptibility to respiratory infections may be linked to the origin of respiratory infections, namely the nasopharynx (van den Bergh *et al.*, 2012). The rich microbial carriage in the nasopharynx is seen as a reservoir for both commensals and potential or invasive pathogens (García-Rodríguez and Fresnadillo Martínez, 2002), which plays an important role in microbial spread as well as disease development. A multitude of microorganisms inhabit this site including viruses and fungi, however it is particularly hospitable to bacterial species (Mizgerd, 2014); all of which form part of the microbiome, a term that was initially used to “signify the ecological community of commensal, symbiotic and pathogenic microorganisms that share our body space”(The NIH HMP Working Group, 2009). The microbiome has been described to aid in maintaining normal host physiology, developing and educating the immune system, metabolizing complex substrates and providing crucial protection against opportunistic pathogens (Shukla *et al.*, 2017). Studies have shown that our microbiomes change us, by promoting health through their beneficial actions or by increasing susceptibility to disease through a phenomenon called dysbiosis (Gerber, 2014).

### 1.6.1 Nasopharyngeal microbiome

The organisms found in in the respiratory tract form part of the respiratory microbiome. The respiratory tract is a composite system that is anatomically divided into the upper and lower respiratory tract. Its surface is completely inhabited by niche specific bacterial communities (as well as viruses and fungi) of which the upper respiratory tract has the highest bacterial densities (Man, de Steenhuijsen Piters and Bogaert, 2017).

The first description of the nasopharyngeal “microbiome” in children was executed by Bogaert *et al* (2011), where 5 prominent phyla were identified, namely *Proteobacteria*, *Firmicutes*, *Bacteroidetes*, *Actinobacteria* and *Fusobacteria* with *Moraxella*, *Haemophilus*, *Streptococcus*, *Flavobacteria*, *Dolosigranulum*, *Corynebacterium* and *Neisseria* being the most predominant genera. This study also showed that the microbiota in the nasopharynx is highly diverse and that the microbiota in a given niche can change (Bogaert *et al.*, 2011).

Man *et al* (2017) reviewed numerous studies and found that a wide variety of microorganisms can be detected in the upper respiratory tract during the first hours of life of healthy term neonates. Furthermore, within the first week of life, niche differentiation in the upper respiratory tract leads to the accumulation of bacterial such as *Staphylococcal*, *Corynebacterium*, *Dolosigranulum* and *Moraxella spp*; of which microbiota profiles characterized by *Corynebacterium* and *Dolosigranulum* early in life and *Moraxella spp* at 4-6 months of age correspond with a stable bacterial composition and respiratory health. However, early life microbiota are highly dynamic and can be attributed to both host and environmental factors (Figure 1.2) such as mode of delivery, feeding type, siblings, season and antibiotic exposure (Bogaert *et al.*, 2011; Man, de Steenhuijsen Piters and Bogaert, 2017; Esposito and Principi, 2018) with environmental factors having the largest described influence.

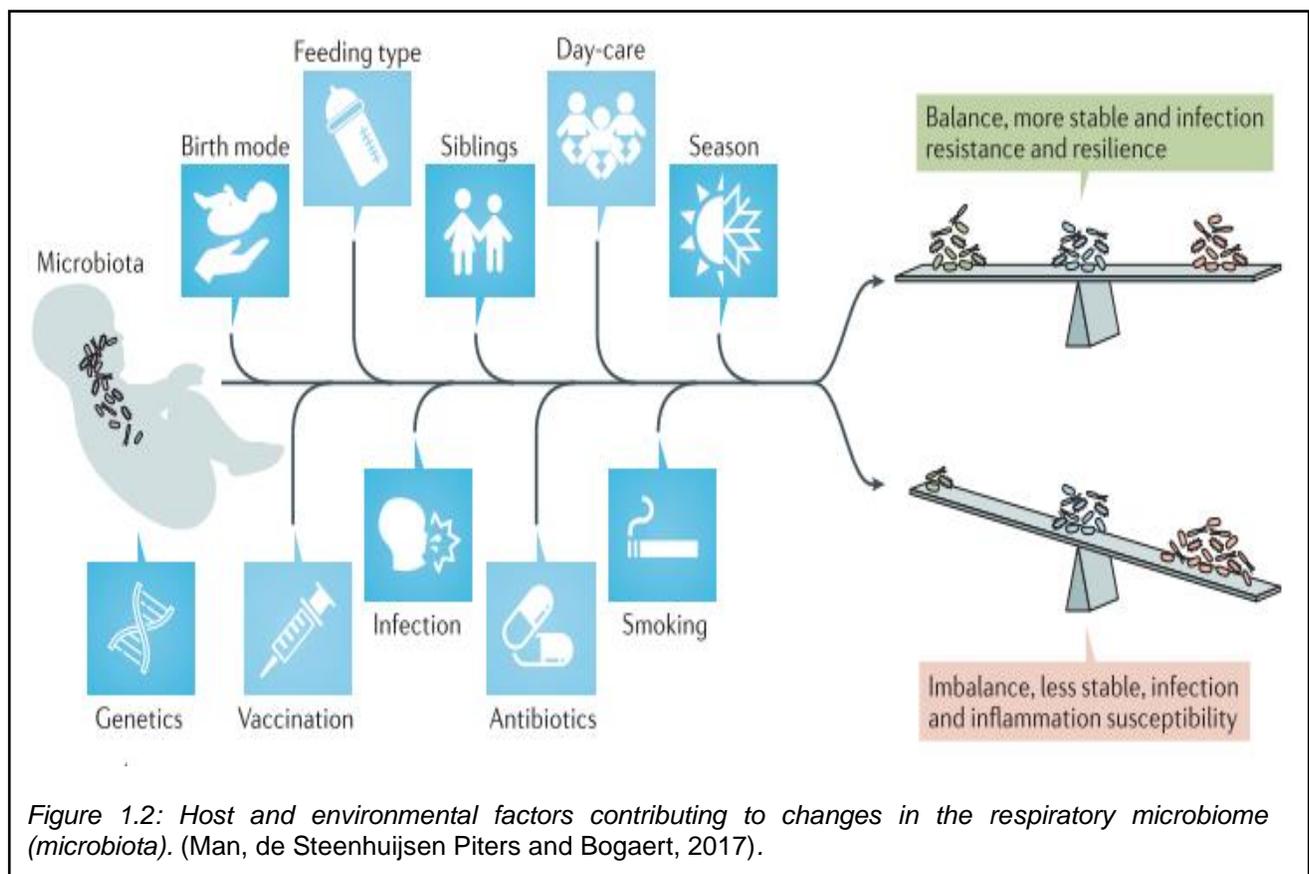


Figure 1.2: Host and environmental factors contributing to changes in the respiratory microbiome (microbiota). (Man, de Steenhuijsen Piters and Bogaert, 2017).

### 1.6.2 The respiratory microbiota and clinical presentation of PTB

Pediatric TB does not present as obviously as TB in adults, and the challenge is that PTB presents similarly to other lower respiratory tract infections. Studies have shown that lower respiratory infections are often polymicrobial (Dube *et al.*, 2016; Zar *et al.*, 2016), and that other respiratory microorganisms (or pathogens) may be the cause of the initial clinical suspicion of TB or contribute to the severity of PTB. Dube *et al.* (2016) described the prevalence of respiratory pathogens in children hospitalized with suspected PTB in Cape Town, South Africa. They found that in 97% of all children with suspected PTB, another respiratory pathogen, including bacteria and viruses, could be detected. The children with PTB had a microbial profile which consisted of *C. pneumoniae*, hMPV, coronavirus 043, influenza C virus, rhinovirus and cytomegalovirus and those without TB had a microbial profile which consisted of *P. jirovecii*, *H. influenzae spp*, RSV, *M. pneumoniae*, influenzae B virus and enteroviruses. However, their results of differences in microbial profile between children with and without PTB failed to reach statistical significance and warranted further investigation (Dube *et al.*, 2016). Schaaf *et al.* (1995) found that 42% of the children with initial suspicion were found not to have TB and were diagnosed with bacterial or viral pneumonia, bronchopneumonia or asthma, which implies that the presentation of these diseases are similar and symptoms can overlap (Schaaf *et al.*, 1995). However, limited data is available with regard to other respiratory pathogens and their role in pediatric PTB. Therefore, investigating the prevalence of respiratory co-infections in children with TB may provide insight into their role in the clinical presentation and the pathogenesis of PTB.

### 1.6.3 The microbiome and diagnosis of PTB

The diagnosis of pediatric TB is challenging, as previously discussed. Microbiome research has shown that during certain diseases specific microbial profiles can be identified. The composition of microbial communities in the nasopharynx seems to differ between different disease states with certain phyla and genera associated with different diseases. Infants with cystic fibrosis have been shown to have nasopharyngeal microbial profiles with *Staphylococcus aureus*, *Streptococcus mitis*, *Corynebacterium accolens* and bacilli as the most abundant organisms, while in healthy controls, *Corynebacterium spp* and *Haemophilus influenzae* were more abundant (Prevaes *et al.*, 2016). Microbial profiles were also found to differ between children with pneumonia and healthy children. Children with pneumonia were found to have less diverse microbial communities in comparison to healthy participants (Sakwinska *et al.*, 2014). The decrease in richness and diversity of the microbiota was shown to be associated with disease and is a common theme in many conditions, particularly for nasopharyngeal and nasal microbiota (Sakwinska *et al.*, 2014).

To our knowledge there is no data available on the respiratory microbiome of children with PTB; before the start of TB treatment, during and after treatment. A better understanding of the respiratory microbiome of children with and without PTB could contribute to the understanding of the role that the microbiome plays in TB pathogenesis.

#### 1.6.4 TB treatment and the microbiome

The main objective of combined TB treatment regimens is to eliminate *M.tb* while preventing resistance with minimum risk of toxicity. However, drug resistance and drug toxicity may not be the only issues to consider. Most of the antibiotics used in the TB treatment regimen, especially the second line drugs, have broad-spectrum activity (Table 1.2). As such, other organisms besides *M.tb* may become unintentional targets during the process of eradicating *M.tb*. Recent microbiome research has shown that although antibiotics are necessary to combat disease or infection, they present an external interference that contributes to microbial imbalances on or inside the body as a consequence of dysbiosis. This is a well described phenomenon that has been described in gut microbiome research, but has since been suggested to occur on any exposed surface or mucus membrane such as the vagina, skin or the respiratory system (Martín *et al.*, 2014). Like the gut microbiome, the respiratory microbiota is established at birth with subsequent changes continuing for several months, and it has been suggested that early respiratory microbiota composition determines respiratory health in children (Man, de Steenhuijsen Piters and Bogaert, 2017; Esposito and Principi, 2018). However, the microbial communities that inhabit different niches in our bodies change throughout our lives and these changes are attributed to many factors; progression during childhood, altered diets, travel, illness and treatment regimens. The risk of treatment regimens such as TB treatment regimens causing dysbiosis is concerning in children since their microbiomes are still being established. Especially since dysbiosis causes the disruption of either the composition or overall numbers of “normal microbiota” which can result in the outgrowth of dominant, usually pathogenic bacterial genera over the diverse microbial community, which may lead to the development, progression or exacerbation of disease (Shukla *et al.*, 2017). This is especially important for TB treatment as treatment is taken for an extended period (for a minimum of six months) compared to other antibiotic courses that usually last a maximum of 1-2 weeks.

The effects of first line TB treatment have previously been described to cause intestinal microbiome dysbiosis in humans and mice. Wipperman *et al* (2017) found that TB treatment did not affect the overall diversity of the intestinal microbiome in humans but that it reduced multiple immunologically relevant commensal bacteria. Also, that treatment can have long lasting effects on the microbiome since dysbiosis was seen to persist after treatment (Wipperman *et al.*, 2017). In the mouse model study, they found that TB treatment had a temporary effect on intestinal microbial diversity and that the altered microbial structure as a result of therapy persisted up to 3 months after therapy ended. The study also compared monotherapy and combination therapy of the first line TB drugs and found that rifampin (RIF) was a major contributor to the altered microbial structure (Namasivayam *et al.*, 2017). These studies show that although FLDs (INH, PZA EMB and RIF) primarily target *M.tb* they can cause dysbiosis which persists after treatment. Determining the effect of FLDs on the nasopharyngeal microbiome is important since RIF belongs to the rifamycin family, which is said to have broad spectrum of activity on Gram positive bacteria of the skin and respiratory microbial communities (Hong *et al.*, 2016). This is important as during suspicion of TB children may be placed on TB treatment without bacteriological confirmation of disease and monitored for improvement (South African National Department of Health, 2013), which could potentially have adverse effects on the nasopharyngeal microbiome. Recognizing TB treatment as a potential cause of antibiotic induced dysbiosis may potentially allow for interventional and treatment strategies, such as the use of probiotics or specific vaccinations, to be implemented.

## 1.7 Problem statement

TB is a worldwide problem that results in morbidity and mortality and has a devastating social and economic impact. Pediatric TB is particularly challenging due to difficulties in diagnosis as a result of non-specific signs and symptoms, the paucibacillary nature of the disease and the overlap with other common childhood illnesses in these age groups. The nasopharyngeal microbiome (microbiota) is critical for respiratory health and impacts on the development, presentation and diagnosis of TB disease. In addition, TB treatment may result in dysbiosis of the nasopharyngeal microbiome which may lead to the development, progression or exacerbation of other diseases. There is limited data describing the nasopharyngeal profiles of children with and without TB, and the effect of TB treatment on the nasopharyngeal microbiota/microbiome.

## 1.8 Aims and objectives

This is a pilot study which aims to

- (1) Compare the respiratory microbiota of children with bacteriologically confirmed, clinically diagnosed and unlikely PTB
- (2) Describe the effect of TB treatment on the respiratory microbiota in children with PTB.

The aims will be achieved by completing the following objectives:

1. Determine the presence of various bacterial pathogens and the fungal pathogen *Pneumocystis jirovecii* in respiratory samples collected from children with bacteriologically confirmed, clinically diagnosed and unlikely PTB.
2. Describe the respiratory microbiota in children with PTB (bacteriologically confirmed/clinically diagnosed) and in those without PTB (well defined ill controls) at baseline.
3. Describe the effect of TB treatment on the respiratory microbiota in children with PTB (bacteriologically confirmed/clinically diagnosed) after 2 and 6 months in comparison to those without PTB (well defined ill controls).

## 1.9 Study Population

This substudy forms part of a larger ongoing prospective cohort study conducted by the Desmond Tutu TB Centre, which aims to improve the diagnosis of pulmonary tuberculosis (PTB) in children with suspected PTB. Children with suspected PTB are enrolled from the Tygerberg and Karl Bremer (respectively, tertiary and secondary level) provincial hospitals in Cape Town and are routinely followed up for six months.

Ethical approval for this study was obtained from the Health Research Ethics Committee of Stellenbosch University (parent study N11/09/282- PI Dr E Walters and substudy N15/04/034).

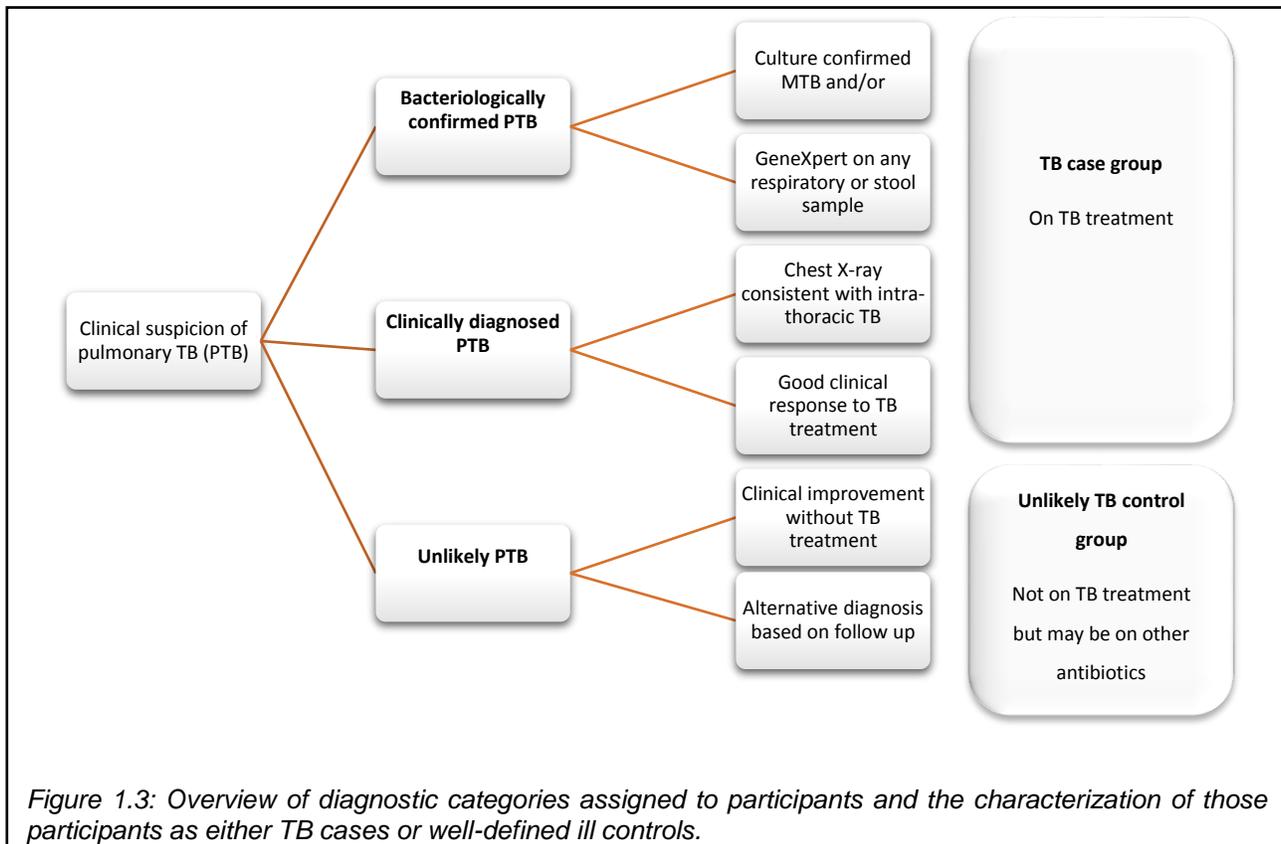
### Inclusion Criteria:

1. Children aged 0-13 years
2. A child with more than one of the following (Graham *et al.*, 2012):
  - i. Persistent unremitting cough (or cough significantly worse than usual in child with chronic lung disease, including HIV-related) of >2 weeks duration, unresponsive to a course of appropriate antibiotics.
  - ii. Poor growth documented over the preceding 3 months [clear deviation from a previous growth trajectory and/or documented crossing of centile lines in the preceding 3 months; and/or weight-for-age Z score of  $\leq 2$  in the absence of information on previous/recent growth trajectory AND not responding to nutritional rehabilitation (or to antiretroviral therapy if HIV infected).
  - iii. Persistent unexplained lethargy or reduced playfulness/activity reported by the caregiver.
  - iv. Any duration of cough with at least one of the following:
    - i. Documented exposure to a known TB source case (regardless of smear status) OR
    - ii. Reactive Mantoux skin test OR
    - iii. Chest radiograph suggestive of TB (Marais *et al.*, 2004).
3. Written consent provided by the parent/ legal guardian for study participation, including HIV testing.

Exclusion Criteria:

1. Presence of only extra-thoracic TB without signs of PTB.
2. Receipt of TB treatment for >2 days in the previous two weeks.
3. Severe illness resulting in unstable clinical condition.
4. Any condition which would constitute an absolute contra-indication to any of the sampling procedures required by the study e.g. acute severe asthma, pertussis-syndrome or raised intracranial pressure as contra-indications for sputum induction.
5. Residence in remote areas with no ready access to transport for follow-up visits.

Following enrolment, participants with suspected PTB based on the eligibility criteria were classified as having “bacteriologically confirmed PTB” (TB confirmed by liquid mycobacterial culture and/or GeneXpert MTB/RIF), “clinically diagnosed PTB” (without bacteriological confirmation but with clinical and radiological evidence of PTB disease), or “unlikely PTB” (TB excluded after intensive investigation based on alternative diagnosis and/or clinical improvement without TB treatment) (Figure 1.3). Diagnostic categories were only assigned once all diagnostic test results were obtained (after a maximum of 8 weeks after enrolment) and after careful follow up. Treatment as per standard care was determined by a hospital clinician after baseline sampling at the entry to the study. Participants who started TB treatment were defined as TB cases and included participants that were assigned to either clinically diagnosed or bacteriologically confirmed PTB categories. Those not on TB treatment were defined as well-defined ill controls (unlikely PTB group), which included participants that were determined not to have PTB and were therefore not on TB treatment (but may be on other antibiotics) (Figure 1.3). Clinical assessment during the period of follow up was systematically documented.



### 1.10 Sample collection

Gastric aspirates (GA) and induced sputa (IS) were collected from all of the enrolled children to detect *Mycobacterium tuberculosis* (*M.tb*) using routine diagnostic procedures. In addition, a nasopharyngeal aspirate (NPA) was collected from a subgroup of children for viral, bacterial and microbiome analysis; and used in this sub study. NPAs were obtained from study participants at the following time points: baseline (entry into study) and follow-up at 1, 2 and 6 months after enrolment or start of TB treatment. In some cases, IS samples were collected instead of NPAs. The samples were collected between December 2015 and January 2017.

NPAs were collected by the study team after a minimum of 2 hours nil per os (NPO) (fasting). The collection of the NPA samples was done according to the standard operating procedure (SOP PC012- Nasopharyngeal aspiration) by well-trained study nurses. At least 1 ml of specimen was collected and stored out of direct sunlight, at 4-8°C, until it was transported to the laboratory. Commercial viral transport medium (Davies Diagnostics, Grenada Spain) was added to the samples at the Division of Medical Virology, Stellenbosch University/National Health Laboratory Service.

## CHAPTER 2:

# Detection of “other” respiratory pathogens in children suspected of pulmonary tuberculosis

## 2.1 Introduction

TB is of concern globally; however South Africa has been recognized as one of the high TB burden countries (World Health Organization, 2018). Davies and colleagues (2005) suggested that two aspects likely contribute to the risk of TB development: (1) the risk of an individual being infected is contingent on the incidence of TB in the community (i.e. work and living) and (2) the risk of infection leading to disease is contingent on several factors that impinge on an individual (i.e. age, maturity of the immune system, genetics and environmental factors). Respiratory co-infection may play an important role in the risk of progression towards TB disease by influencing the immune response of the host. However, limited data is available with regard to the involvement of other respiratory pathogens in PTB, particularly in children (Dube *et al.*, 2016).

Microbial colonization of both the upper and the lower respiratory tract plays a major role in respiratory health. The significance of the upper respiratory airway is that it serves as an entry point for microbes into the body and a connective channel between the outside world and the lower respiratory tract. The nasopharynx forms part of the upper respiratory tract and is particularly hospitable to bacteria (Mizgerd, 2014). It is densely colonized by a wide range of microorganisms including commensal bacteria and pathobionts (potential pathogens) such as *Streptococcus pneumoniae*, *Haemophilus influenzae* and *Moraxella catarrhalis* (García-Rodríguez and Fresnadillo Martínez, 2002). It is likely that most individuals are colonized with these pathogens at least once early in life and as many as 54% and 33% of children are colonized with *S. pneumoniae* and *H. influenzae* respectively by the age of one (Faden *et al.*, 1997). These bacteria can be carried without causing clinical symptoms, however when conditions in the host are altered, invasion of adjacent sites and/or the bloodstream can lead to disease (García-Rodríguez and Fresnadillo Martínez, 2002).

Community acquired pneumonia (CAP) is an important cause of morbidity and mortality worldwide. Both *S. pneumoniae* and *H. influenzae* have been reported as etiological agents of CAP, with *S. pneumoniae* as the most common cause (Ruiz *et al.*, 1999; Cillóniz *et al.*, 2011). The Pneumonia Etiology Research for Child Health (PERCH) study group found that in children, *S. pneumoniae* was the most common bacterium isolated from culture and that viruses were a

major cause of pneumonia, except in severe pneumonia cases where bacteria were more common (O'Brien *et al.*, 2019). Other bacterial agents involved in CAP are atypical pathogens *Mycoplasma pneumoniae*, *Chlamydia pneumoniae*, *Coxiella burnetii* and *Legionella pneumophila* which have been described in up to 35% of CAP episodes (Cillóniz *et al.*, 2011). Other organisms that are less likely to cause pneumonia but have been reported are *Bordetella pertussis* and *Bordetella parapertussis* (Elahi *et al.*, 2008). These are closely related Gram-negative bacterial species that typically cause whooping cough.

The fungal pathogen *Pneumocystis jirovecii* is also a cause of pneumonia in immunocompromised hosts, such as cancer patients undergoing chemotherapy, individuals taking immunosuppressants and most commonly, HIV-infected people (Truong and Ashurst, 2019). Immunocompetent infants infected with *P. jirovecii* can be asymptomatic carriers or have mild respiratory disease (Morris *et al.*, 2008).

These pathobionts and pathogens are clinically significant causes of respiratory diseases to which children are particularly susceptible, however the relationship between these organisms and TB disease in children has yet to be explored. Co-infection or colonization with other respiratory pathogens may influence susceptibility and possibly the severity of respiratory diseases such as TB. Various studies have shown the involvement, interaction and implication of respiratory pathogens in respiratory diseases such as pneumonia, asthma and otitis media, and even in suspected TB in adults (Bosch *et al.*, 2013, 2016; Brealey *et al.*, 2015; Mhimbira *et al.*, 2018). However, less focus has been placed on the identification of respiratory pathogens in children with suspected PTB (Dube *et al.*, 2016). Dube *et al.* (2016) found that in children under the age of 15 years (median age 36 months) presenting with PTB in South Africa, the most common bacteria identified were *M. catarrhalis*, *S. pneumoniae*, *H. influenzae* spp and *Staphylococcus aureus*, with less common bacteria identified as *M. pneumoniae*, *B. pertussis* and *C. pneumoniae*. The study also included viruses and found the most common to be metapneumovirus, rhinovirus, influenza virus C, adenovirus, cytomegalovirus and coronavirus O43. Both viruses and bacteria were identified. This study showed that multiple potential pathogens are present in the nasopharynx of children presenting with TB. The identification of these organisms may contribute to our understanding of the clinical presentation in children during suspicion of TB disease and to our understanding of the pathogenesis of tuberculosis disease in children.

The aim of this chapter was to describe the presence of respiratory pathogens in respiratory samples from children with suspected PTB, classified as bacteriologically confirmed, clinically diagnosed or unlikely PTB after diagnostic evaluation. Respiratory samples were subjected to two

real-time Polymerase Chain reaction (PCR) assays; the Seegene Allplex™ Respiratory Panel 4 and an “in-house” real-time PCR assay for the detection of multiple respiratory pathogens that are of clinical significance. In addition, we studied the associations between various risk factors and the carriage of respiratory pathogens.

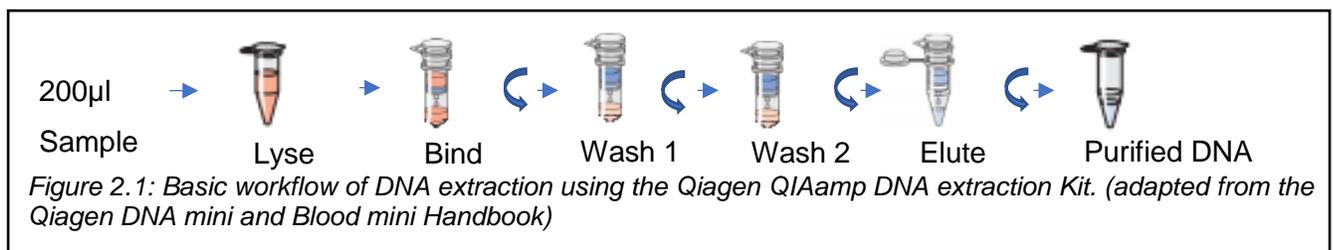
## 2.2 Materials and Methods

### 2.2.1 Sample collection

Baseline respiratory samples were obtained from children with suspected TB recruited for the study. The baseline samples included nasopharyngeal aspirates (NPA) and induced sputum (IS) samples, as outlined in Section 1.10 (Chapter 1). Samples were stored at -20°C prior to DNA extraction, as it was the temperature at which the samples were stored prior to receipt for our study. Although not ideal, samples were kept at this temperature to ensure consistency, for the microbiome analysis (Chapters 3-4).

### 2.2.2 DNA extraction

DNA was isolated from the NPA and IS samples using the QIAamp DNA mini kit (Qiagen, Netherlands) as per adaptation of the Body Fluid Protocol in the Qiagen QIAamp DNA Mini and Blood Mini Handbook (Addendum 1). The basic workflow of the DNA extraction procedure is described in Figure 2.1. Two hundred microliters of each sample was added to 20 µl Proteinase K and lysed with 200 µl AL lysis buffer in a microcentrifuge tube. The contents were then transferred to a mini spin column, where the DNA was adsorbed onto the QIAamp silica membrane during a brief centrifugation step. The DNA bound to the QIAamp membrane was washed twice to remove residual contaminants. Purified DNA was eluted in 50 µl AE buffer and incubated at room temperature for 2 min to increase the yield. The extracted DNA was stored at -20°C.



The purity (A260/A280 and A260/A230 ratios) and DNA concentration (ng/μl) of the isolated DNA was measured using the BioDrop μLite (Biodrop, United Kingdom). Based on the DNA yield, samples with a DNA concentration >50 ng/μl were diluted to approximately 50 ng/μl to avoid possible PCR inhibition due to high starting template nucleic acid concentrations.

### 2.2.3 Seegene Allplex Respiratory Panel 4 PCR

The Seegene Allplex™ Respiratory Panel 4 (Seegene Allplex RP4) assay is a qualitative (*in vitro*) multiplex real-time PCR assay that allows for the simultaneous amplification and detection of single or multiple pathogens, namely, *S. pneumoniae*, *L. pneumophila*, *H. influenzae*, *B. parapertussis*, *B. pertussis*, *M. pneumoniae*, *C. pneumoniae*, with a limit of detection of 100 copies.

The Seegene Allplex Respiratory Panel assays use Seegene's proprietary technologies: Multiple Detection Temperature (MuDT), Tagging Oligonucleotide Cleavage Extension (TOCE), Dual Priming Oligonucleotides (DPO), Real Amplicon Detection (ReAD) and Annealing Control Primer (ACP) to enable the detection of multiple targets in a single fluorescence channel without melting curve analysis (Lee *et al.*, 2015). The Seegene Allplex RP4 primarily exhibits MuDT™ based on multi-Ct (threshold cycle) values in a single fluorescence channel without melt curve analysis for the detection of multiple targets, on real-time PCR instruments. This technology is advantageous because it overcomes the present technology barrier of "one channel, one Ct". TOCE utilizes primer oligonucleotide (DPO) pairs, pitchers and catchers which are designed to detect a specific DNA target (Lee *et al.*, 2015). The real-time PCR preparation was executed as described by the manufacturer. Two separate master mixes were prepared. The Respiratory Panel-Bacteria Internal Control (RP-B IC) was included in the master mix for the samples to detect PCR inhibition, but not for the controls as the IC is premixed in the positive control sample (RP4 PC) and is excluded from the negative control where no amplification is expected. All reagents were briefly vortexed and centrifuged before use.

The PCR master mix was prepared as follows for the samples (per reaction): 5 μl 5X MuDt Oligo Mix (RP4 MOM, primer), 5 μl 5X Anyplex PCR master mix (with UDG, premix), 2 μl RP-B Internal Control and 5 μl RNase-free water. Following PCR master mix preparation, 8 μl of sample nucleic acid was added to each of the white eight-strip low-profile PCR tubes (Biorad Laboratories, USA) containing an aliquot of PCR master mix. The PCR master mix for the controls was prepared in the same manner but without the RP-B IC, and the nucleic acid was substituted with 8 μl of RP4

Positive Control (positive control) or 8 µl of RNase free water (negative control). The eight strip PCR tubes were sealed with optical flat eight-cap strips, as fluorescence is detected from above.

The Seegene Allplex Respiratory Panel assay was performed on the CFX-96™ Real Time PCR system (CFX, Bio-Rad, United States). The CFX setup prior to running the assay is divided into two main steps: Step A: protocol setup where the thermal profile was set (Table 2.1) and Step B: plate setup where the fluorophores were selected (FAM, HEX, Cal Red 610 and Quasar 670) (Table 2.2); followed by the selection of plate size (96 well) and plate type (BR White).

Table 2.1: Thermal profile setup on CFX-96™ Real Time PCR machine. (Allplex Respiratory panel 4 (Respiratory bacteria), Cat. No. RP9803Y)

Step	No. of cycles	Temperature	Duration
1	1	50 °C	20 min
2		95°C	15 min
3	45	95°C	10 secs
4*		60°C	1 min
5*		72°C	10 secs
6			Go to step 3, 44 more times

\* Fluorescence detection.

Table 2.2: Fluorophores used for the detection of analytes.

Analyte		
Fluorophore	Graph 1	Graph 2
FAM	<i>S. pneumoniae</i>	<i>L. pneumophila</i>
HEX	<i>H. influenzae</i>	<i>B. parapertussis</i>
Cal Red 610	<i>M. pneumoniae</i>	<i>B. pertussis</i>
Quasar 670	Internal Control (IC)*	<i>C. pneumoniae</i>

\*Detection of Internal Control in the Quasar 670 channel is not required for a positive result for target pathogens as a high titer of another analyte may result in reduced detection or the absence of the internal control signal.

The quantitation data obtained from the Seegene Allplex RP4 PCR run on the CFX was exported using the Seegene Export tool, where folders were automatically created (QuantStep4 and QuantStep5) to save the amplification curve data. The Seegene Viewer for Real Time PCR instruments (V2.0) was used to open the exported QuantStep4 data file, followed by the selection of test kit (Seegene Allplex Respiratory panel 4-8 strip) from the product menu, from which the results could be checked for each well that was selected. A Ct value of  $\leq 45$  indicates a positive result and a Ct value of  $>45$  or N/A (not detected) is considered a negative result.

#### 2.2.4 Pneumocystis jirovecii real time-PCR

A real-time PCR for the detection of *P. jirovecii* (PCP) has previously been developed (Huang *et al.*, 1999) and optimized in the Division of Medical Microbiology at Stellenbosch University (D. Banda, MSc 2016). This assay detects the Major Surface Glycoprotein (MSG) gene of *P. jirovecii* using the intercalating fluorescent DNA dye, SYBR Green I, and melt-curve analysis. Due to the large copy number,  $>100$  copies/genome, MSG is considered an appropriate target to establish a sensitive method for the detection of *P. jirovecii* in clinical specimens (Huang *et al.*, 1999).

MSG Heminested primers (Table 2.3) were used to amplify a 249 bp fragment of a highly conserved region of the *P. jirovecii* (*Pneumocystis carinii*) MSG gene (Huang *et al.*, 1999). The PCR master mix preparation was done as previously described (D. Banda 2016); 12.5  $\mu$ l of 2x Qiagen Rotor-Gene SYBR Green I master mix, 0.5  $\mu$ l (50  $\mu$ M) of each primer (1  $\mu$ M final concentration) and 9.5  $\mu$ l of RNase-free H<sub>2</sub>O per reaction. Two microliters of sample nucleic acid was added to each reaction. Each experimental run included the pCR2.1 plasmid containing HUMSG14 (GenBank accession nF AF033205) and a PCP positive clinical sample (determined by immunofluorescence as part of routine diagnostic procedures) as positive controls and a no template control.

Table 2.3: MSG Heminested primer sequences and product sizes. (Huang *et al.*, 1999)

Primer name	Sequence 5'-3'	Product size (bp)
JKK14/15 (upstream)	GAATGCAAATCYTTACAGAGACAACAG	249
JKK17 (downstream)	AAATCATGAACGAAATAACCATTGC	

Real-time PCR amplification was carried out using the Rotor-Gene Q analyzer (Qiagen) that was manually set to the following cycling program; an initial enzyme activation of 95°C for 5 minutes, followed by 40 cycles of denaturation for 5 seconds and annealing/extension/data acquisition at 60°C for 10 seconds. The amplified products were analyzed using melt curve analysis software that is programmed on the Rotor-Gene software (version 2.02.4). The Rotor-Gene software calculated the derivative of intensity of fluorescence (acquired to Melt A on green) where the derivative peak (dF/dT) represented the melting temperature value ( $T_m$ ) of the MSG gene fragment. For the melt curve analysis the following criteria was used; the melting temperature ramp was set from 70-90°C rising by 0.5°C each step. The  $T_m$  value of the PCP positive plasmid control (84.5°C) was used to define the derivative peak bin which was set to 4°C (2°C on either side of the  $T_m$ ). All samples with peaks which conformed to the following defined thresholds were considered positive: (1) dF/dT threshold of >1 and (2) derivative peak within the defined derivative peak bin.

### 2.2.5 Clinical data collection

Case report forms (CRF) were completed for each participant enrolled in the study. The information collected included social demographics, presenting symptoms, history of TB, HIV and TB exposure, HIV status (HIV exposed uninfected, HIV infected, HIV unexposed), concurrent illnesses, current and recent antibiotic treatment, clinical examination and the TB result obtained after intensive investigation. In addition, information on well known risk factors for respiratory illnesses were included; age, gender, breastfeeding (including duration), prenatal or postnatal exposure to smoke, mode of delivery (caesarean section (CS) or natural vaginal delivery (NVD)), day care attendance, household size and whether the child had siblings or not.

### 2.2.6 Statistical Analysis

Statistical analysis was done using IBM SPSS Statistics 25 software. Chi square and Fischer exact tests (where applicable) were done to determine the statistical significance between the detection of respiratory pathogens in relation to the TB and unlikely TB groups and any other risk factors or two-way associations. Results were considered statistically significant if the p value was 0.05 or less. The statistical analysis was done in consultation with a biostatistics consultant from the Division of Epidemiology and Biostatistics, Department of Global Health, Stellenbosch University through support from the Faculty of Medicine and Health Science's Dean's fund in collaboration with the Biostatistics Unit, Stellenbosch University.

## 2.3 Results

### Study population

From the 75 participants enrolled in the parent study between December 2015 to January 2017 (Chapter 1, Section 1.10), 72 baseline respiratory samples were received for this sub-study, these included NPA (n= 68) and IS samples (n=4). Three participants refused sample collection or withdrew from the study. Of the 72 participants from whom samples were obtained, 42 were defined as TB cases (58.3%), based on either bacteriological confirmation (19/42, 45.2%) or clinical diagnosis (23/42, 54.8%). The remaining 30 (41.7%) participants were defined as well-defined ill controls (unlikely TB). There were no statistically significant differences between the demographics of the different groups. The demographic details are described in Table 2.4. Of the participants, 18.6% (13/70) were HIV-infected of whom 15.3% (2/13) had bacteriologically confirmed PTB, 53.8% (7/13) were clinically diagnosed with PTB and 30.8% (4/13) did not have PTB. Two of the 72 baseline samples were excluded due to insufficient sample for DNA extraction. Both samples were from the unlikely TB control group.

The median age of participants was 21.5 months (interquartile range (IQR), 9-45 months). Fifty-six percent (39/70) of the participants were under the age of two, 30% (21/70) were 2-5 years of age and 14% (10/70) were over the age of 5 years. 55.7% (39/70) of participants included in this study were male of which 53.8% (21/39) were cases (Table 2.4). Table 2.4 describes participant information and the risk factors that were evaluated.

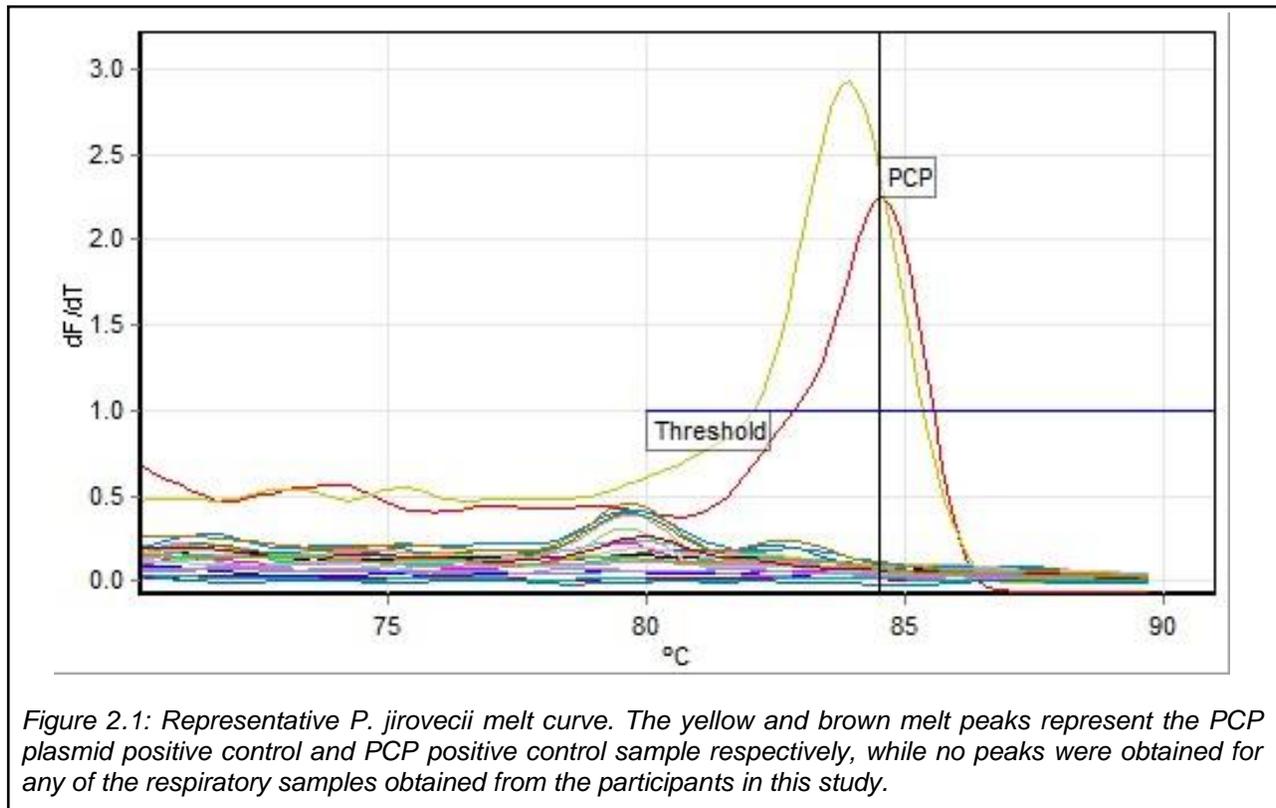
Table 2.4: Participant information and risk factors.

Category	Total study population (n, %)	Bacteriologically confirmed PTB (n, %)	Clinically Diagnosed PTB (n, %)	Unlikely TB (n, %)
<b>Number of participants</b>	70 (100)	19 (27.1)	23 (32.9)	28 (40.0)
<b>Age in months<sup>1</sup> (median, IQR)</b>	21.5 (36)	21.5 (39)	24 (38.5)	21.5(39)
<b>Sex<sup>2</sup></b>				
Male	39 (55.7)	9 (47.4)	12 (52.2)	18 (64.3)
Female	31 (44.3)	10 (52.6)	11 (47.8)	10 (35.7)
<b>TB treatment<sup>2</sup></b>				
Yes	42 (60)	19 (100)	23 (100)	N/A
Multi Drug Resistant	4 (9.5)	3 (15.8)	1 (4.3)	N/A
<b>HIV status<sup>2</sup></b>				
Yes	13 (18.6)	2 (10.5)	7 (30.4)	4 (14.3)
No	54 (77.1)	15(78.9)	16 (69.6)	23 (82.1)
Unknown	3 (4.3)	2 (10.5)	0 (0)	1 (3.6)
<b>HIV Exposure<sup>2</sup></b>				
Yes	15 (21.4)	4 (21.1)	6 (26.1)	5 (17.9)
No	53 (75.7)	15 (78.9)	15 (65.2)	23 (82.1)
Unknown	2 (2.9)	0 (0)	2 (8.7)	0 (0)
<b>HIV Infected<sup>2</sup></b>				
Yes	13 (18.6)	2 (10.5)	7 (30.4)	4 (14.3)
No	57 (81.4)	17 (89.5)	16 (69.6)	24 (85.7)
<b>Smoking exposure (prenatal)<sup>2</sup></b>				
Yes	26 (37.1)	7 (36.8)	10 (43.5)	9 (32.1)
No	26 (37.1)	6 (31.6)	10 (43.5)	10 (35.7)
Unknown	18 (25.7)	6 (31.6)	3 (13)	9 (32.1)
<b>Mode of Delivery<sup>2</sup></b>				
Normal Vaginal Delivery	54 (77.1)	15 (78.9)	17 (73.9)	22 (78.6)
Caesarean Section	14 (20)	4 (21.1)	4 (17.4)	6 (21.4)
Unknown	2 (2.9)	0 (0)	2 (8.7)	0 (0)
<b>Breast fed<sup>2</sup></b>				
Yes	58 (82.9)	16 (84.2)	21 (91.3)	21 (75)
No	11 (15.7)	3 (15.8)	2 (8.7)	6 (21.4)
Unknown	1 (1.4)	0 (0)	0 (0)	1 (3.6)
<b>Length of breastfeeding<sup>2</sup></b>				
0-4 months	28 (40)	8 (42.1)	10 (43.5)	10 (35.7)
5-10 months	18 (25.7)	5 (26.3)	4 (17.4)	9 (32.2)
>10 months	8 (11.4)	2 (10.5)	3 (13)	3 (10.7)
Unknown	16 (22.9)	4 (21.1)	6 (26.1)	6 (21.4)
<b>Day Care Attendance<sup>2</sup></b>				
Yes	25 (35.7)	4 (21.1)	10 (43.5)	11 (39.3)
<b>Siblings<sup>2</sup></b>				
Yes	55 (78.6)	14 (73.7)	17 (73.9)	24 (85.7)

<sup>1</sup> Continuous data is expressed as a median with interquartile range (IQR) in between brackets. <sup>2</sup> Proportions are expressed as a number with percentage in between brackets

### 2.3.1 Pathogen detection

The Seegene Allplex RP4 kit was used to detect the presence of seven clinically significant bacterial pathogens in the 70 respiratory samples. Valid results were obtained for all of the samples based on the detection of the RP-B internal control with or without detection of other bacterial pathogens. The in-house real-time PCR targeting the Major Surface Glycoprotein (MSG) gene was used to detect *P. jirovecii* in the 70 respiratory samples. Based on the thresholds defined in section 2.2.4, the MSG gene was only detected in the positive control samples (Figure 2.2).



The most frequently detected pathogens were *H. influenzae* and *S. pneumoniae* which were detected in 74.2% (52/70) and 60% (42/70) of the samples, respectively. The bacterial pathogens, *C. pneumoniae* (8/70; 11.4%), *M. pneumoniae* (4/70; 5.7%) and *B. parapertussis* (1/70; 1.4%) were less common, while *L. pneumophila*, *B. pertussis* and *P. jirovecii* (fungal pathogen) were not detected in any of the samples (Figure 2.3).

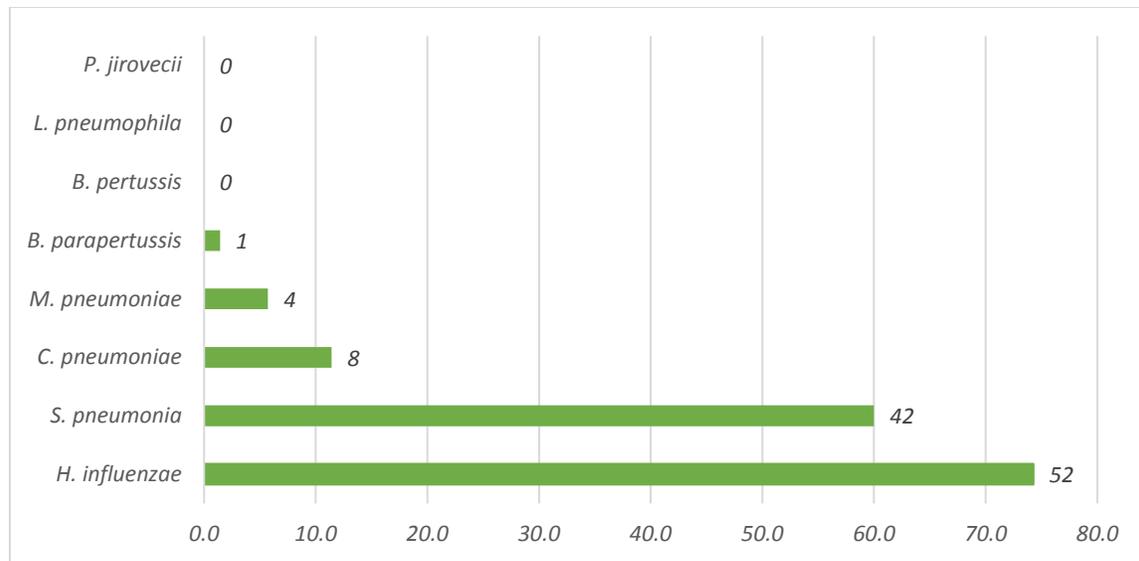


Figure 2.2: The percentage of samples ( $n=70$ ) in which pathogens were detected

Similar trends in the frequencies of the various pathogens were seen in the TB cases (bacteriologically confirmed and clinically diagnosed) and the unlikely PTB group (Figure 2.4). There was no significant difference between the presence of *S. pneumoniae* or *H. influenzae* in the TB cases and unlikely TB groups ( $p=0.921$  and  $0.655$  respectively). *H. influenzae* was detected 74.3% (52/70) of samples; 76.2% (32/42) in the TB group and 71.4% (20/28) in the unlikely TB group ( $p=0.655$ ). *S. pneumoniae* was detected in 60% (42/70) of samples; 59.5% (25/42) in the TB group and 60.7% (17/28) in the unlikely TB control group ( $p=0.921$ ). *C. pneumoniae* was most frequently detected in the unlikely TB group, while *M. pneumoniae* was more common amongst the TB cases, however the low numbers of positive samples precluded statistical analysis. *B. parapertussis* was detected in a single sample from a bacteriologically confirmed PTB case (Figure 2.4).

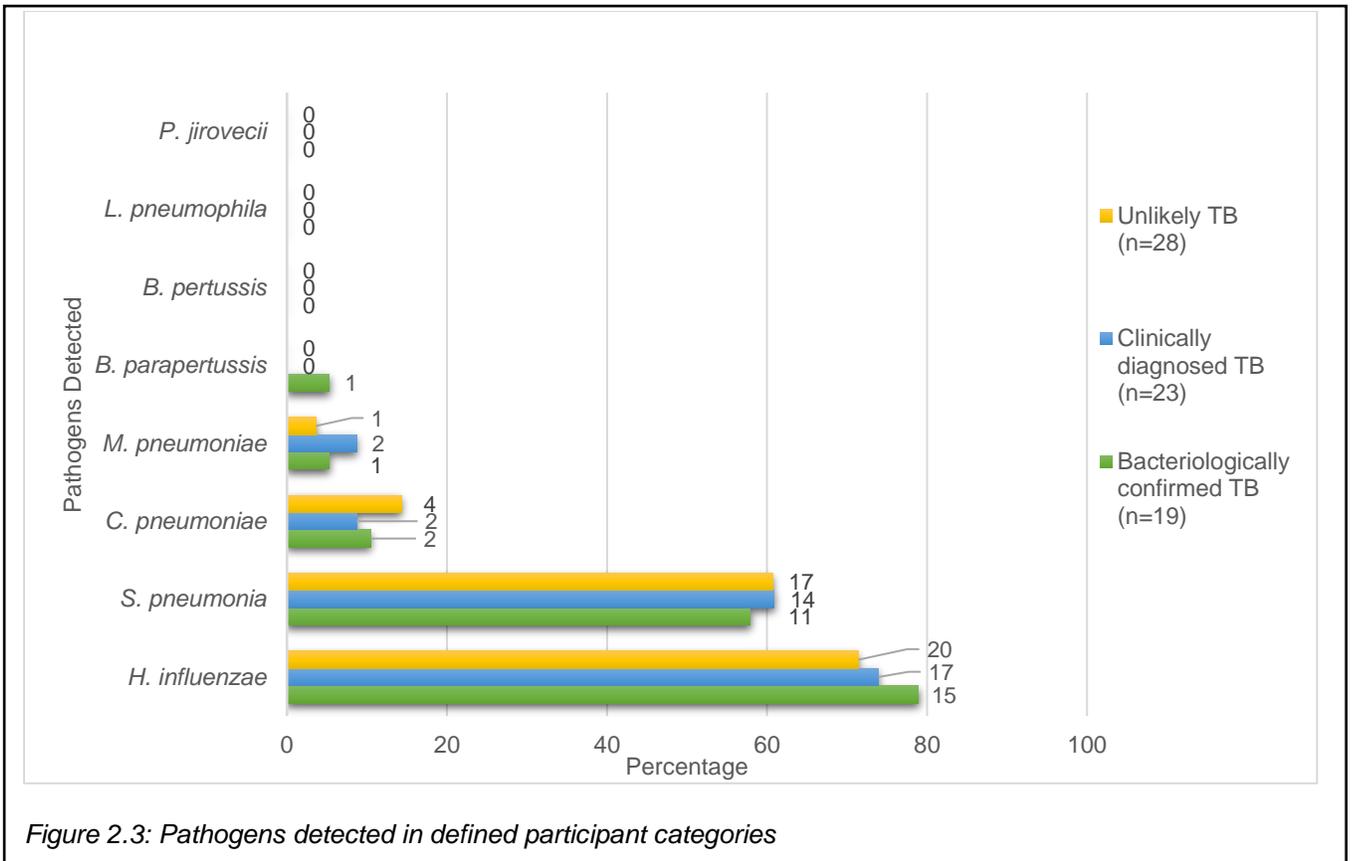


Figure 2.3: Pathogens detected in defined participant categories

In 88.6% (62/70) of samples at least one respiratory pathogen was detected, with two or three bacterial pathogens detected in the majority of samples (57.1%, n=40/70) (Table 2.5). There was no significant difference between the number of pathogens detected in the TB and unlikely TB groups (Table 2.5).

Table 2.5: The number of bacteria detected in the TB and unlikely TB groups.

		TB group (n,%)	Unlikely TB Group (n,%)	Total (n)	P value (Fischer Exact test)
Number of Bacteria	None	4 (9.5)	4 (14.3)	8	0.553
	One	15 (35.7)	7 (25)	22	
	Two	19 (45.2)	16 (57.1)	35	
	Three	4 (9.5)	1 (3.6)	5	
	<b>Total</b>	<b>42</b>	<b>28</b>	<b>70</b>	

*H. influenzae* was detected alone in 15/70 samples (21.4%), while *S. pneumoniae* (6/70, 8.6%) and *C. pneumoniae* (1/70, 1.4%) were less commonly detected alone. *H. influenzae* and *S. pneumoniae* was the most frequently detected bacterial pathogen pair, present in 29/70 (41.4%) samples, while *H. influenzae* and *C. pneumoniae* (2/70, 2.9%), *S. pneumoniae* and *C. pneumoniae* (2/70, 2.9%) and *H. influenzae* and *M. pneumoniae* (1/70, 1.4%) were less commonly detected. In samples where three bacteria were detected, both *H. influenzae* and *S. pneumoniae* were identified with either *C. pneumoniae* (2/70, 2.9%), *M. pneumoniae* (2/70, 2.9%) or *B. parapertussis* (1/70, 1.4%). There were no differences between confirmed and clinically diagnosed cases.

### 2.3.2 Evaluation of risk factors for the presence of respiratory pathogens

The presence of bacterial pathogens in relation to multiple risk factors was evaluated using univariable analysis. Due to the small sample size, no multivariable analysis was performed. Normal vaginal delivery (NVD) was associated increased risk for the presence of bacterial pathogens, in comparison to caesarean section delivery ( $p=0.006$ ), while any breastfeeding was associated with a reduced risk for the presence of bacterial pathogens ( $p=0.01$ ). However, no statistically significant association was observed between length of breastfeeding and the presence of bacterial pathogens ( $p=0.29$ ).

A trend toward significance was observed for both smoking exposure and day care attendance as risk factors for the presence of bacterial pathogens ( $p=0.17$  and  $0.145$ , respectively) (Table 2.6). Gender, age, having siblings and HIV infection were not associated with the presence of bacterial pathogens (Table 2.6).

Table 2.6: Risk factors for the presence of bacterial pathogens.

Risk factor	Total population (n = 70)	No bacterial pathogen detected	One or more bacterial pathogen detected	p-value
<b>Mode of delivery<sup>1</sup></b>				
C/S	14	5 (35.7)	9 (64.3)	0.006*
NVD	54	3 (5.6)	51 (94.4)	
Unknown	2	0 (0)	2 (100)	
<b>Breastfeeding (any length of time)<sup>1</sup></b>				
No	11	0 (0)	11 (100)	0.01*
Yes	58	7 (12.1)	51 (87.9)	
unknown	1	1 (100)	0 (0)	
<b>Length of breastfeeding<sup>1</sup></b>				
0-4 months	28	4 (14.3)	24 (85.7)	0.29
5-10 months	18	2 (11.1)	16 (88.9)	
>10 months	8	2 (25)	6 (75)	
Unknown	16	0 (0)	16 (100)	
<b>Prenatal smoke exposure<sup>1</sup></b>				
No	26	3 (11.5)	23 (88.5)	0.17
Yes	26	1 (3.8)	25 (96.2)	
Unknown	18	4 (22.2)	14 (77.8)	
<b>Attended day care<sup>1</sup></b>				
Yes	25	1 (4)	24 (96.0)	0.145
<b>Siblings<sup>1</sup></b>				
Yes	55	5 (9.1)	50 (90.9)	<b>0.239</b>

Table 2.6: Continued: Risk factors for the presence of bacterial pathogens

<b>Age<sup>1</sup></b>				
<2 years	39	4 (10.3)	35 (89.7)	0.65
2-5 years	21	2 (9.5)	19 (90.5)	
> 5 years	10	2 (20)	8 (80)	
<b>Gender (male)<sup>1</sup></b>				
Male	39	5 (12.8)	34 (87.2)	0.681
Female	31	3 (9.7)	28 (90.3)	
<b>HIV infected<sup>1</sup></b>				
Yes	13	2 (15.3)	11 (84.6)	0.461
No	57	6 (8.8)	51 (89.5)	

Risk factor percentages calculated as row percentages in brackets. \* Statistically significant result (Chi-square).

## 2.4 Discussion

### Detection of pathobionts and pathogens

Of the seven bacterial species that can be detected using the Seegene assay only *H. influenzae*, *S. pneumoniae*, *M. pneumoniae*, *C. pneumoniae* and *B. parapertussis* were detected, and the fungal pathogen *P. jirovecii* was not detected in any of the respiratory samples. There was a high burden of respiratory pathogens in children suspected of PTB, of which *H. influenzae* (74.4%) and *S. pneumoniae* (60%) were the most commonly detected. The presence of *H. influenzae* and *S. pneumoniae* is not surprising as they are among the frequent colonizers of the nasopharynx and are often identified in respiratory samples. Moreover, studies have found that acquisition and colonization by *H. influenzae* and *S. pneumoniae* (among others) can occur as early as a few days to a few months after birth (García-Rodríguez and Fresnadillo Martínez, 2002). Another study conducted in Cape Town also found that *S. pneumoniae* and *H. influenzae* were commonly detected in children with suspected PTB (Dube *et al.*, 2016), however *H. influenzae* was more frequently detected in our study (74.4% vs 29% in their study), and *S. pneumoniae* was more frequently detected in their study (42%). This may be a result of the age group ( $\pm 36$  months versus  $\pm 22$  months in our study), the type of DNA extraction that was conducted, the type of kit used for the detection of nucleic acid targets (FTD resp33 Kit, Fast-track Diagnostics, Luxembourg vs Seegene Allplex Respiratory Panel 4 in our study) or the specific sample type that was used (NP swabs versus NPAs in our study).

The other respiratory pathogens, *M. pneumoniae*, *C. pneumoniae* and *B. parapertussis*, were less frequently detected in the respiratory samples. Generally, the detection of these pathogens in respiratory samples is associated with cases of infection and disease, rather than colonization. Both *M. pneumoniae* and *C. pneumoniae* are atypical pathogens that can cause disease ranging from mild to severe acute respiratory infections. They are also pathogens that are more frequently seen in older children (Principi *et al.*, 2001). The clinical manifestation of disease ranges from tracheolobronchitis to atypical pneumonia followed by extrapulmonary complication and from pharyngitis, bronchitis and sinusitis to community acquired pneumonia (CAP), respectively (Del Valle-Mendoza *et al.*, 2017). Both *M. pneumoniae* and *C. pneumoniae* are important causes of atypical CAP, but *C. pneumoniae* has been identified as a more frequent cause, that is often detected in children and the elderly (Del Valle-Mendoza *et al.*, 2017).

*B. parapertussis* was only detected in 1 sample. Its involvement in respiratory disease has been linked to the less severe form of whooping cough. The incidence of infections caused by *B. parapertussis* is not well known. Liko *et al* (2017) established that this is due to laboratories not distinguishing between *Bordetella* species and *B. parapertussis* either not being reported or being misreported as *B. pertussis*, while another study suggested that it is neglected because of its lower incidence and milder symptoms and its similarities with *B. pertussis* (Elahi *et al.*, 2008). One study suggested that it may even be as common as *B. pertussis*, especially in children who have yet to attend school (He, 1998). Despite this, studies have reported that this pathogen can cause up to 20-30% of cases of whooping cough (Elahi *et al.*, 2008) which indicates that it might become a serious health problem.

*B. pertussis* and *L. pneumophila* were not detected in any of the samples. Numerous studies have successfully used nasopharyngeal samples for the detection of *B. pertussis* by PCR. Protection against *B. pertussis* infection is possible through vaccination which may explain why the participants in this study were negative. Although the samples used in this study were negative for *B. pertussis*, this pathogen is still not under control in any country and represents the most prevalent vaccine-preventable childhood disease (Locht, 2016). This was evident in the study conducted by Dube *et al* (2016) where *B. pertussis* was found to be fairly common (12/214, 6%) in children, which could imply that the detection of this pathogen in our study could also be limited by the small sample size (sample size of 70 vs 214). The type of clinical specimen also influences the detection of a particular pathogen. A study that investigated the diagnosis and incidence of pertussis in children found that induced sputum samples had a higher sensitivity for the diagnosis of *Bordetella* spp, where they identified more confirmed cases on IS samples than on NP specimens (Rudzani Muloiwa; Felix S. Dube; Mark P. Nicol; Heather J. Zar; Gregory D. Hussey, 2016).

*L. pneumophila* causes CAP but is one of the less frequently involved pathogens and contributes to approximately 8% of CAP in South Africa. Individuals in this study may be negative because those at higher risk of infection are persons over the age of 50. Additionally, unlike the other respiratory pathogens in this study, human to human transmission of *L. pneumophila* has not been reported and one usually becomes infected as a result of exposure to contaminated water droplets (NICD, 2016). The presence or absence of this bacterium in respiratory samples would therefore provide insight into the efficacy of water management systems that are currently in place in certain communities. Perhaps in other regions where water system management is poor the probability of identifying this pathogen would be higher.

*P. jirovecii* was included in this study as it is a clinically significant pathogen. Pneumocystis pneumonia is mostly identified in HIV infected individuals, and HIV is a major driver of TB in South Africa, thus the reason for the inclusion. Additionally, it has been reported that HIV uninfected infants with underlying predisposing factors such as HIV exposure and malnutrition are also susceptible to pneumocystis pneumonia (McNally *et al.*, 2007; Morrow *et al.*, 2010). Lower respiratory tract samples such as bronchoalveolar lavage or induced sputum are predominantly used for the detection of *P. jirovecii* during standard laboratory diagnosis (Samuel *et al.*, 2011). However, in children, obtaining these specimens can be challenging as invasive collection is not well tolerated. Therefore, upper respiratory tract specimens (nasopharyngeal aspirates) which are easier to obtain have been used in combination with PCR for the detection of *P. jirovecii* in children. PCR has been shown to be more sensitive than the standard diagnostic measures in both upper and lower respiratory samples (Samuel *et al.*, 2011). This allows more desirable clinical specimen in children and improves on the overall turnaround time for diagnosis .

Of the participants in this study, 18.6% were HIV-infected, however none of the participants in this study were positive for this fungal pathogen. Studies have found that PCR sensitivity is dependent on the type and nature of the specimen, the degree of immunosuppression in infected individuals and the gene targets used (mtLSU rRNA, ITS locus, MSG) (Olsson *et al.*, 1993; Gupta *et al.*, 2009). However, the use of the MSG gene has previously been shown to be useful for the detection of *P. jirovecii* in upper respiratory tract samples (NPA) from children <14 years, in Cape Town (Samuel *et al.*, 2011). The main differences between their study and our study is the use of a different DNA extraction kit (Nuclisense Easy Mag) and the use of fluorescence resonance energy transfer probes in a qualitative touchdown PCR. In agreement with the present study, Gupta *et al* (2009) detected no *P. jirovecii* in NPA samples when using MSG, mtLSU or ITS as gene targets. Although different factors can contribute to the detection of *P. jirovecii* in samples, the use of antimicrobial agents such as co-trimoxazole could also be a factor to consider. This antimicrobial agent has contributed to the decline of Pneumocystis pneumonia since its use as a prophylaxis against Pneumocystis pneumonia in HIV exposed infants (Zar, 2010). However, the use of co-trimoxazole in this study population was not determined and therefore the statement may or may not be a valid reason for the decline in the detection of *P. jirovecii* in the participants, especially in the HIV exposed infants.

The detection of no pathogens in a few of the samples (n=8,11.4 %), may be that the samples did not have the specific pathogens present or that these samples contained other respiratory pathogens, including viruses and other fungi which are not detected by the tests used in this study. Microorganisms within a specific niche such as the nasopharynx can include numerous other commensals or pathobionts (potential pathogens) that all interact with one another. Bosch *et al* (2013) suggested that microbes have developed various interaction tools that lead to positive and negative interactions, where positive associations generate favorable conditions for microorganisms via mutualism, commensalism, symbiosis or the assistance with immune evasion. On the other hand, negative associations may be a result of interspecies interactions, where there is direct competition between organisms for a specific niche, or when the hosts immune response disproportionately affects a competing organism. In this study negative associations between microorganisms could have resulted in the loss of certain organisms in the participants (Bosch *et al.*, 2013). Additionally, Rodriguez *et al* (2002) reported that viridans streptococci can antagonize colonization by other streptococci, such as  $\alpha$ -haemolytic streptococci, which can inhibit the growth of *S. pneumoniae*, *H. influenzae* and *M. catarrhalis*. Of those samples that were negative for all pathogens, one was an ill-control, which could imply that other respiratory organisms may have influenced the clinical presentation and brought about the suspicion of disease. More so, antibiotics majorly influence the presence of commensals and pathogens. Participants in this group were on antibiotics which may have contributed to the loss of these pathogens, although the use of antibiotics was not taken into consideration when the analysis was done.

#### The detection of pathogens in relation to TB

No associations could be made between the absence or presence of one or more respiratory pathogens and TB disease. This may be due to the fact that all participants in this study were ill and may have been exposed to antibiotic treatment, or that a healthy control group was not included.

Similar trends in the detection of various pathogens were seen in TB cases (bacteriologically confirmed and clinically diagnosed) and unlikely TB groups with no significant difference between the most prevalent organisms (*S. pneumoniae* and *H. influenzae*) and the defined categories. The detection of *C. pneumoniae* was equally distributed in both categories (unlikely-TB control group and TB cases), while *M. pneumoniae* appeared to be more common among the TB cases. However, the small sample size precluded statistical analysis and therefore warrants the need for further investigation. A larger sample size of participants provides a better estimate of the true

population and more robust statistical analysis. The information obtained from a larger study would be more informative as it would provide a more accurate representation of the community and the defined categories.

The detected pathogens such as *C. pneumoniae* and *M. pneumoniae* in the unlikely TB group may have contributed to respiratory disease in these children. However, whether these pathogens caused disease in these patients should be further evaluated as the detection of pathogens does not necessarily imply infection and could represent asymptomatic colonization. A method that could assist in differentiating between infection vs asymptomatic colonization is quantitative PCR as it can be used simultaneously to detect and quantify the pathogen and pathogen load, respectively. Studies have shown that quantitative PCR can assist in diagnosis and determining disease severity during viral or bacterial respiratory infections from nasopharyngeal or sputum samples (Borg *et al.*, 2003; Gadsby *et al.*, 2015). A challenge with this may be determining appropriate cutoff values to differentiate between infection vs asymptomatic colonization. Furthermore, it should be recognized that the nasopharynx is a complex niche that is constantly exposed to different pathogens, including viruses which are common pathogens in respiratory diseases in children and are therefore vital to explore. Bosch *et al* (2013) reviewed numerous studies and summarized how the presence of viruses can influence bacterial colonization, (1) through predisposing bacterial adherence as a result of an altered mucosal surface, (2) disrupting the first line of defense of the respiratory tract; the epithelium barrier, (3) the upregulation of adhesion proteins which alters the hosts defense mechanism, (4) the production of viral factors such as neuraminidase which allows bacteria to enter host cells and (5) impairing the hosts immune system components (Bosch *et al.*, 2013). It would be interesting to determine whether viruses or viral-bacterial co-infections contributed to the clinical suspicion of respiratory disease and what type of viral-bacterial interactions exist (if any) in these children.

The number of bacteria detected in the TB and ill control group were found to not be statistically different. The detection of most of the atypical pathogens (*C. pneumoniae*, *M. pneumoniae* *B. parapertussis*) in the TB group could suggest that the atypical pathogens identified may have contributed to the clinical presentation or symptoms and possibly allowed for the early detection of TB. However, the association between the presence of atypical pathogens and the clinical presentation and/ or symptoms was not determined because only a few positive samples were identified and would not be sufficient for analysis.

Not many atypical pathogens were identified in the ill control group which may indicate that the cause of the initial suspicion of TB may have been influenced by other respiratory pathogens;

such as viruses. Considering other respiratory pathogens may be more informative in determining if respiratory pathogens influenced the suspicion of TB in the ill control group, especially since lower respiratory tract infections in children are believed to be polymicrobial in nature. At this point investigation of the relationship between bacterial pathogens and clinical presentation or suspicion of TB was limited by sample size, the lack of healthy controls and the limited bacterial pathogens that could be detected with the Seegene Allplex RP4 assay. The detection of viruses in these samples is currently being conducted and has yet to be compiled with the bacteriological findings in this study, which may provide insight into the interactions between the bacteria and viruses in the TB case group in addition to providing insight into what contributed to the initial suspicion of PTB in the control group.

Mhimbira *et al* (2018) conducted a study that investigated the clinical significance of bacteria and viruses in adult TB participants and household contacts, and found that TB patients had a lower prevalence of bacterial pathogens than the controls. Our findings were contradictory in that the TB group had more bacteria present than the ill control group (although not significant). This may be because our study focused on children, and respiratory pathogens in the nasopharynx differ between adults and children as it is in a constant state of flux as bacteria are acquired and eliminated or reacquired throughout life. Additionally, Mhimbira *et al* (2018) found that TB patients were more likely to have severe TB when they were co-infected with both viruses and bacteria, which raises the importance of including viruses in this study and also the significance of polymicrobial disease.

There are limited studies which have investigated the role of respiratory pathogens in children with PTB. The study from Dube and colleagues found no clear association between TB categorisation and the detection of specific pathogens but did see a dominant microbial profile in children with PTB that consisted of *C. pneumoniae*, hMPV, coronavirus O43, influenza C virus, rhinovirus and cytomegalovirus, which failed to reach statistical significance. Whereas in those without TB, *P. jirovecii*, *H. influenzae spp*, RSV, *M. pneumoniae*, influenza B virus and enteroviruses were more consistently detected (Dube *et al.*, 2016).

### Evaluation of risk factors

Various risk factors for the presence of bacterial pathogens were investigated. The presence of bacterial pathogens was negatively associated with being breastfed, although length of breastfeeding did not affect this correlation. Normal vaginal birth was associated with increased risk for the presence of bacterial pathogens. However, it should be mentioned that these findings violate the assumption of the chi-square test that not more than 25% of the cells have a count of

less than 5 and therefore these results should be interpreted with caution, and may be attributed to the small sample size.

Studies have hypothesized that breastfeeding offers a protective effect for respiratory tract infections. Biesbroek *et al* (2014) was one of the first to describe a correlation between infant feeding and microbial composition of the upper respiratory tract, where it was suggested that lactic acid bacteria may contribute to the protective effect of breastfeeding against respiratory infections. However, other studies have found that breastfeeding doesn't substantially influence nasopharyngeal colonization with respiratory pathogens (Kaleida *et al.*, 1991; Bakhshae *et al.*, 2015). The duration of breastfeeding has been shown to reduce the risk of respiratory tract infections when breastfed for 6 months and longer (Tromp *et al.*, 2017; Wang *et al.*, 2017), however this was not observed in our study, possibly due to the small sample size.

Mode of delivery can influence nasopharyngeal microorganisms by favouring the development of potentially protective or negative microorganisms (Esposito and Principi, 2018). This is important as early microbial composition has been shown to influence respiratory health in children (Biesbroek, Tsivtsivadze, Elisabeth A M Sanders, *et al.*, 2014). Microbiome studies found that children delivered by natural vaginal birth acquire microorganisms resembling their mothers vaginal microbiota, where those born via C-section harbor bacterial communities similar to those found on the skin surface (Dominguez-Bello *et al.*, 2010).

Numerous factors can influence nasopharyngeal carriage rates in children, ranging from age to sleeping position (García-Rodríguez and Fresnadillo Martínez, 2002). In this study no significant associations were detected for the remaining risk factors, although associations between prenatal smoke exposure and day care, and presence of bacterial pathogens showed a trend towards significance. A study conducted in Russian children found that both prenatal smoke exposure and postnatal environmental smoke exposure have adverse effects on respiratory health which increases the risk of respiratory illnesses such as asthma, bronchiolitis and respiratory infections (Jaakkola *et al.*, 2006). The increased risk for respiratory illness may be due to an increase in susceptibility to viral and bacterial colonization as a result of damaged and inflamed nasopharyngeal mucosa (García-Rodríguez and Fresnadillo Martínez, 2002). Day care centers provide the perfect environment for pathogen exposure and transmission due to the frequent close person to person contact. This is similar in children with siblings who are also more likely to be any bacteria positive in comparison to those who did not have siblings (García-Rodríguez and Fresnadillo Martínez, 2002). HIV infection did not seem to influence the presence of pathogens in this study, most (89.5%) of the participants were HIV negative.

## 2.5 Conclusion

Pulmonary TB is a well-studied disease; however, little is known about respiratory co-infections and their role in TB disease. Due to the fact that respiratory pathogens play a huge role in overall respiratory health, only a few studies have looked at respiratory pathogens in relation to TB disease in adults (Lockman *et al.*, 2003; Mhimbira *et al.*, 2018), but less so in children (Dube *et al.*, 2016). This chapter investigated associations between the presence of respiratory pathogens (and pathobionts) in children with suspected PTB, by detecting a variety of respiratory pathogens in children categorized as TB cases and children with unlikely TB.

This pilot study found no association between the presence of respiratory pathogens (and pathobionts) in the nasopharynx and the defined TB categories (bacteriologically confirmed PTB, clinically diagnosed PTB and unlikely PTB cases). This may be due to the limited sample size, the limited respiratory pathogens that could be detected with the techniques used, and the fact that children with other respiratory illnesses were used as control group or the involvement of other respiratory pathogens such as viruses. Further studies describing the nasopharyngeal microbiota in the same population are described in Chapter 4 to determine whether microbial diversity may be associated with PTB in children.

## CHAPTER 3:

# The description of the respiratory microbiome using 16S rRNA gene sequencing

### 3.1 Introduction

In 2007, the Human Microbiome Project (HMP) initiated multiple projects that aimed to fulfill various goals, namely (1) to use improved sequencing technology to characterize the microbiome of multiple body sites (the gastrointestinal tract, vagina, mouth, skin and nasal cavity), (2) to determine whether associations exist between the microbiome and health or disease in different medical conditions, and (3) to provide both a standardized data resource and new technological approaches to allow microbiome studies to be broadly undertaken in the scientific community (The NIH HMP Working Group, 2009). Since then, numerous microbiome studies have been done on multiple body sites describing marked differences between the microbiome in healthy and diseased states, using different technological approaches. More so, the microbiome has since been dubbed the human health biomarker as it aids in maintaining normal host physiology, developing and educating the immune system, metabolizing complex substrates and providing crucial protection against opportunistic pathogens (Shukla *et al.*, 2017).

Traditionally, microscopy and culture-based techniques were used to study and differentiate microorganisms, but it was soon realized that these techniques provided a limited scope of the vast microbial world. The introduction of molecular approaches in diagnostic laboratories improved on the turnaround time, specificity and sensitivity for the identification and characterization of medically relevant microorganisms that were usually not identifiable with traditional culture techniques (i.e. phenotypic identification). This included fastidious, non-viable, slow growing and non-cultivable organisms (Shilts *et al.*, 2016). Conventional molecular techniques such as denaturing gradient gel electrophoresis, terminal restriction fragment length polymorphism and fluorescent in situ hybridization are nucleic acid based techniques which were commonly used to study bacterial communities and diversity before the introduction of high throughput sequencing technologies (Case *et al.*, 2007). High throughput sequencing such as next-generation sequencing (NGS) has revolutionized the study of microbial communities as a result of its ability to generate a comprehensive catalogue of microbial sequences present in different ecological niches within a large host organism such as humans (Martín *et al.*, 2014). With the advent of and improvement in sequencing technologies, characterizing microbial diversity in samples has become more efficient and cost effective. This has become apparent in

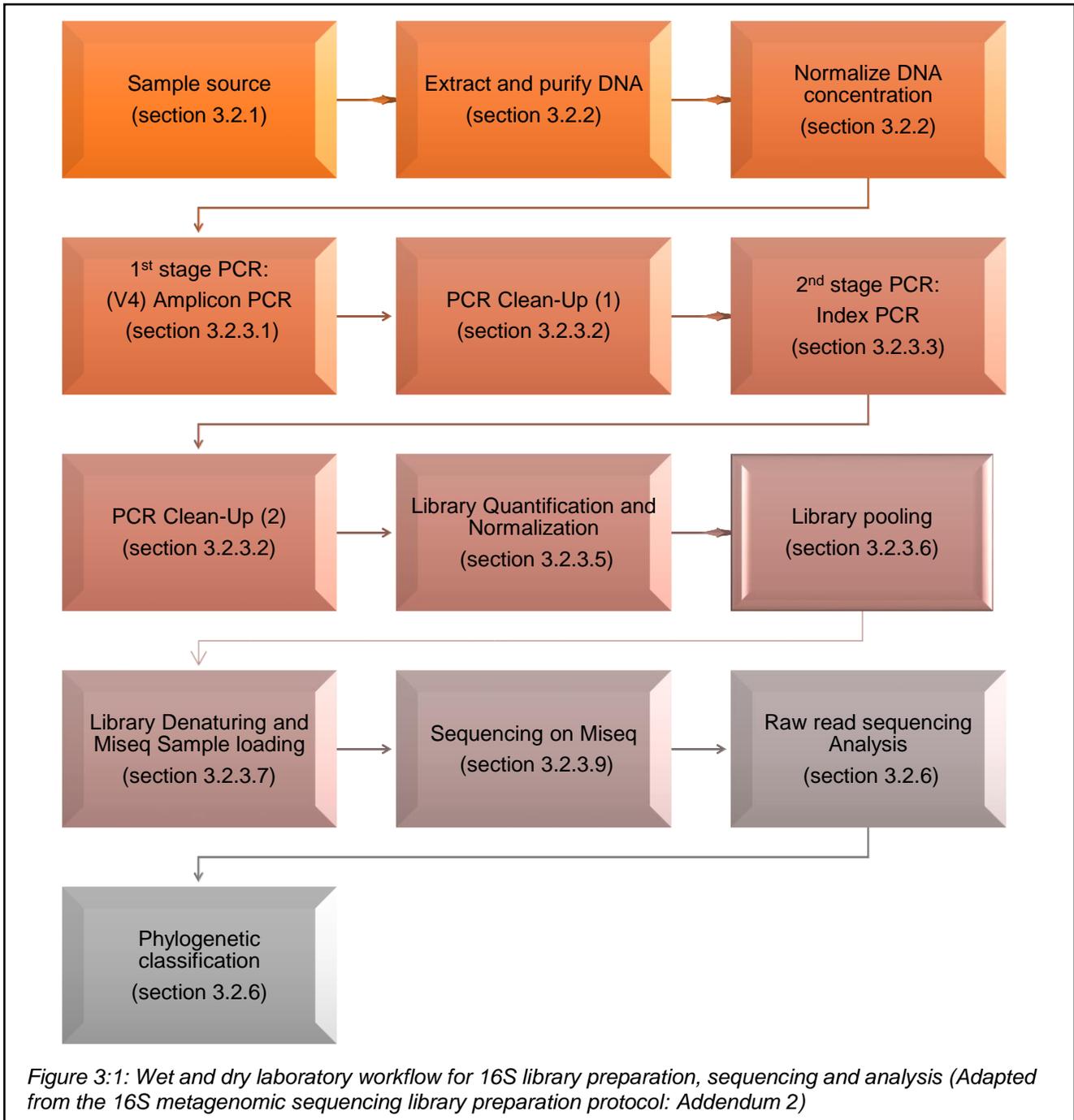
Illumina sequencing technologies, which with its reduced costs, comparatively high sequencing depth (Klindworth *et al.*, 2013) and increase in sequencing length, has become the most widely used sequencing platform (Ravi, Walton and Khosroheidari, 2018).

The 16S ribosomal ribonucleic acid (rRNA) gene has become the most commonly used genetic marker to study bacterial phylogeny and taxonomy. This may be attributed to its presence in all bacteria, the fact that 16S rRNA functionality has not changed over the years suggesting that random sequence changes are a more accurate measure of time, and its size: the 16S rRNA gene is large enough (1500 bp) for informatics purposes (Janda and Abbott, 2007). The full-length 16S rRNA gene can be used for accurate taxonomic identification; however, due to cost constraints and limitations of NGS technology, one or a few of the 9 hypervariable regions are usually targeted. Amplicon based studies depend on the annealing of amplification primers to conserved regions that flank the variable regions (V1-V9) of the 16S rRNA gene (Myer *et al.*, 2016). Once the 16S rRNA gene or variable region is amplified the amplicon sequences are assessed qualitatively and highly similar sequences are clustered into operational taxonomic units (OTUs), which are then aligned against standard reference database, such as Greengenes (McDonald *et al.*, 2012), SILVA (Yilmaz *et al.*, 2014) or Ribosomal Database Project (RDP) (Cole *et al.*, 2014). This is attained through various bioinformatic pipelines for example, mothur (Schloss *et al.*, 2009), QIIME (Caporaso *et al.*, 2010) or UCHIME (Edgar *et al.*, 2011). Potentially, OTUs may be classified to the species level, where others may only be classified up to genus or family level, due to the “varying resolution in sequencing reads of specific regions of the 16S rRNA gene used for distinguishing different types of bacteria” (Shukla *et al.*, 2017). The OTUs can then be analyzed in terms of presence or absence, (relative) abundance or phylogenetic diversity (Morgan, Segata and Huttenhower, 2013).

The purpose of this chapter is to describe and assess the technical approach used to analyze the respiratory microbiome in children with and without PTB, as well as to describe the challenges of microbiome sequencing in resource limited settings. These results were used to achieve the aims of Chapter 4.

### 3.2 Materials and Methods:

16S rRNA gene sequencing was performed on the baseline and two- and six-month follow-up respiratory samples from children with drug susceptible TB and well-defined ill controls to describe the respiratory microbiomes in these population groups. The wet and dry laboratory workflow utilised for the 16S rRNA microbiome sequencing and analysis is outlined in Figure 3.1.



### 3.2.1 Sample selection

For the microbiome pilot study, nasopharyngeal aspirate (NPA) and induced sputum (IS) samples were obtained from study participants at baseline and follow-up visits at months two and six after enrolment or start of TB treatment, as described in Chapter 1, section 1.10. Twenty-six participants were selected for the microbiome analysis based on the availability of baseline, month two and month six follow up samples. NIH 2015 case definitions were used retrospectively to classify children in the diagnostic categories; confirmed TB cases, unconfirmed TB cases and unlikely TB cases. The confirmed TB cases were bacteriologically confirmed by culture or GeneXpert MTB/ Rif on one or more respiratory samples. The unconfirmed TB cases were not bacteriological confirmed and had to have at least 2 of the following: 1) Well defined symptoms and signs suggestive of TB); 2) Chest radiograph consistent with tuberculosis (dual read by experts); 3) Close tuberculosis exposure or immunologic evidence of *M. tuberculosis* infection; 4) Positive response to tuberculosis treatment (requires documented positive clinical response on tuberculosis treatment). In the unlikely TB cases (well-defined ill controls), TB was ruled out after intensive TB investigation and careful clinical follow-up (Graham *et al*, 2015). To be consistent with the terminology used in this study, the term clinically diagnosed was used for the unconfirmed TB cases. All but one had all three NPAs available (Figure 3.2).

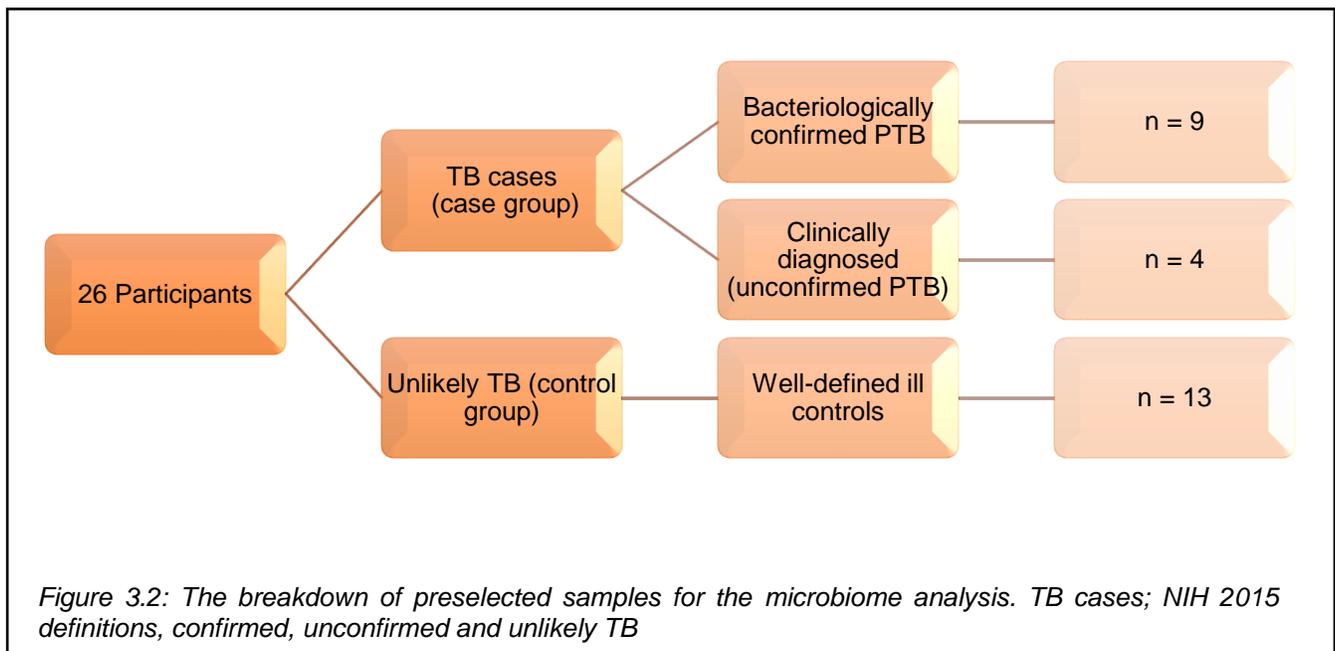


Figure 3.2: The breakdown of preselected samples for the microbiome analysis. TB cases; NIH 2015 definitions, confirmed, unconfirmed and unlikely TB

### 3.2.2 DNA Extraction

DNA was extracted from the respiratory samples using the QIAamp DNA mini kit (Qiagen, Netherlands) and the purity and DNA yield was measured using the BioDrop  $\mu$ Lite (BioDrop, United Kingdom) as described in Chapter 2, section 2.2.2.

The extracted DNA was diluted to 20 ng/ $\mu$ l in 10 mM Tris pH 8.5 (Inqaba Biotechnical Industries (pty) Ltd.) and 2  $\mu$ l of each diluted DNA sample was sent for Qubit analysis at the Central Analytical Facility, Stellenbosch University. The concentration was determined using the Qubit™ 1X dsDNA HS (high sensitivity) Assay Kit (Thermo Fischer Scientific) on the Qubit™ 4 Fluorometer. Qubit analysis is a fluorometric quantification method that uses double stranded DNA (dsDNA) binding dyes to select for dsDNA (over RNA) and accurately quantifies DNA in the range of 10 pg/ $\mu$ l to 100 ng/ $\mu$ l. Based on the Qubit results the DNA samples were further diluted with 10 mM Tris pH 8.5 to 10 ng/ $\mu$ l to ensure that all DNA samples were of similar concentration and to minimise possible PCR inhibitors. The diluted DNA was stored in low DNA binding 0.2 ml PCR tubes at -20°C.

The purity and DNA yield were measured using the BioDrop  $\mu$ Lite after diluting the DNA to 10 ng/ $\mu$ l. The DNA purity was assessed according to standard DNA purity absorbance ratios at A260/A230 (DNA=2.0-2.2) and A260/A280 (DNA= $\sim$ 1.8-2.0).

#### Sequencing controls

Several controls were included to assess for contamination that could have been introduced during the sample preparation (example: viral transport medium) or during DNA extraction (example: elution buffer) for each of the kits that was used (Table 3.1). The storage medium and negative controls were spiked with DNA from a known bacterium (*Serratia* spp.) at a concentration representative of the biological samples (10 ng/ $\mu$ l). This was done so that possible contaminants present in the controls would not be overamplified during PCR i.e. spiking the controls allows the “competing DNA” from a known bacterium to reduce the over amplification of contaminants. *Serratia* spp. was chosen as it was not expected to be a contaminant in the no-template controls and unlikely to be observed in the samples. Four separate QIAamp DNA mini kits were used for DNA extraction and therefore a sample storage control (Elution Buffer) and a negative extraction control (NEC) were included for each DNA extraction kit. In addition, an internal sequencing control (PhiX) was included (and is further described in section 3.2.4.8). All mock control, negative control and clinical samples were prepared in the student research laboratory and the non-template controls (NTC) were prepared in a clean PCR laboratory.

Table 3.1: Sequencing controls.

Control	Experiment Stage	Description	Purpose
<b>Mock community</b>			
Mock 1 - Equal volume	DNA extraction	5 known ATCC bacterial strains obtained from the NHLS	Sequencing control
Mock 2 - Equal concentration			
<b>Within-run repeat</b>	DNA extraction	Any sample in duplicate in a run (random sample selected- Sample 40)	Measure reproducibility
<b>Sample storage medium</b>			
1. Elution buffer* (Kit 1,2,3,4)	DNA extraction	Elution buffer obtained from each kit	Contamination control
2. Viral transport medium*		Virus transport medium put through the entire DNA extraction process for each kit	
3. Tris pH Buffer*		Tris pH buffer used to dilute samples	
<b>PCR grade water</b>			
Nuclease free H2O	PCR	Water used during PCR setup for primer dilution and PCR setup	Contamination control
<b>Negative controls</b>			
1 NEC (Kit: 1, 2,3,4)*	DNA extraction and PCR	Negative extraction control obtained from each kit during sample DNA extraction	Contamination control
2 NTC (PCR Batches 1, 2,3,4)		Non-template control included with each PCR batch	
<b>Positive control</b>	PCR	<i>Escherichia coli</i>	PCR control
<b>5% PhiX</b>	Sequencing run	Served as an internal control for the low diversity library	Internal control

\* Controls spiked with a known bacterium (*Serratia* spp).

The bacterial strains used for the preparation of the mock and spiked controls were obtained from the National Health Laboratory Service diagnostic laboratory at Tygerberg Hospital. DNA from a *Serratia* spp clinical isolate was used for the spiked controls and *Staphylococcus aureus* (ATCC BA 1026), *Klebsiella pneumoniae* (ATCC BAA 1706), *Escherichia coli* (ATCC 25922), *Enterococcus faecalis* (ATCC 51299) and *Pseudomonas aeruginosa* (ATCC 27953) were used for the mock controls (Table 3.1). DNA was extracted from individual pure cultures using the Qiagen QIAamp DNA mini kit using the protocols for Gram-negative and Gram-positive bacterial plate cultures as per the manufacturer's instructions. The purified DNA was eluted in 50µl AE buffer and stored at -20°C until further use.

### 3.2.3 16S rRNA library preparation

The library preparation was performed as described in the Illumina 16S Metagenomic Sequencing Library preparation protocol (Addendum 2) with only minor modifications, as described in the following sections. The basic 16S Library preparation workflow is included in Figure 3.1. The 16S rRNA gene amplification and sequencing was done using the next generation sequencing Illumina Miseq™ (California, USA) platform at the Institute for Microbial Biotechnology and Metagenomics, Department of Biotechnology, University of the Western Cape.

#### 3.2.3.1 16S rRNA Amplicon Polymerase Chain Reaction (PCR)

The 16S rRNA amplicon PCR was performed using modified primers, 515F\_short and 805R\_short (Claassen-Weitz *et al.*, 2018) (Table 3.2), targeting the V4 hypervariable region of the 16S rRNA gene. These primers were modified from primers 515F and 806R (Caporaso *et al.*, 2011) by incorporating ambiguous bases, to allow for more diversity to be detected. Overhang adapter sequences were appended on the 5' end of the primers for compatibility with the Illumina indexing and sequencing adapters.

Table 3.2: 16S rRNA V4 PCR primers.

Primer Name		Primer Sequence 5'-3'
<b>515F</b>	forward overhang primer 1 and locus specific sequence	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG <b>GTGCCAGC</b> <b>HGCYGCGGT</b>
<b>805R</b>	Reverse overhang primer 2 and locus specific sequence	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG <b>GGACTAC</b> <b>NNGGGTNTCTAAT</b>

The locus specific sequences are indicated in blue and the 5' overhang adapters in black. Ambiguous bases are H= A/C/T, Y= C/T, N= A/T/C/G

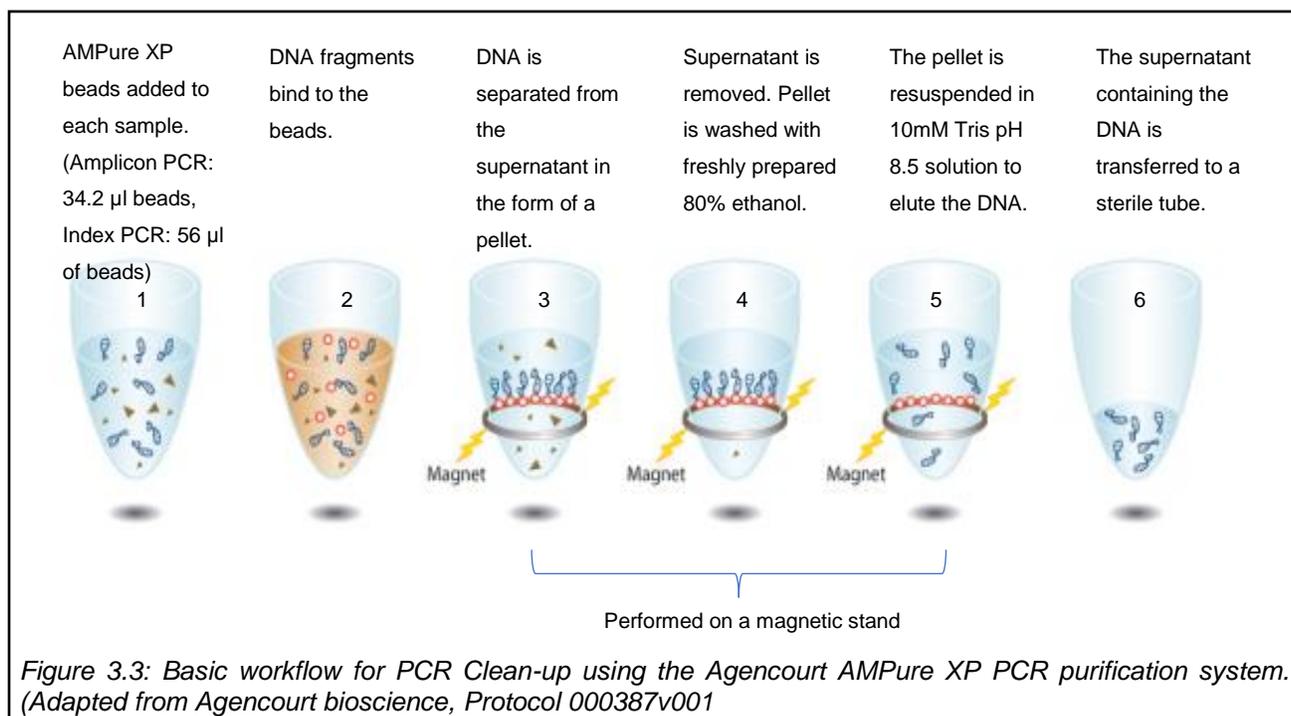
Each PCR reaction consisted of 12.5 µl of Phusion Hot start 2X mastermix (New England Biolabs, USA), 5 µl of the diluted sample DNA (10 ng/µl in 10 mM TRIS), 1.25 µl each of amplicon PCR 515 Forward and 805 Reverse primers (10 µM) and 5 µl of PCR grade H<sub>2</sub>O, to a final volume of 25 µl per sample. The Kapa HiFi Hot start Ready mix (Kapa Biosystems, South Africa) recommended in the protocol was substituted with Phusion Hot start 2X mastermix as it offered more robust high-fidelity performance. The procedure was conducted in low DNA binding 0.2 ml PCR tubes using the Applied Biosystems Proflex PCR system (Life Technologies, United States) that was manually set to the touchdown PCR cycling conditions described in Table 3.3. All baseline and follow up samples from a participant were included in a single PCR batch to avoid PCR bias between samples obtained from the same participant.

Table 3.3: 16S rRNA V4 touchdown Amplicon PCR cycling conditions.

	Temperature (°C)	Time	Number of cycles
<b>Enzyme activation</b>	98	30 sec	X1
<b>Initial denaturation</b>	98	5 sec	
<b>Annealing</b>	65	30 sec	X10
<b>Extension</b>	72	30 sec	
<b>Initial denaturation</b>	98	5 sec	
<b>Annealing</b>	55	30 sec	X20
<b>Extension</b>	72	30 sec	
<b>Final Extension</b>	72	5 min	X1
<b>Hold</b>	4	∞	

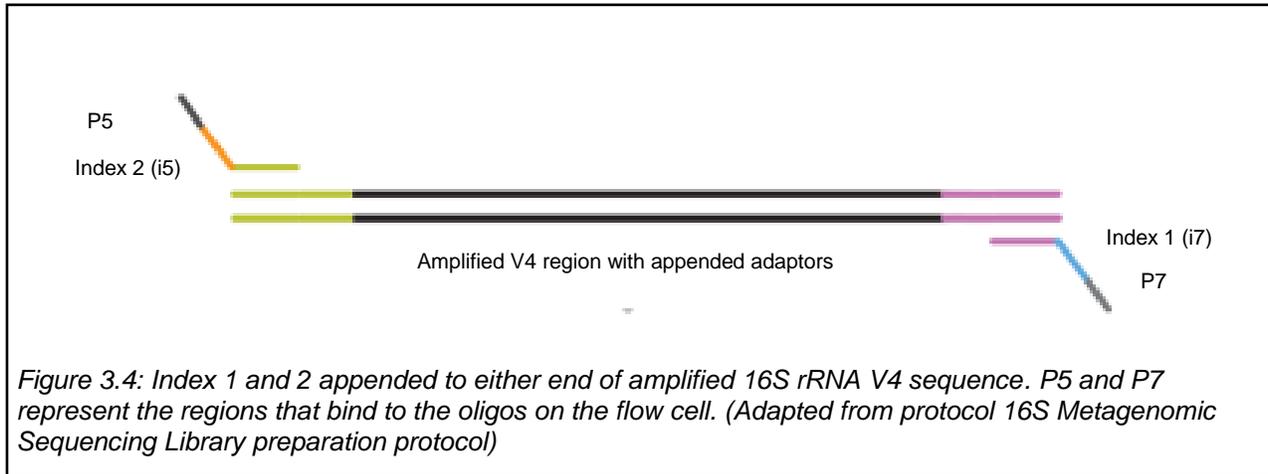
### 3.2.3.2 PCR clean-Ups (1 and 2)

PCR clean up steps were incorporated after each PCR (Amplicon PCR and Index PCR), to purify the amplicon, remove free primers and primer dimers, and to prepare the final library before quantification. This was achieved using the Agencourt AMPure XP PCR purification system, where double (and single stranded) DNA fragments are bound to paramagnetic AMPure XP beads (Beckman Coulter Genomics) in an optimized buffer, freeing the samples from excess primers, nucleotides, salts and enzymes after washing with 80% ethanol. The ratio (v/v) of AMPure XP beads to PCR product used for the Amplicon PCR was 1.8X in 19  $\mu$ l of PCR product, while a ratio of 1.2X in 45  $\mu$ l PCR product was used for the Index PCR. The basic constituents and steps followed for the PCR cleanup are described in Figure 3.3.; the full procedure is described Addendum 2.



### 3.2.3.3 Nextera XT Index PCR

The Index PCR was done to attach dual indices and Illumina sequencing adapters at either end of the amplified sequences (Figure 3.4) using the Nextera Index Kit (Illumina Inc, USA) to create a library. Each sample was labelled with one of each of the 12 unique Index 1 (i7) and eight index 2 (i5) indices, to enable the identification of 96 distinct sample libraries. Each sample and control was assigned an index pair, based on the 96 well plate design (Addendum 3 and 4, respectively).



Each Index PCR reaction consisted of 25 µl of 2x KAPA HiFi Hot start Ready mix, 5 µl Amplicon DNA, and 5 µl each of Nextera XT Index primer 1 (N7xx) (orange caps, yellow solution) and Nextera XT Index primer 2 (S5xx) (White caps, clear solution) and 10 µl PCR grade water (Qiagen, Germany), to amount to a final volume of 50 µl per sample. The procedure was conducted in low DNA binding 0.2 ml PCR tubes and carried out using the Applied Biosystems Proflex PCR system. The PCR cycling conditions were manually set to the following cycling conditions; a denaturation step at 95°C for 3 minutes, an amplification step at 95°C, 55°C and 72°C for 30 seconds each for 8 cycles, followed by a final extension step at 72°C for 5 minutes. The final library was purified using 56 µl of AMPure beads, as previously described in Section 3.2.4.1.

All PCR products were visualized on a 2% agarose gel using the UVitec Cambridge Alliance 2 gel documentation system after each PCR amplification and clean up step.

#### 3.2.3.4 Library Validation

To validate the library construction, 11 samples were selected for Agilent 2100 Bioanalyzer (model G2939B) analysis at the Central Analytical Facility (CAF), Stellenbosch University. The Agilent 2100 Bioanalyzer is a commercially available chip-based nucleic acid analysis system. The microchip consists of micro-channels that are filled with a sieving polymer and an intercalating fluorescent dye. This microfluidic technology allows DNA to be separated according to mass and DNA concentration to be determined, when an electrical charge is applied. Four microliters of purified Indexed PCR product for each sample was submitted and amplicon size and concentration were determined using the Agilent High Sensitivity DNA Kit (Agilent Technologies, Germany). Data were recorded using the 2100 Expert software B02.08.s1648 (SR3).

#### 3.2.3.5 Library Quantification and normalization

The final libraries were quantified using the Qubit dsDNA High Sensitivity (HS) Assay kit on a Qubit Fluorometer, at the Institute for Microbial Biotechnology and Metagenomics, University of Western Cape. This was done to determine the concentrations of the sample libraries before normalization and pooling.

The DNA concentration of each sample was converted to nM using the average library size determined by the Agilent 2100 Bioanalyzer. The uniquely indexed final sample libraries were diluted to 4 nM with 10 mM Tris pH 8.5 in a final volume of 50  $\mu$ l before pooling.

#### 3.2.3.6 Sample library pooling

Pooling was done essentially as described in the 16S Metagenomic Sequencing Library preparation protocol (Addendum 5). Eight separate pools (A-H) were prepared containing 10  $\mu$ l of each of 12 sample libraries, at a concentration of 4 nM each. The final pooling step involved aliquoting 20  $\mu$ l from each pooled library group (A-H) into a single micro-centrifuge tube.

#### 3.2.3.7 Library denaturing and Miseq loading

The v3 reagent kit (Illumina, USA) was used for sequencing; it consisted of the reagent cartridge, Hybridization buffer (HT1), Incorporation buffer (PR2 bottle) and the Miseq flow cell. Five microliters of freshly prepared 0.2 N NaOH was combined with the 4 nM pooled library (5  $\mu$ l) in a microcentrifuge tube and vortexed briefly before incubation at room temperature for 5 minutes. Nine hundred and ninety microliters of pre-chilled Hybridization Buffer (HT1) was added to the 10  $\mu$ l of denatured DNA resulting in a 20 pM denatured library in 1 mM NaOH. The

denatured DNA was placed on ice until final dilution to 8 pM (240 µl of denatured library was added to 360µl of pre-chilled HT1).

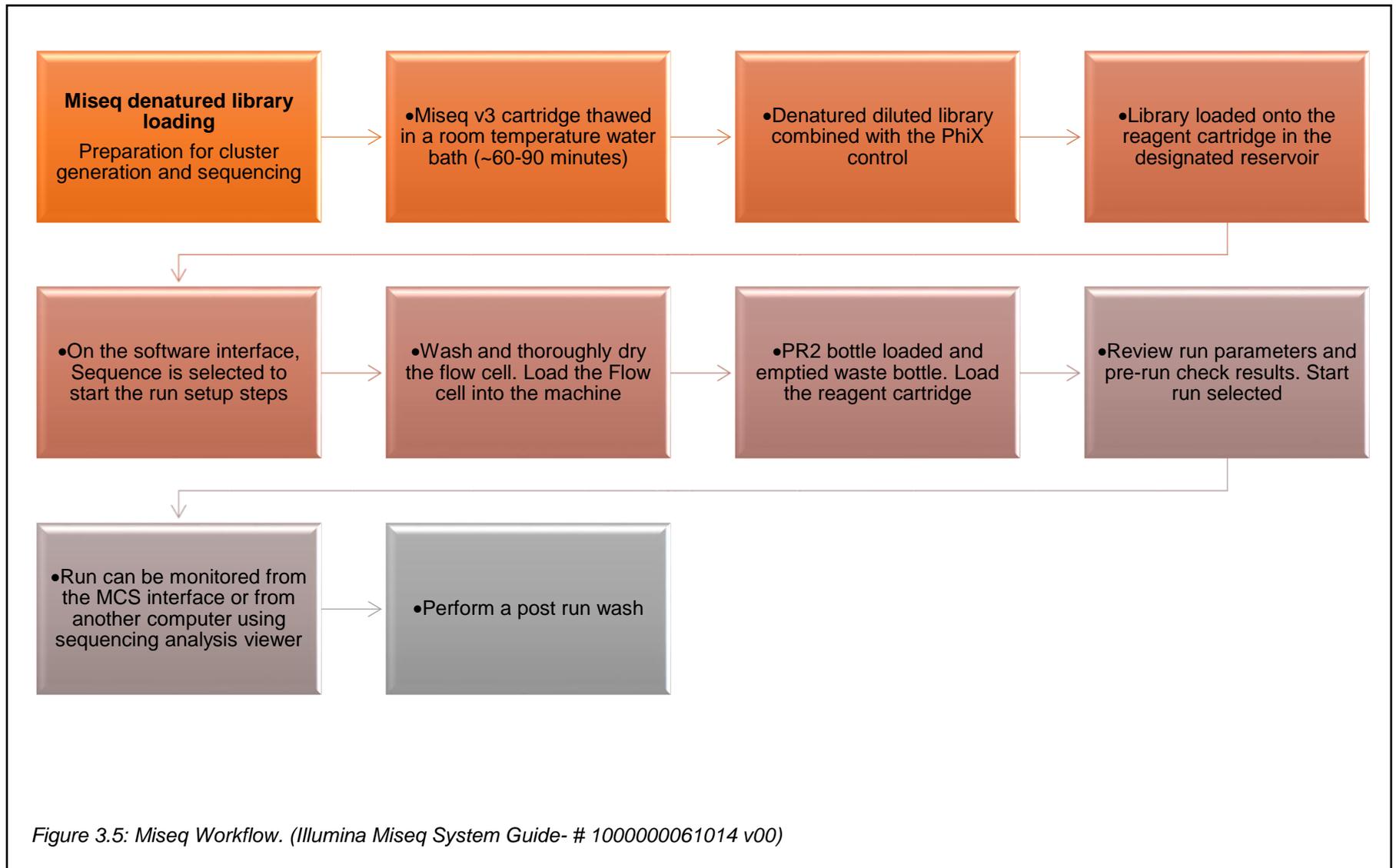
### 3.2.3.8 Denaturation and Dilution of the PhiX control

Ten nanomolar PhiX library (Control Kit v3; FC-110-3001) was diluted to 4 nM by combining 3 µl of 10 mM Tris pH 8.5 with 2 µl of PhiX library and then denatured using 5 µl of 0.2 N NaOH. The solution was quickly vortexed and incubated at room temperature for 5 minutes. To result in a 20 pM PhiX library, 10 µl denatured PhiX library was added to 990 µl of pre-chilled HT1. The denatured 20 pM PhiX library was then diluted to the same concentration as the Amplicon library, by adding 240 µl of the 20 pM denatured library to 360 µl of pre-chilled HT1.

Five hundred and seventy microliters of the pooled amplicon library and 30 µl of diluted PhiX control (both 8 pM) were combined and heat denatured at 96°C for 2 minutes. The final library mixture contained 5% PhiX which served as an internal control for the low diversity library. After incubation the tube was inverted and placed in an ice-water bath for 2 minutes before immediately loading the library into the Miseq reagent cartridge. This was done to ensure efficient template loading on the Miseq flow cell.

### 3.2.3.9 Preparing for sequencing on the Miseq machine

Prior to sequencing, the Miseq machine was set up according to the Miseq system guide and the final components required for sequencing were loaded into the machine. 300 bp paired end sequencing was performed for a total of 602 cycles (2x301 cycles; an extra cycle was added for phasing and prephasing calculations). Ideally, the ends of each read could be overlapped to generate high quality, full length reads of the V4 region in an ~ 65 hour run. The protocol assumed that for 96 indexed samples >100 000 reads could be generated per sample. According to Illumina, at 2 X 300bp paired end sequencing >70% of bases will have a quality score higher than 30 (Q30) (<https://emea.illumina.com/systems/sequencing-platforms/miseq/specifications.html>). The basic sequencing workflow is shown in Figure 3.5; more detail is provided in Addendum 2.



### 3.2.4 Generation of *fastq* files

Sequence information (sequence ID, sequence read, blank and quality score) was stored in the form of a basic file type (text file); *fastq* file. Paired-end data had two files (forward and reverse sequence) that should have the same number of lines and sequence order. *Fastq* files were generated by calculating the amount of successfully completed and extracted cycles and requeuing analysis as explained in Addendum 6.

### 3.2.5 FAST Quality Control (QC)

FastQC was performed on the raw sequencing reads to determine the quality of the sequencing reads and to identify any issues or biases in the data which could affect the downstream analyses. FASTQ files (containing both the biological sequences and quality control information) obtained for each sample were inputted into an open source web-based platform, Galaxy, that used the FastQC package tool version 0.11.7 (Babraham Institute informatics) to determine the sequencing quality. FastQC provides a quality control report which identifies issues which originate either in the sequencer or in the starting library material. The Phred score was used to describe the quality of the output and was presented as an integer value i.e. Q=10, Q=20, Q=30. These values represent the probability that a base is called incorrectly  $\left(10^{\left(-\frac{Q}{10}\right)}\right)$  and therefore the larger the integer value the greater the confidence in the output. The quality control checks the per base sequence quality which shows an overview of the range of quality values across the bases (warning sign indicates that 25% of data is less than 10 or has a median less than 25, failed sign indicates that 25% for any base is less than 5 or if the median for any base is less than 20), per sequence quality informs about sequences with universally low quality values (good sign indicates that the mean Q >27, warning sign indicates that the mean Q <27 (equates to 0.2% error rate) and failed sign indicates that the mean Q <20 (equates to a 1% error rate)), per base GC, per sequence GC and per base N content (N is substituted where the sequencer is unable to confidently make a base call) (warning sign indicates any position shows an N content of >5%, failed sign indicates any position with an N content of >20%) as well as identifying duplicate and overrepresented sequences of a certain length (Kmers) (<https://www.bioinformatics.babraham.ac.uk>). MultiQC (version 1.7) was used to aggregate FastQC output quality indicators from the entire dataset, which was saved as html reports.

The MultiQC reports were used to summarize the number of samples that had forward and reverse reads that met the quality control criteria for per base quality, per sequence quality, the per base N content (conventional bases substituted with N) and to determine the average sequence length of the forward and reverse reads.

### 3.2.6 Data analysis Pipeline

Quantitative insight into microbial ecology (QIIME) is an opensource bioinformatics pipeline that was used to pre-process the raw DNA sequencing data (Caporaso *et al.*, 2010). This section briefly describes the bioinformatic analysis pipeline used to analyze single reads using QIIME (version 1.9.1).

Single (forward) read analysis was done essentially as described by (Dumbrell, Ferguson and Clark, 2016) and subsequent steps were followed according to an established bioinformatics pipeline used at the Division of Molecular Biology and Human Genetics, Stellenbosch University. The samples were demultiplexed immediately after the sequencing run at the University of the Western Cape, therefore the file names were used to identify which sequences belonged to which sample. A quality filtering step was implemented using Phred scores of  $Q \geq 20$  in QIIME before assigning OTUs. OTUs were assigned using an open reference picking approach, where the sequences were clustered at 97% similarity using UCLUST (version 1.2.22) and aligned against the Greengenes reference database (gg\_13\_8, modified 15 August 2013). Sequences that did not match the reference sequence collection were subsequently clustered *de novo*. Open reference picking in QIIME included taxonomy assignment, sequence alignment and tree-building steps. The commands used for generating the outputs can be found in Addendum 7.

Files obtained from QIIME were imported into R studio (version 1.2.5001) to graphically represent and interpret the sequencing data. The packages used in R studio included Phyloseq (version 1.28.0) and the Microbiome R package (Leo Lahti et al (Bioconductor, 2017-2019)). The minimum and maximum number of reads for the 95 samples included in the sequencing run was determined using these packages. OTU counts for the samples and controls were transformed to the relative abundance which included a filter (Abundance > 0.02) that filtered out low abundance taxa. Datafiles were subset for the samples and controls and aggregated at the different taxonomic ranks (phylum, class, order, family and genus). Alpha diversity (e.g. Simpson and Shannon), beta diversity (e.g. Bray-Curtis) and other statistical analysis were also conducted in R and further explained in Chapter 4.

### 3.3 Results

This nasopharyngeal microbiome pilot study included seventy-seven respiratory samples (one participant had no month 2 follow up sample) obtained from 26 participants, and eighteen control samples.

#### 3.3.1 DNA quality assessment of samples

The median DNA concentration of the samples based on Qubit analysis was 54.3 ng/ $\mu$ l (Interquartile range (IQR) 27.6 ng/ $\mu$ l-153.9 ng/ $\mu$ l) and the average A260/A230 and A260/A280 DNA purity ratios were 2.0 (Standard deviation (SD) = 0.9) and 1.9 (SD= 0.5), respectively. Approximately 17% (n=13/77) and 78% (n=60/77) of the samples were not in the desired ranges for the A260/A230 and A260/A280 respectively. Despite the poor quality of some of the samples, all were subjected to PCR because of the limited samples available.

#### 3.3.2 16S rRNA library preparation

The V4 region of the 16S rRNA gene was successfully amplified in all clinical samples and spiked negative controls; no products were observed for the negative template controls. A substantial loss of PCR product was seen after the clean-up steps for each PCR (amplicon and index) for some samples; however, all cleaned amplicon PCR products for samples and controls were used for the index PCR. Two samples failed to produce visible bands after the Index PCR. All amplicon PCR products were between 300-400 bp and the indexed PCR products were between 400-500 bp, as expected. An example of PCR products from the mock control is given in Figure 3.6, indicating an 80-100bp increase in size between the amplicon and index products, resulting from the addition of the adapter and indices.



*Figure 3.6: Representative gel of amplified amplicon and indexed PCR products from a mock control sample: Lane 1: amplified Amplicon PCR product, Lane 2: cleaned amplicon PCR product, Lane 3: Indexed PCR product, Lane 4: cleaned indexed PCR product, Lane L: 100 base pair ladder (New England BioLabs).*

A subset of cleaned V4 amplicon and Index PCR products were subjected to Bioanalyzer analysis to confirm the product sizes and to ensure that indexing was successful. The cleaned amplicon PCR product size was determined to be approximately 375 bp and the indexed PCR product size was approximately 460 bp, consistent with the anticipated increase of 83 bp, based on the size of the index and adaptors.

### 3.3.3 Illumina Miseq sequencing

Sequencing of the pooled 16S rRNA libraries was performed on the Illumina Miseq at UWC. The sequencing run was expected to take approximately 65 hours, producing read lengths of 2 x 301bp, and an output of >20 million reads; with the assumption of >100 000 reads per sample for 96 indexed samples. However, due to load-shedding and failure of the back-up generator during the sequencing run; the run was prematurely aborted resulting in incomplete sequencing of the reverse reads. Five hundred and ninety cycles completed successfully, and the length of the reverse reads was determined to be ~272bp in comparison to the expected fully sequenced read length of 301bp. In addition, the secondary analysis was not performed on the Miseq reporter, therefore no information was obtained from the sequencing run with regard to cluster graphs, sample table summaries and the cluster pie chart representing the classification breakdown for the samples. As a result of the incomplete sequencing run, *fastq* files were manually generated as described in section 3.2.4.

### 3.3.4 Sequencing quality: FASTQ quality control (QC) assessment

The average length of the forward read across samples and controls was 296bp whereas the average length of the reverse read was 271bp based on the MultiQC reports. The per sequence quality scores showed that 95% of the forward reads had a mean Q >27 in comparison to the reverse reads where none passed QC; 72.6% (n=69/95) failed and 27.3% had warning signs. Based on the per base N content 97.9% of the forward reads passed QC, however all of the reverse reads had warning signs indicating that some positions in the read had an “N” content of more than 5%. None of the forward or reverse reads for the samples and controls met the per base sequence content quality criteria indicating that the lower quantile for any base was less than 5 or the median of any base had a Q <20.

Based on the individual FASTQ quality assessment, quality scores for the forward reads averaged Q= >30 up to position 175bp on the read length, while the average quality score for the reverse reads was Q= >20 up to position 100bp, after which the quality score decreased further. The length of the forward read ranged from 35-301bp while the length of the reverse read ranged from 35-272bp for the samples and controls. Figure 3.7 is a graphical representation of the quality scores across the bases for the forward and reverse read from the mock control. Due to the poor quality and length of the reverse reads, subsequent analysis was based on the analysis of only the forward reads obtained from each sample and sequencing controls included in the study.

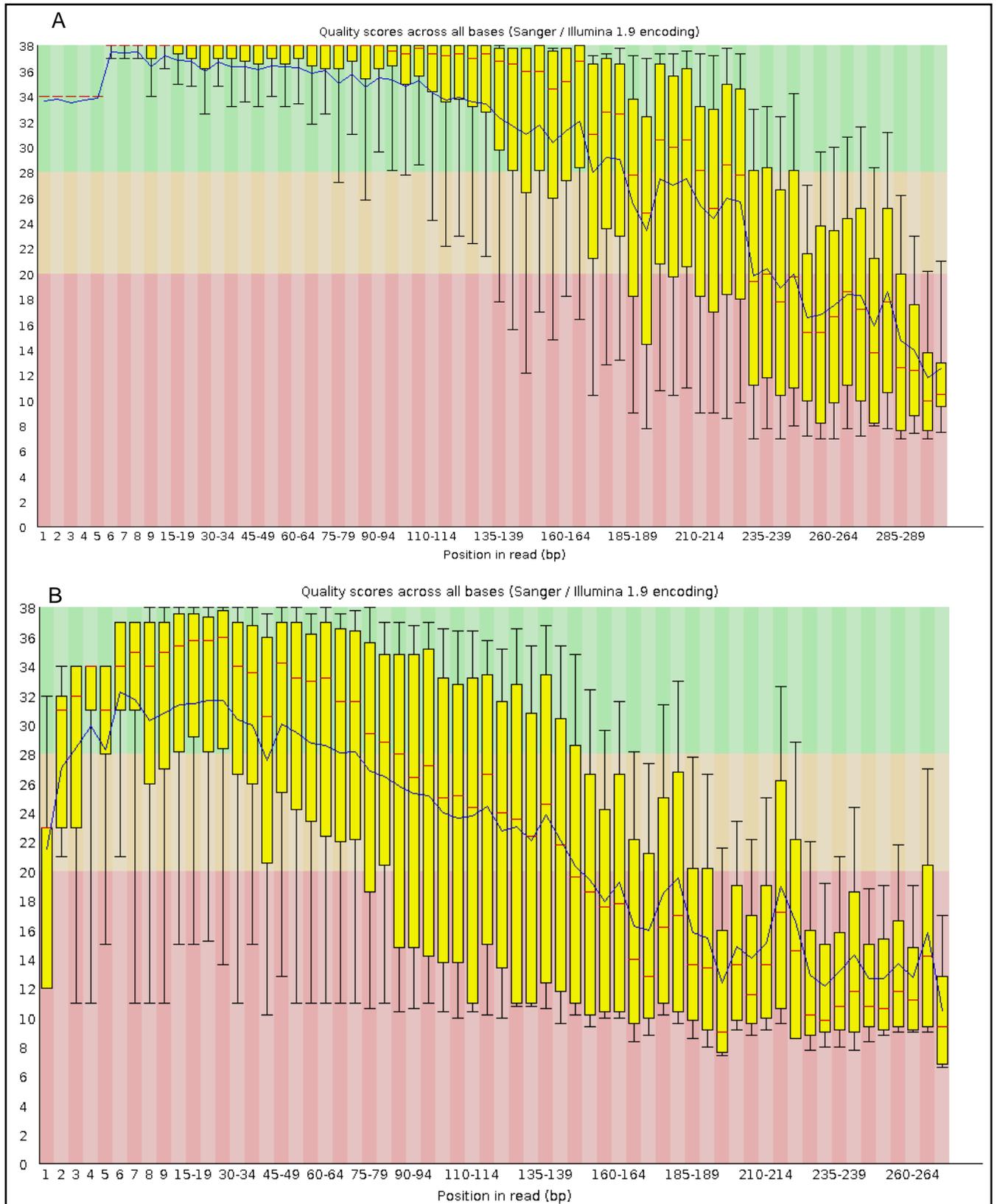


Figure 3.7: A graphical representation of the Fast Quality Control (QC) report for the forward (A) and reverse reads (B) obtained from a mock control

### 3.3.5 Taxonomic classification

The median number of forward reads for the samples was 34 942 (IQR= 25 569). The median number of forward reads for un-spiked controls (H<sub>2</sub>O, MCKEC, MCKEV, NTC1-3, PCR, WRR) was 37 106 (IQR= 43 972) where the median number of reads for the spiked controls (EB1-4, NEC1-4, TRIS and VTM) was 59 203 (IQR= 32 649).

OTUs were clustered at 97% similarity and aligned against the Greengenes database. The taxonomic ranks investigated were phylum, class, order, family and genus; the main focus being on the phylum, family and genus ranks. At the phylum level the most abundant phyla observed across samples and sequencing controls belonged to Proteobacteria, Firmicutes, Fusobacteria, Bacteroidetes and Actinobacteria (Figure 3.8). The data was sub-set into samples and sequencing controls and analyzed separately. This chapter focuses on the classification of OTUs from the controls to inform the analysis of the samples in Chapter 4.

The sequencing controls included two mock controls, a within run repeat, elution buffer, viral transport media, H<sub>2</sub>O, negative extraction control, non-template control and a positive PCR control. The two mock controls were composed of DNA from 5 known bacteria namely *S. aureus*, *K. pneumoniae*, *E. coli*, *E. faecalis* and *P. aeruginosa* and were included as sequencing controls (Table 3.1). Mock control 1 had equal volumes of DNA from each organism regardless of the concentration and Mock 2 had equal concentrations of DNA from each organism. The PCR positive control consisted of a single bacterium, *E. coli*.

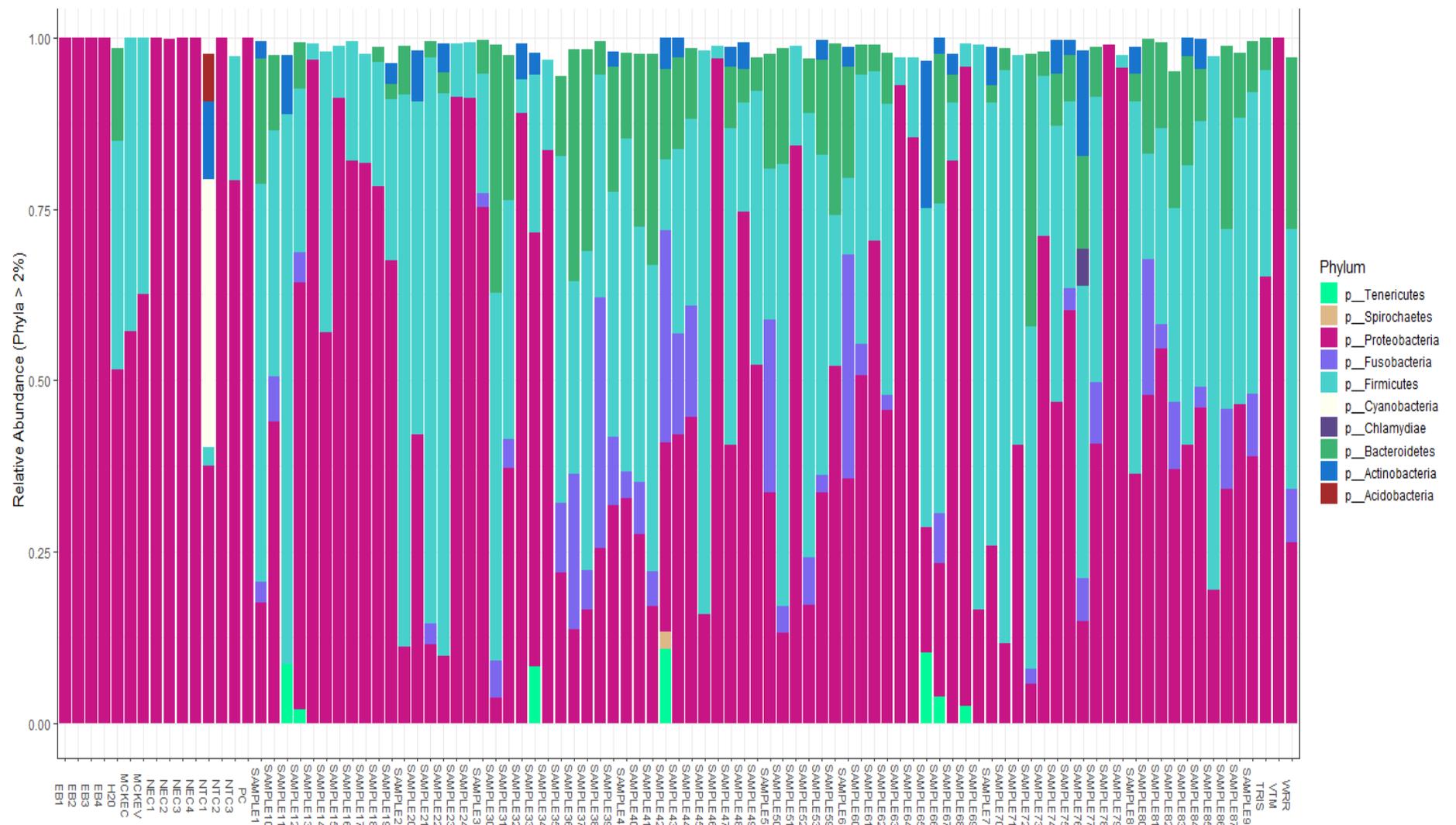


Figure 3.8: The relative abundance of phyla observed across samples and sequencing controls. WRR- Within run repeat, VTM- Viral transport media, TRIS- Tris buffer, PC- Positive control, NTC- Negative template control, NEC- Negative extraction control, MCKEV- Mock control (equal volume), MCKEC- Mock control (equal concentration) and EB- Elution buffer.

Proteobacteria and Firmicutes were the dominant phyla identified in the mock controls, while only Proteobacteria was identified in the positive control (Figure 3.9). At the genus level, *Staphylococcus* sp, *Pseudomonas* sp and *Klebsiella* sp were successfully identified in both mock controls (Figure 3.10). However, *Escherichia* sp and *Enterococcus* sp were not identified and may be included in the large portion of unclassified genera. Similarly, *E. coli* could not be classified in the positive control and a large portion of unclassified genera was observed. *Serratia* sp was also identified in both the mock and positive controls, which was not expected, but it was present at a lower relative abundance in comparison to the other genera detected. The number of reads associated with the genus *Serratia* sp in the mock controls (number of reads: 914 (MCKEC) and 1264 (MCKEV)) and positive control (number of reads: 2884) were less than the spiked controls which had a median read count of 54887 (IQR 4543-71773) associated with the genus *Serratia* sp.

Due to the large proportion of unclassified genera in the mock and PCR positive controls, classification to a higher taxonomic rank (family) was considered to provide a more accurate description of taxa. However, the expected proportions of each family (Enterobacteriaceae: Enterococcaceae: Staphylococcaceae: Pseudomonadaceae: 2:1:1:1) were not observed in the mock controls (Figure 3.11). In addition, Planococcaceae were detected at the family level, which was not expected.

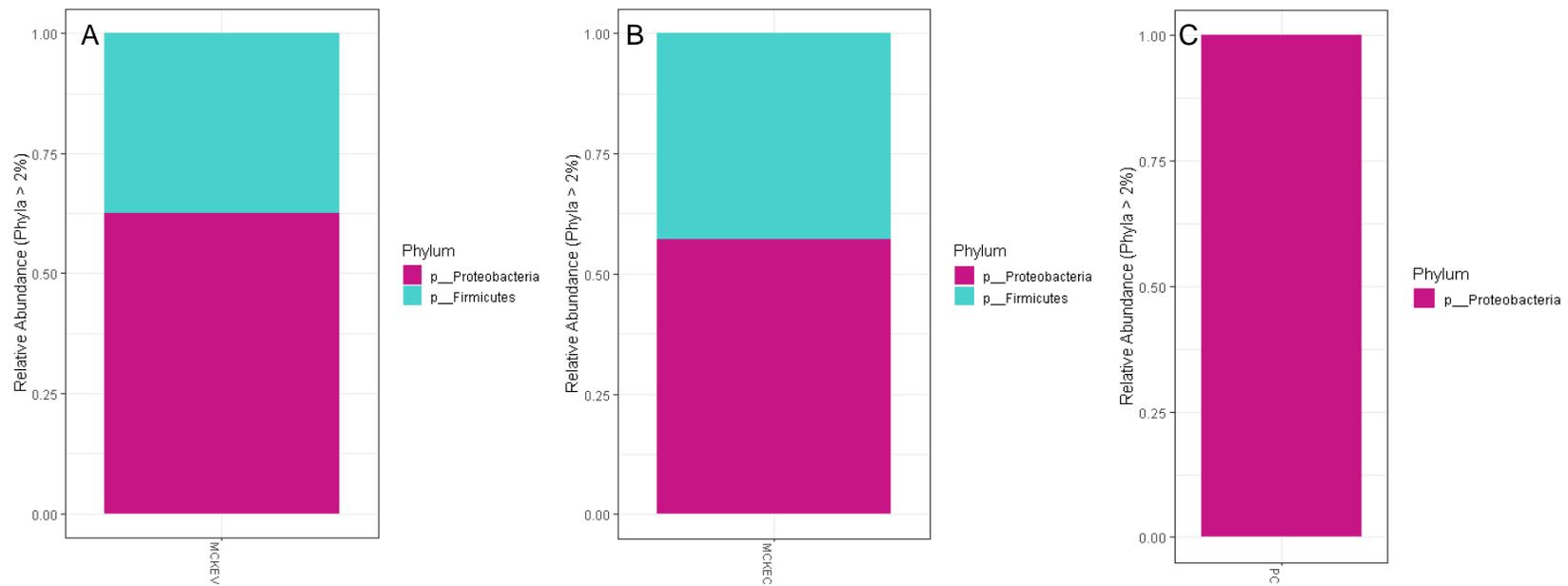
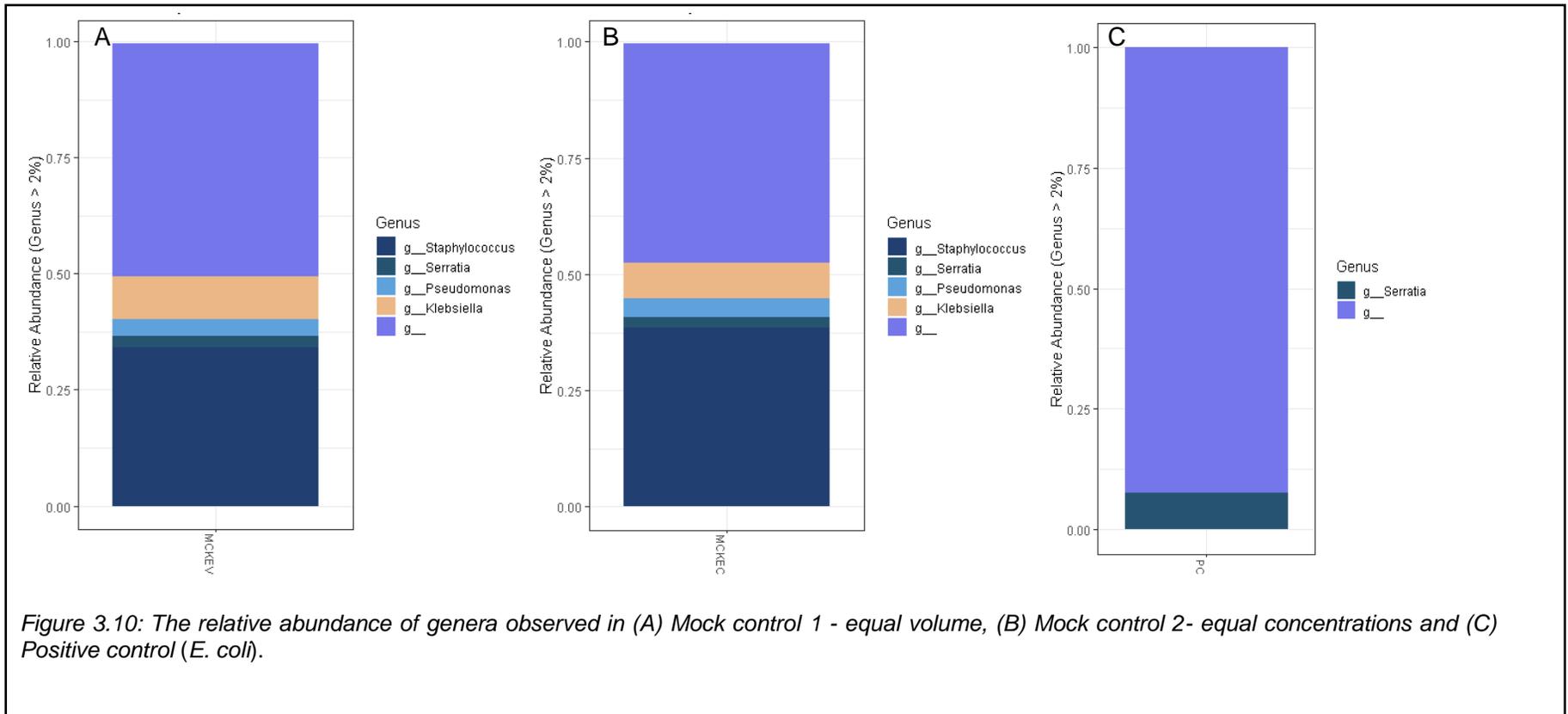
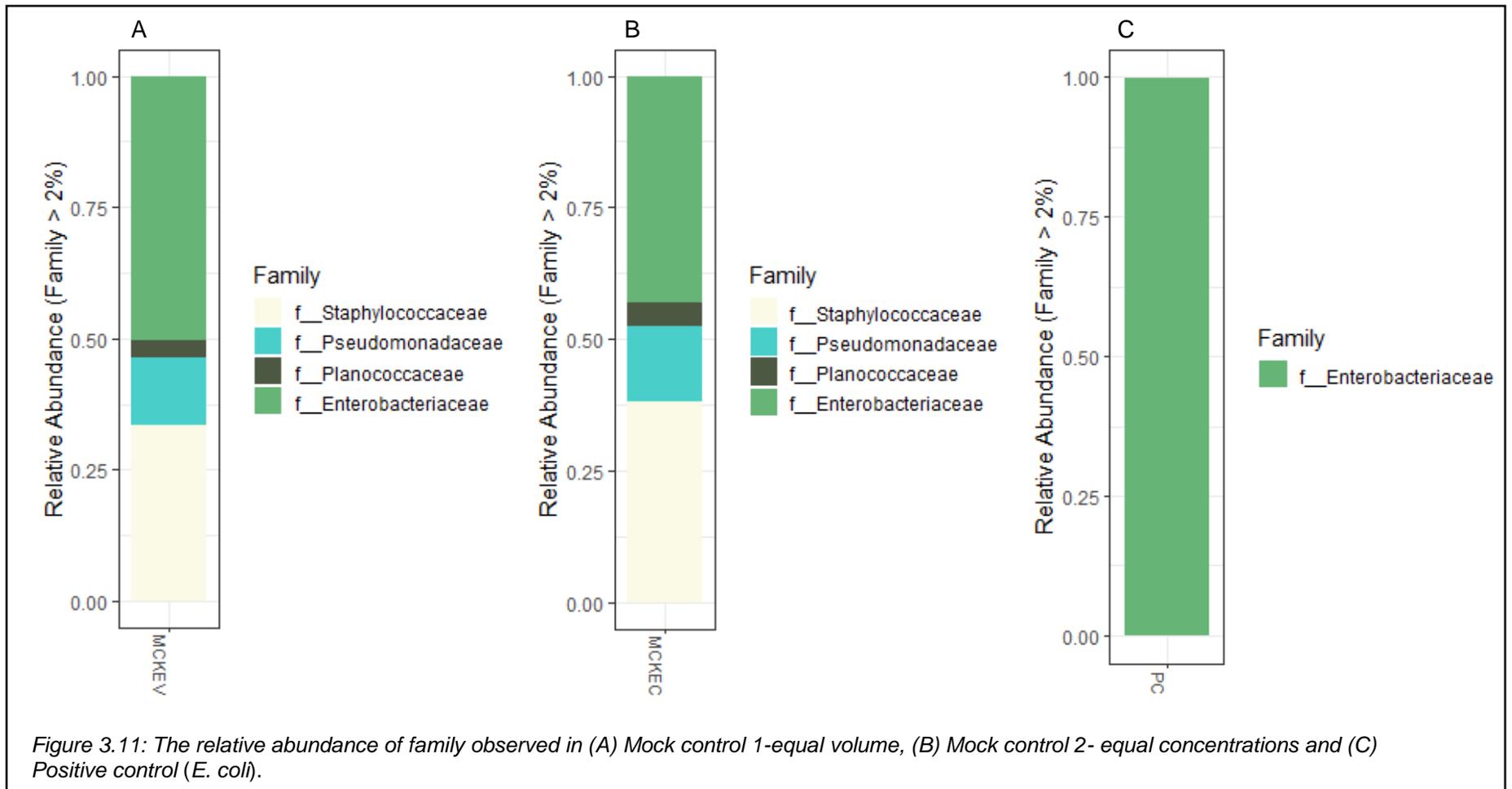
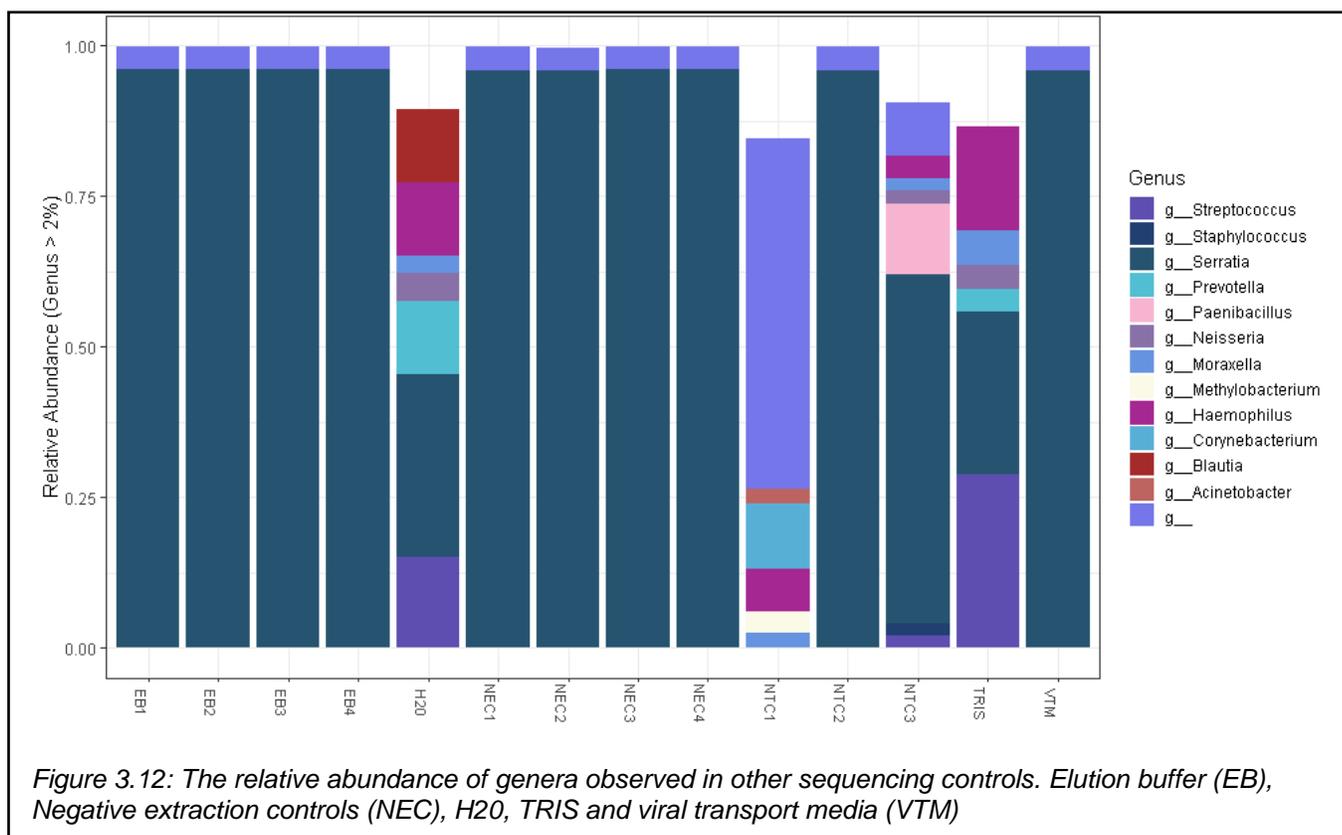


Figure 3.9: The relative abundance of phyla observed in (A) Mock control 1-equal volume, (B) Mock control 2- equal concentrations and (C) the PCR positive control (*E. coli*).





The other controls included were used to identify possible contaminants introduced during DNA extraction (NEC), PCR amplification (H<sub>2</sub>O and Non-template control (NTC) and storage (EB, TRIS and VTM). All of the spiked controls (EB, NEC, VTM) except TRIS produced similar results (Figure 3.12), with *Serratia* as the predominant genus; consistent with the spiked DNA. Only 63 reads were obtained for the TRIS control, which may have contributed to the difference in profile. NTC 2 also resembled the spiked controls although it was not spiked. This may have been a result of human error: the controls may have been indexed incorrectly during the indexing PCR. While H<sub>2</sub>O, NTC1 and NTC3 had the most artefacts or genera detected, they were also the controls with the smallest number of reads present (number of reads 66 (H<sub>2</sub>O), 335 (NTC1) and 255 (NTC3) which may indicate that contamination is low. Analysis to the family level did not provide any additional information regarding the OTUs detected (Addendum 9), especially for the large proportion of unclassified genera observed in NTC 1.



### 3.4 Discussion

Chapter 3 focused on describing the technical aspects involved in conducting the microbiome pilot study; the DNA quality, sequencing quality based on the assessment of the sequencing controls and the challenges faced in low -resource settings.

#### DNA quality

Most of the DNA isolated from the respiratory specimens did not meet the desired DNA purity requirements. This may be due to the low bacterial burden in some of the respiratory specimens (low biomass samples) included and the fact that low DNA concentrations can influence DNA quality scores (Lucena-Aguilar *et al.*, 2016), in addition to the duration and condition of storage which may have compromised the DNA in the samples. Also, a column-based kit was used which may have contributed to the poorer quality; this type of kit usually produces poorer quality DNA; however, they are generally accepted to be the best for these sample types. Although the DNA purity was suboptimal, all except 2 samples were PCR amplifiable. Sample 65 was one of the samples with the poorest purity measures (A260/A280 and A260/A230) where sample 76 had somewhat acceptable purity readings but may not have amplified due to the presence of PCR inhibitors.

#### 16S rRNA amplification

In our study the V4 region of the 16S rRNA gene was amplified. The change in the target region from the standard V3 – V4 Illumina protocol implemented at UWC was based on the fact that with a shorter target region, a full overlap of the target region could be achieved with the Illumina paired end sequencing and possibly reduce sequencing error (Schirmer *et al.*, 2015). The V4 region is also one of the sub regions that best represents the full length of the 16S rRNA gene (Yang, Wang and Qian, 2016). The starting DNA concentration for the Amplicon PCR was increased from 5ng/µl to 10ng/µl to increase the PCR yield. This was done to reduce the impact of the substantial loss of PCR product that occurred during the clean-up steps. The reason for the loss of PCR product could not be determined, but may include the extended period of time (i.e. 3-4 weeks) over which the clean-ups were conducted, or the AMPure beads being over dry (drying beads for longer than 5 minutes after ethanol wash step) as suggested by Quail *et al* (2009), as the protocol followed in our study included drying the beads for 10 minutes.

It should be mentioned that the PCR (Amplicon and Index) steps were conducted in our setting (Medical Microbiology research laboratory, Stellenbosch University) and then the amplified PCR products were transported on ice to the University of the Western Cape to do the clean-up and

subsequent library preparation steps. Under ideal circumstances these steps would be conducted in a single facility with designated laboratory stations for each step in the protocol; however, this was not possible due to limited resources and time-constraints regarding equipment availability.

### Illumina Miseq sequencing

The sequencing run was expected to run for 602 cycles (additional cycles added for phasing/ pre-phasing calculations), however, due to well documented problems with electricity generation and infrastructure in South Africa, our sequencing run was aborted at around 590 cycles as a result of load shedding and failure of the back-up generator. As a result, no information was received with regard to the sequencing run based on the Miseq software outputs. Furthermore, the interrupted sequencing run may have impacted the length of the reverse read but it did not necessarily contribute to the poor sequence quality of the reverse read or the average quality of the forward read.

It is likely that the sequencing quality of the forward and reverse reads was a result of the poor starting DNA quality, but it may also be due to other factors. A useful source of information regarding sequencing quality may have been explained in the Miseq sequencing run report as it would have provided useful insight into the overall sequencing quality (the sequencing run information was not provided by the facility). This would have provided information on the cluster density which is a critical metric for measuring sequencing performance, as it impacts data quality and yield from a run (influences run quality, reads passing filter, Q30 scores and total data output) (Illumina, 2019). Fadrosch *et al* (2014) reported that although Illumina is useful for the characterization of microbial communities using 16S rRNA amplicon sequencing, it does not perform well on low sequence diversity samples (or in samples dominated by only a few bacteria); the overall quality was reported to be significantly lower in comparison to a more diverse or random library. This is a result of issues inherent to the technology, during cluster identification and phasing/pre-phasing rate determination, where a balanced base composition through the initial 12 to 18 cycles of the run is required to generate high quality data (Fadrosch *et al*, 2014). A PhiX control was proposed to modulate the overall sample base composition in low diversity samples to facilitate a successful sequencing run (Fadrosch *et al*, 2014). In our pilot study a 5% PhiX control was included to provide a balanced fluorescent signal at each cycle to improve the quality of the sequencing run. Since the overall sequencing quality was not optimal it may indicate that the PhiX control added was insufficient for this dataset. This should be investigated further as a spike-in of up to 50% can be included for low diversity samples (Fadrosch *et al*, 2014) depending on the Miseq software used and the specific sequencing technology (Genome

Analyzer or HiSeq). Furthermore, Fadrosch *et al* (2014), proposed incorporating a heterogeneity spacer to primers that contains your target region, index and linker sequence and is optimized for the Illumina sequencing platform, to address the low sequence diversity issue, as it offers a more balanced composition throughout the sequencing run. This method was however optimized for the V3-V4 region and may have to be optimized when using paired end sequencing for other target regions. It is known that sequencing quality falters towards the end of the read and that the reverse read quality is lower than the forward read due to the depletion of reagents with increased cycles. However, the exceptionally poor quality of the reverse read in this study was not expected. Personal communication with our collaborator at the University of the Western Cape informed that a similar issue was encountered specifically when using the v3 reagent kit, in comparison to when using the v2 kit which sequences shorter reads (2x250bp).

The length of the reverse read was determined to be ~272 bp at ~ 590 cycles. Considering that the number of cycles is equivalent to the length of the read, the reverse is shorter than it should have been at 590 cycles by ~17 bp; in addition to the fact that sequencing was aborted before the remaining cycles could be repeated. The shorter read length may be due to phasing occurring during the sequencing run. Phasing causes sequencing of some molecules to lag behind due to problems with enzyme kinetics such as the 3' terminator or the fluorophore not being removed. With increased sequencing cycles the number of affected sequences increases, therefore limiting the overall length of the read (Schirmer *et al.*, 2015). Some clusters in the forward read could also have encountered this phenomenon since the average read length was 296 bp, although it did not vary from the fully sequenced read length by as many base pairs. Considering the aborted sequencing run and the poor sequencing quality, most service laboratories may have re-sequenced the library, however this was not feasible for this study as there was insufficient DNA, time and resources available to do so.

### Data analysis

Paired end sequencing could be viewed as a potential error correcting measure when forward and reverse reads are overlapped to form a consensus sequence that is less erroneous (Schirmer *et al.*, 2015). Therefore, with the 2x 300 bp paired end sequencing of the V4 region (291 bp Amplicon product size) a full overlap of the region could have been expected and thus provided sufficient confidence in the consensus sequence that could be generated after aligning the forward and reverse reads. However, since the reverse reads were of suboptimal quality, a consensus sequence was not generated and therefore subsequent analysis was based on the forward read alone. The forward reads were imported into QIIME in the form of *fastq* files where

default quality filtering steps were implemented. This was done to improve the sequencing quality of the reads without losing too many sequences because of stringent quality control measures. Additionally, since the output of reads was low, including more filtering steps could result in more data loss. This process is somewhat subjective as it becomes a tradeoff between quality cutoffs and having sufficient data to work with. The absence of more stringent quality control steps may have contributed to the inability to classify all taxa at the genus level (discussed further below) and may have contributed to bias in downstream analyses as a result of potentially incorrectly called nucleotides not being removed. In future studies, sequencing errors caused by Illumina sequencing could be reduced by trimming sequences based on the quality/ or length of the reads (FASTQ/A trimmer from the FASTX toolkit was used in this study to trim sequences to a specific length i.e. all sequences >301 bp), overlapping reads using either PANDAseq, which has been shown to be effective in removing errors but reduces the number of aligned sequences, or PEAR which retains the number of aligned sequences but reduces errors to a lesser extent, and lastly using an error correction program such as BayesHammer (Schirmer *et al.*, 2015).

### Contamination

Respiratory samples are considered to be low microbial biomass samples, which are prone to reagent and laboratory contamination, which means that microbial reads that are detected could potentially be from environmental sources. Therefore, in this pilot study several sequencing controls were included to assess possible contamination as well as the quality of sequencing, based on mock controls. Negative controls (NEC, VTM, TRIS and EB) included in our study were spiked at the same volume and concentration as our samples, so that the level of contamination in the negative controls would be representative of the study samples. All the spiked negatives had a high relative abundance of the spiked bacteria (*Serratia*) present with a smaller proportion which could consist of some contaminating DNA sequences. As the relative abundance threshold excluded all sequences with less than 2% abundance, contaminating sequences that were less abundant would have been excluded and are therefore not likely to negatively influence the sample data. Although all surfaces were decontaminated and all materials used were subjected to sterilization by UV light, possible contaminants were detected in the non-spiked controls (H<sub>2</sub>O, NTC 1 and 3) which were also the controls with the least reads present (in addition to TRIS). Some of the contaminants such as *Methylobacterium*, *Acinetobacter*, *Corynebacterium* and *Streptococcus* have previously been reported in literature (Salter *et al.*, 2014). The other contaminants (*Staphylococcus*, *Serratia*, *Prevotella*, *Paenibacillus*, *Neisseria*, *Moraxella*, *Haemophilus*, *Blautia*) present in these controls could be due to external contamination from

research subjects, investigators, laboratory surfaces, air or even have been present in laboratory reagents (Davis *et al.*, 2018). Additionally, other reasons for the presence of unexpected OTU's may be a result of spillover and/or microdroplet sprays that occurred during library preparation and index hopping that occurred during sequencing (DNA indices switched from one sample to another resulting in the misassignment of indices to samples/controls). Hornung *et al* (2019) reviewed numerous studies to show the issues and standards of controls included into microbiome research. For negative controls it was advised to focus on the number of reads; a negative control could be considered clean if it has fewer reads and therefore major contamination could be excluded. In our study the non-spiked controls had fewer reads than the samples, which suggests that contamination was not a major problem in this study. The outcome of negative controls should strongly be considered in low microbial biomass samples because low level of contamination can impact the results. Simply removing OTUs from negatives has been suggested but is only applicable if one is certain that these OTUs are actually contaminants and not truly present in the samples (Hornung, Zwartink and Kuijper, 2019). In this study common respiratory pathogens (*Staphylococcus*, *Neisseria*, *Moraxella*, *Streptococcus* and *Haemophilus*) were identified in some of the negative controls and therefore simply removing OTUs would not be applicable.

The mock controls and PCR positive controls were used as PCR and sequencing controls. These controls were included to evaluate the sequencing run. Based on the mock and positive controls some genera (*Escherichia* and *Enterococcus*) could not be classified. This may have been a consequence of the poor sequence quality or the fact that the sequencing read was not specific enough to distinguish the taxa at the genus level. It could imply that the sequencing quality was not sufficient for some organisms. However, the family level could be used to improve the interpretation, especially for the unclassified genera. The importance of including controls in microbiome research has become evident in recent years, and although there is no standard method for processing controls, it is highly recommended that they be included in microbiome studies (Hornung, Zwartink and Kuijper, 2019). Although some of the control taxa weren't classified, they were not expected in the high abundance in the study samples. Therefore, based on the analysis discussed in this chapter, OTU classification to the phylum and family levels was used to address the aims of Chapter 4.

## Limitations

As previously described, the DNA extraction quality in this pilot study was not optimal and may have influenced the sequencing quality. No sequencing run reports were obtained from the sequencing facility; this may have provided more insight into the overall sequencing quality. No samples were excluded because of the limited number of samples available for the study, as well as insufficient sample quantities for the re-extraction of DNA; this also restricted the possibility of resequencing the libraries. Nasopharyngeal samples are low biomass samples, which may have impacted on the poor DNA quality. Samples were stored at suboptimal temperatures (-20 °C). It is imperative to ensure that in future studies, samples are stored at optimal conditions (-80 °C) and that DNA quality is of acceptable purity and yield, as both of these factors contribute to the robustness and reproducibility of sequencing results. Another limitation of this study is that even though paired end sequencing was done, only the forward read was used for generating OTUs; because of this the sequencing reads may not have been sufficient quality to obtain reproducible sequencing data.

Microbiome research has shown that consistency is key for ensuring robust and reproducible results. Due to the fact that the study was conducted in a resource limited setting, certain challenges were faced including conducting the library preparation across two laboratories, transporting samples between the laboratories and load shedding. This could contribute to contamination as one does not have complete control of the working environment, and sample deterioration could occur as a result of fluctuating temperatures during transportation. Additionally, as a consequence of load shedding the sequencing run was interrupted and this highlighted a challenge of conducting research within research limited settings.

### 3.5 Conclusion

Next generation sequencing is fast evolving and is being implemented in research facilities for microbiome research. Microbiome research is a complex research field that requires numerous methodological approaches to be carefully considered to obtain adequate results, from sample collection to sequencing analysis. However, there is no consistency as various methodological approaches can be used. Furthermore much of what is known about microbiome research over the past 10 years has been pioneered by gastrointestinal microbiome studies, which accounts for about 40% of microbiome research (NIH Human Microbiome Portfolio, 2019). This implies that other microbiome niches such as the respiratory microbiome, which was previously thought to be a sterile site (Gallacher and Kotecha, 2016), are still being established and appropriate and adequate methodological approaches remain to be determined. It is therefore necessary to check the validity and the effects of certain methodological processes to ensure rigorous scientific research for the particular microbiome niche being investigated, especially for low biomass niches such as the respiratory tract.

To improve on these challenges, particularly for respiratory microbiome research, samples should be collected and stored at a consistent temperature, DNA extraction techniques should be evaluated to determine which method provides the highest yield and purity for low biomass samples, and the library preparation optimized depending on the sequencing technology and where the samples will be sequenced. Lastly, the data analysis pipeline should include quality control steps that would allow for cleaner sequencing data without losing too much sequencing information.

In this study it was observed that the overall DNA quality and sequencing quality was suboptimal and may have impacted the sequencing results and hence the generation of OTUs. This chapter highlighted the limitations to conducting microbiome research in resource limited settings. Due to the inconsistency of classification at genus level during optimization it was decided to perform further taxonomic classification on the sample set at higher taxonomic levels such as phylum and family level; and these analyses will be applied to address the aims of Chapter 4: (1) to compare the baseline respiratory microbiota of children with bacteriologically confirmed, clinically diagnosed and unlikely PTB and (2) to describe the effect of TB treatment on the respiratory microbiota in children with PTB.

## CHAPTER 4:

### Tuberculosis and the Nasopharyngeal microbiome (microbiota)

#### 4.1 Introduction

The respiratory tract plays a role in respiratory health, as described in Chapter 1. More specifically, the nasopharynx could be involved in susceptibility to respiratory infections and studies have found associations between respiratory microorganisms and respiratory tract infections, including pulmonary tuberculosis (PTB). However, the focus of previous research has been on the lung microbiome in adult tuberculosis (TB) patients and healthy controls. Research has shown that the microbial diversity differs between TB patients and controls (Cui *et al.*, 2012; Botero *et al.*, 2014), specifically that the relative proportion of respiratory microbiota at phylum and genus level can differ between TB case and control groups (Eshetie and Van Soolingen, 2019). Additionally, these studies highlight that *Mycobacterium tuberculosis* (*M.tb*) may not be the sole agent responsible in TB disease, but that interactions with other microbial pathogens and immunological factors also play a role (Eshetie and Van Soolingen, 2019).

Although the lung is the major organ involved in PTB, disease development may occur prior to reaching this point. Some studies have found changes in the nasopharyngeal microbiome during respiratory diseases such as pneumonia and cystic fibrosis (Sakwinska *et al.*, 2014; Prevaes *et al.*, 2016). One study found unique bacterial colonization patterns associated with cystic fibrosis in infants prior to prophylaxis and antibiotic use (Prevaes *et al.*, 2016) while a second study found that pneumonia etiologies could be associated with certain microbiota (Sakwinska *et al.*, 2014). The nasopharyngeal microbiota has also been shown to determine the spread of infection to lower airways, the severity of accompanying inflammatory symptoms as well as the risk of future asthma development in infants (Teo, Mok, Pham, Kusel, Serralha, Troy, Barbara J. Holt, *et al.*, 2015).

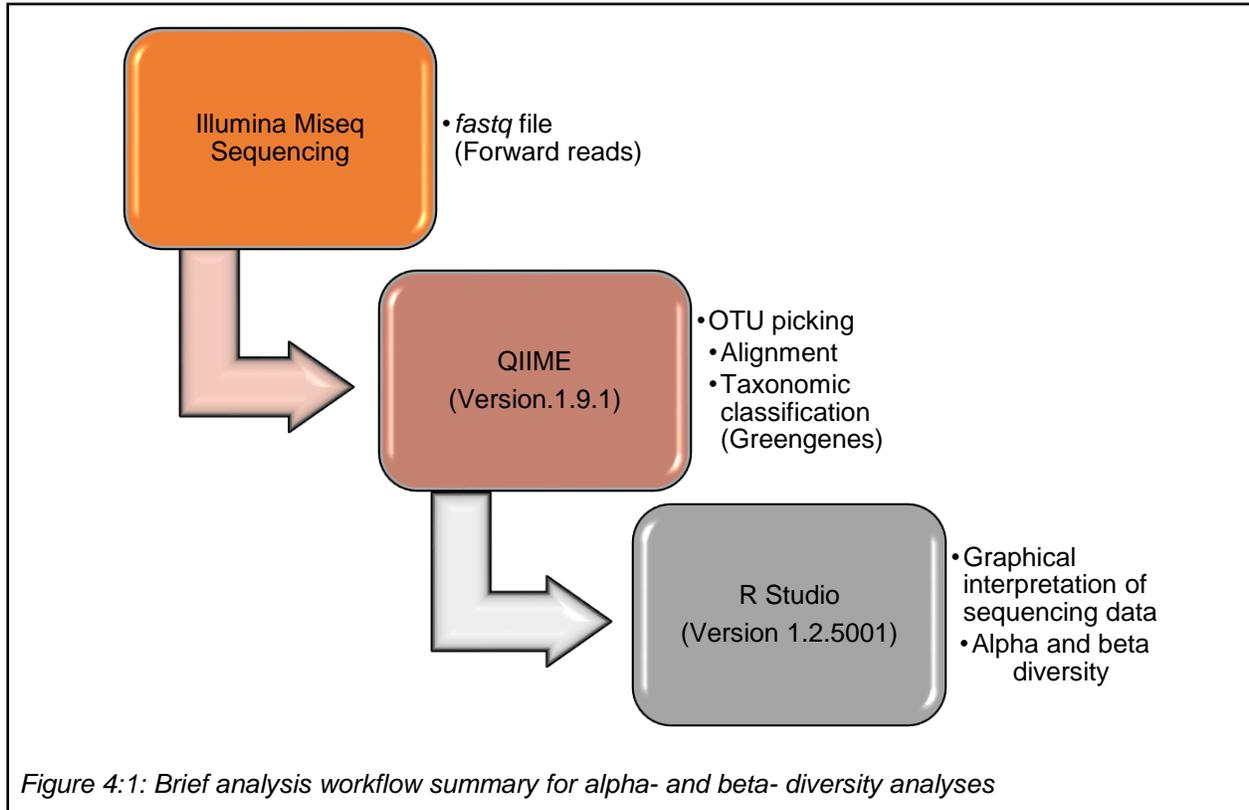
When it comes to the nasopharyngeal microbiome and PTB, limited data is available. The first detailed study to describe respiratory “microbiota” in children with suspected TB in a TB endemic setting was conducted in Cape Town, South Africa (Dube *et al.*, 2016). However, this study based the description of respiratory pathogens present in nasopharyngeal swab samples detected using a multiplex real time PCR based method, which does not allow characterization of the full range and complex network of microorganisms (especially bacteria) present in the nasopharynx, which is possible when using high throughput sequencing. Furthermore, the effect of TB treatment on the nasopharyngeal microbiome has not yet been investigated.

This is important since antibiotics are known to induce dysbiosis which could contribute to future disease as a result of an unstable microbiome or loss of diversity; and studies have shown that TB treatment has the potential to cause dysbiosis (Namasivayam *et al.*, 2017; Wipperman *et al.*, 2017).

Chapter 4 aims to compare the respiratory microbiota of children with PTB, bacteriologically and clinically diagnosed PTB, and unlikely PTB and to describe the effect of TB treatment on the respiratory microbiota in children with PTB. Microbiota is used to describe microbial communities (bacteria, archaea, protists, fungi and viruses) (Eshetie and Van Soolingen, 2019) but the focus for this study is on the bacterial assemblage of microorganisms in the nasopharynx.

## 4.2 Material and Methods

The wet laboratory and dry laboratory processing for the characterization of the nasopharyngeal microbiota in children with suspected TB is fully described in Chapter 3, section 3.2. Figure 4.1 serves as a brief analysis workflow summary for this chapter for determining alpha and beta diversity.



Microbiome diversity can be described in terms of alpha and beta diversity, which describe the within and between sample diversities respectively (Wagner *et al.*, 2018). Two factors contribute to microbiota diversity, namely richness and evenness, where richness is the number of different taxa observed without regard to their frequency and evenness refers to the equitability of the taxa frequencies in a community (Wagner *et al.*, 2018).

Alpha diversity was assessed using the most commonly used alpha diversity measures, the Shannon index and Simpson index (Wagner *et al.*, 2018) (Table 4.1). The Shannon and Simpson indices both provide insight into the diversity within a sample, i.e. the community within a sample. With greater diversity a higher index estimate can be expected. The index values range from 1.5 to 3.5 and 0 to 1 for the Shannon and Simpson indices, respectively (Table 4.1).

Table 4.1: Alpha diversity measures, index range and calculation.

Alpha diversity index	Measurement	Index range	Calculation
<b>Shannon Index</b>	Richness and evenness evenly weighted	1.5-3.5	$Shannon (H') = - \sum_{i=1}^s p_i \ln_b p_i$
<b>Simpson Index</b>	Evenness is more heavily weighted	0-1	$Simpson (D_1) = 1 - \sum_{i=1}^s p_i^2$

$p_i =$  proportion of species  $i$ ,  $s =$  number of species,  $\sum_{i=1}^s p_i = 1$ ,  $\ln =$  natural logarithm,  $b =$  base of natural logarithm (Oksanen, 2019)

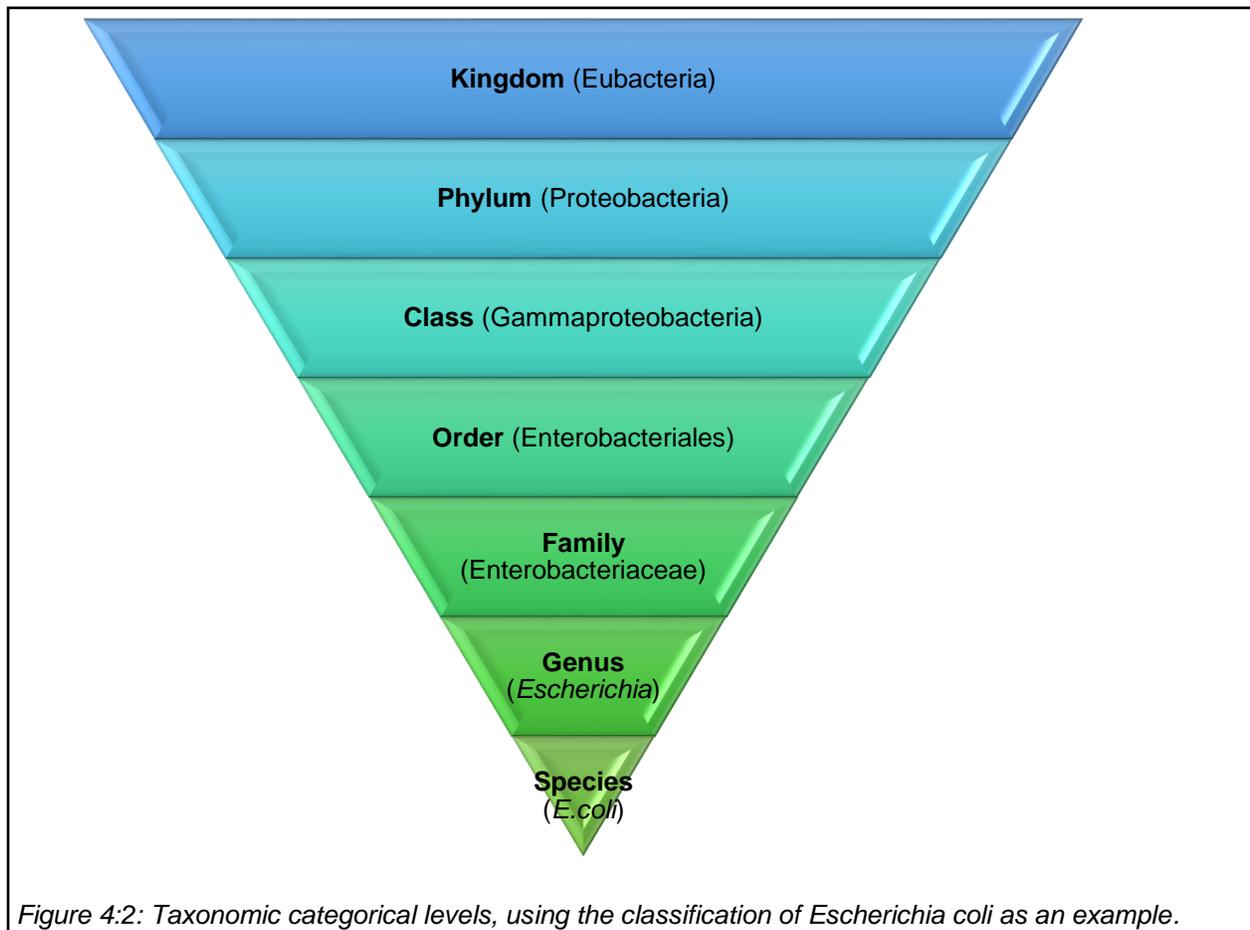
Diversity can also be described across a collection of samples. Beta diversity describes the relationship between two components, namely the local component (alpha; diversity in a single sample) and the regional component (gamma; diversity of a collection of samples) (Wagner *et al.*, 2018). Simply, beta diversity is used to describe the degree to which samples differ from each other and provide information on the microbial ecology that may not be apparent when looking at the composition of individual samples (Goodrich *et al.*, 2014).

To assess the diversity of the community structure across the collection of samples, beta diversity was measured using quantitative measures based on sequence abundance; Bray-Curtis (dissimilarity matrix) and weighted UniFrac, based on the relative abundance of operational taxonomic units (OTUs) at the family level. Dissimilarity results range between 0 and 1; where 0 means that the two site compositions are shared whereas a value of 1 indicates that 2 sites do not share the same composition. UniFrac is a distance metric that differs from Bray-Curtis in that it incorporates the relative relatedness of community members by including the phylogenetic distance between the observed microorganisms (Lozupone and Knight, 2005). Bray Curtis and UniFrac are visualized using Principal coordinate analysis (PCOA) plots and/ or heatmaps.

DeSeq2 package (v. 1.28.1) is a quantitative method that was used to determine differentially abundant taxa in this study. The analysis was done as described using the DeSeq2 with phyloseq tutorial. DeSeq2 uses tests such as negative binomial generalized linear models; dispersion estimates and logarithmic fold changes that allow for differential expression (Love, Huber and Anders, 2014). Raw (non-normalized) count data was used to quantify and determine differentially abundant taxa (at the family level) between the TB cases and ill control group at baseline and pre

and post TB treatment in the TB cases. Baseline and Ill control groups were used as the reference level.

Alpha and beta diversity analyses as well as other non-parametric statistical tests were conducted in R studio. The input files used for further analysis in R included the output files generated in QIIME, namely the biom file and tre file (phylogenetic tree file), and a mapping file that contained sample information. The R packages used included Phyloseq (v. 1.29.09) which is a tool used to import store, analyze and graphically display complex phylogenetic sequencing data that has been clustered into OTUs and Vegan (v.2.5-6) which was used for alpha and beta diversity analysis, ordination methods and tools for the analysis of dissimilarities. All diversity measures were based on the relative abundance of OTUs at the family level (Figure 4.2). A non-parametric multivariate analysis of variance test, measured using the Adonis test in R studio (Version 1.2.5001) (*adonis function | R Documentation, Anderson, 2001*), was used to determine the statistical significance of differences in beta-diversity..



## 4.3 Results

Seventy-seven respiratory samples (nasopharyngeal aspirates (NPA) and some induced sputum (IS)) obtained from 26 participants, at recruitment (baseline), 2 months and six months, were evaluated at the phylum and family level. The most abundant phyla observed across all samples regardless of timepoint were Proteobacteria, Firmicutes, Bacteroidetes and Fusobacteria, with a few samples having Tenericutes and Actinobacteria present (Chapter 3, Figure 3.8). Common familia across all samples included *Streptococcaceae*, *Pasteurellaceae*, *Moraxellaceae*, *Prevotellaceae*, *Veillonellaceae* and *Neisseriaceae*.

### 4.3.1 The respiratory microbiota in TB cases and ill controls

#### 4.3.1.1 Taxonomic classification of TB case and ill control samples at baseline

The baseline samples were categorized as bacteriologically confirmed and clinically diagnosed TB cases (unconfirmed TB) (as described in section 3.2.1) and unlikely TB cases (well-defined ill controls). All but one TB case sample (sample 82) and two ill control samples (samples 30 and 39) were NPAs; the others were IS samples.

No distinct microbial profiles were observed in the TB case or ill control groups (Figures 4.3 and 4.4). Twenty familia were identified in the case group (Figure 4.3). The most common familia included *Streptococcaceae* (84.6%, n=11/13), *Moraxellaceae* (61.5%, n=8/13), *Pasteurellaceae* (61.5%, n=8/13), *Veillonellaceae* (53.8%, n=7/13), *Prevotellaceae* (46.1%, n=6/13) and *Neisseriaceae* (46.1%, n=6/13). Familia identified in fewer samples and at a lower abundance included *Staphylococcaceae*, *Mycoplasmataceae*, *Fusobacteriaceae*, *Corynebacteriaceae*, *Enterobacteriaceae*, *Gemellaceae*, *Chlamydiaceae*, *Paraprevotellaceae*, *Leptotrichiaceae*, *Flavobacteriaceae*, *Enterobacteriaceae*, *Burkholderiaceae*, *Aerococcaceae* and *Mycobacteriaceae*. *Leptotrichiaceae* was identified in the sputum sample (Sample 82).

In the ill control group 19 familia were identified, most of which were seen in the TB case group (Figure 4.4), including *Streptococcaceae* (92.3%, n=12/13), *Pasteurellaceae* (77%, n=10/13), *Moraxellaceae* (77%, n=10/13), *Prevotellaceae* (69.2% n=9/13), *Veillonellaceae* (61.5%, n=8/13), *Neisseriaceae* (38.5%, n=5/13). *Fusobacteriaceae* (38.5%, n=5/13) and *Paraprevotellaceae* (38.5%, n=5/13) were more common in the ill control group and *Spirochaetaeae*, *Micrococcaceae*, *Lachnospiraceae* and *Bifidobacteriaceae* were only detected amongst the control samples. Other less common and less abundant familia observed in the control group samples included *Staphylococcaceae*, *Porphyomonadaceae*, *Mycoplasmataceae*, *Gemellaceae*, *Corynebacteriaceae*, and

*Aeorcoccaceae*, most of which were also detected in the TB case group. The two IS samples differed from the NPA samples by the presence of a single family, *Leptotrichiaceae*.

Similarly, no distinct microbial profiles were identified between the two case categories (bacteriologically confirmed or clinically diagnosed TB) (Figure 4.3). Families observed in the bacteriologically confirmed TB group which were not found in the clinically diagnosed TB group included *Staphylococcaceae*, *Enterobacteriaceae*, *Mycoplasmataceae* and *Mycobacteriaceae*. *Mycobacteriaceae* was only detected in a single sample in the bacteriologically confirmed category. In the clinically diagnosed TB group all the samples differed; two of the samples were dominated by *Moraxellaceae* and *Streptococcaceae*, while samples 76 and 82 were the most diverse. Other families observed only in this category included *Chlamydiaceae* and *Burkholderiaceae* which were not found in the bacteriologically confirmed category.

#### 4.3.1.2 Alpha and beta diversity between TB case and ill control samples at baseline

The alpha diversity appeared to be lower in the TB case group than in the ill control group at baseline (Figure 4.5), although this was not statistically significant for either the Shannon ( $p=0.24$ ) or Simpson ( $p=0.31$ ) indices. For both indices, the alpha diversity of the case group samples appeared to be more variable than that of the ill controls.

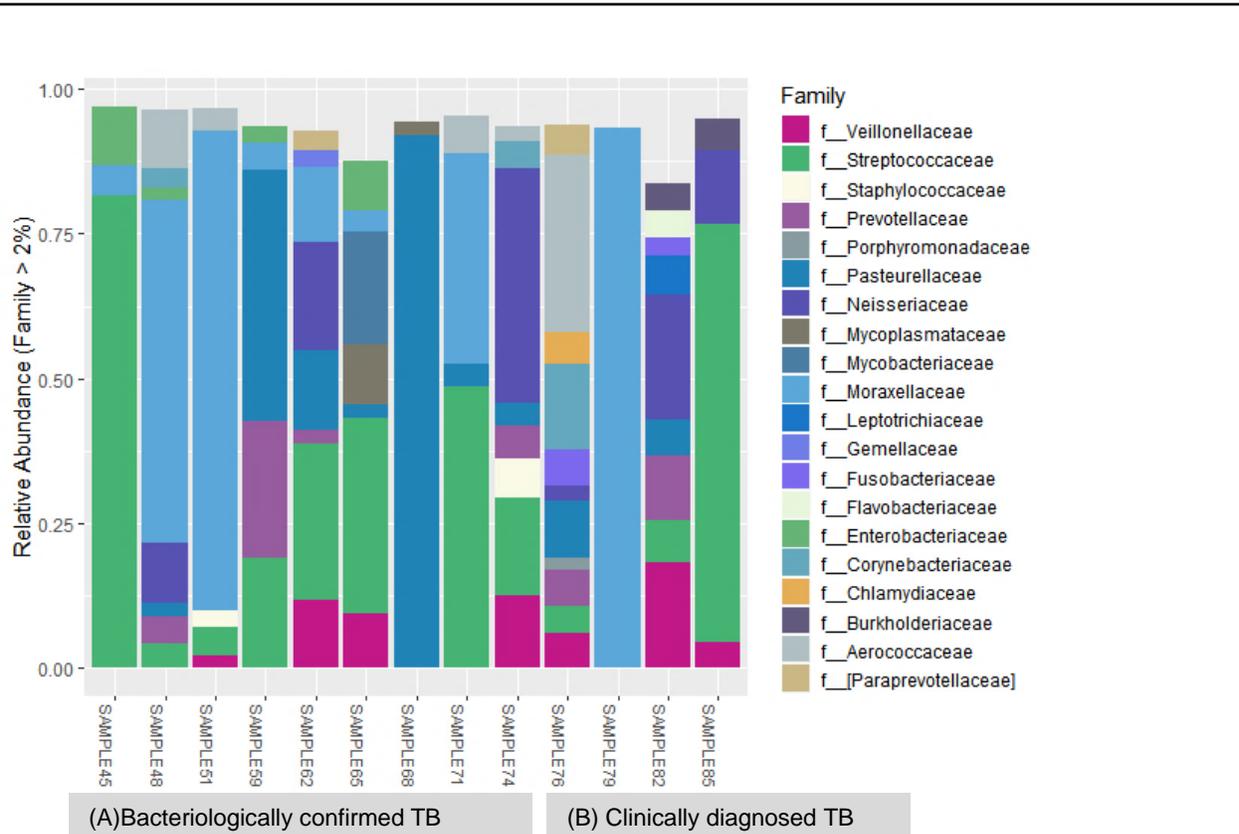


Figure 4:3: The relative abundance of familia observed in the TB case group, categorized into (A) Bacteriologically confirmed TB and (B) Clinically diagnosed TB at baseline. Only families with a minimum of 2 % relative abundance are represented.

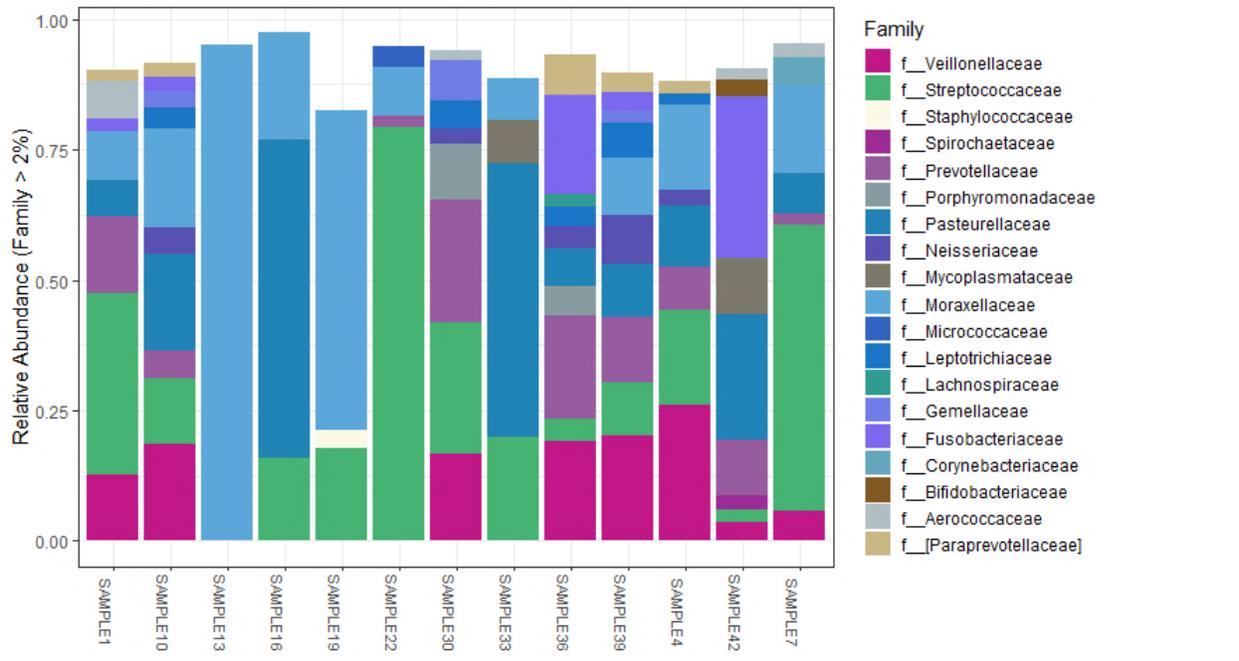


Figure 4:4: The relative abundance of familia observed in the unlikely TB control group (well-defined ill controls) at baseline. Only families with a minimum of 2 % relative abundance are represented.

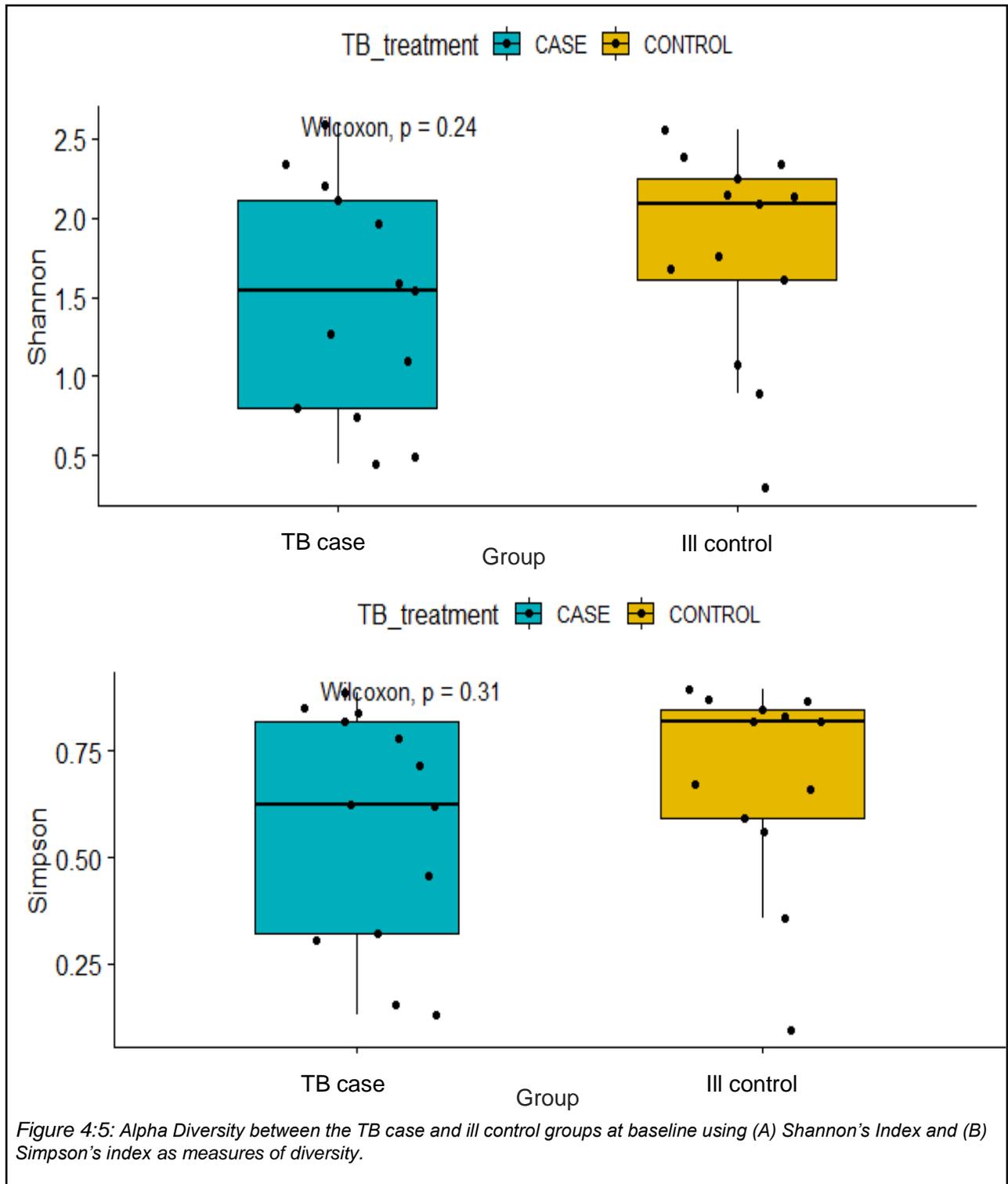


Figure 4:5: Alpha Diversity between the TB case and ill control groups at baseline using (A) Shannon's Index and (B) Simpson's index as measures of diversity.

The beta diversity between TB cases and ill controls showed no distinct clustering based on the Bray Curtis heatmap (Figure 4.6) or Bray Curtis PCOA analyses (Adonis test,  $R^2= 0.242$ ) (Figure 4.7), which suggests that there is a similar taxa composition in the groups. Similarly, no difference was seen between the bacteriologically confirmed and clinically diagnosed TB groups or ill controls at baseline (Adonis test,  $R= 0.802$ ) (Figure 4.8). Weighted UniFrac measures showed no distinct clustering between the TB case (bacteriologically confirmed and clinically diagnosed) and ill control groups at baseline (Adonis test,  $R^2=0.02$  and  $Pr (>F)= 0.72$ ) (Addendum 10). Based on differential abundance testing using DeSeq2, TB cases were compared to the ill control group. Based on this a 5- and 4-fold lower abundance of Pasteurallaceae and Prevotellaceae was observed in the TB cases compared to the ill controls at baseline (Figure 4.9). Suggesting that these two families were lower in abundance in the TB cases compared to the ill control group.



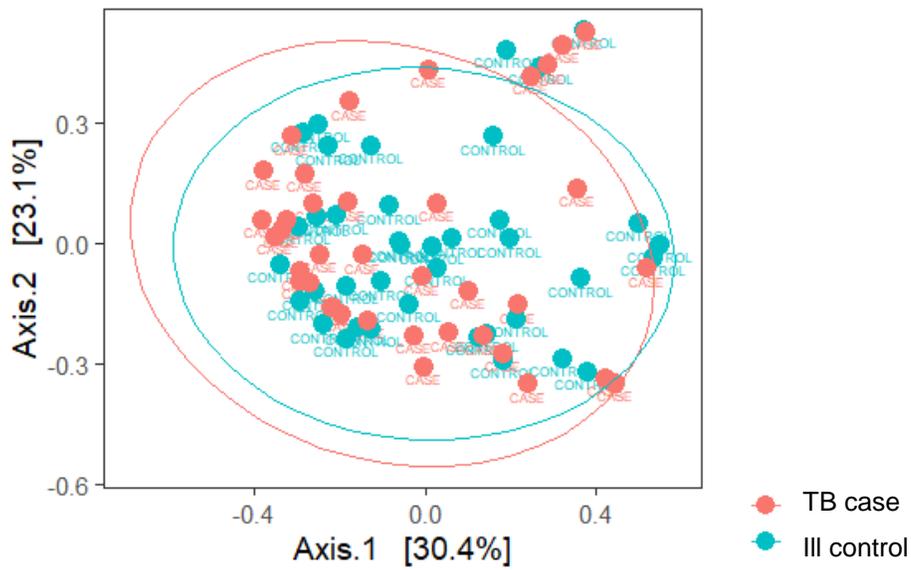


Figure 4:7: Principal coordinate analysis (Bray-Curtis) showing the difference between respiratory samples obtained from the TB case and ill-control groups at all time points. The PCOA plots were based on OTUs classified at the family level.

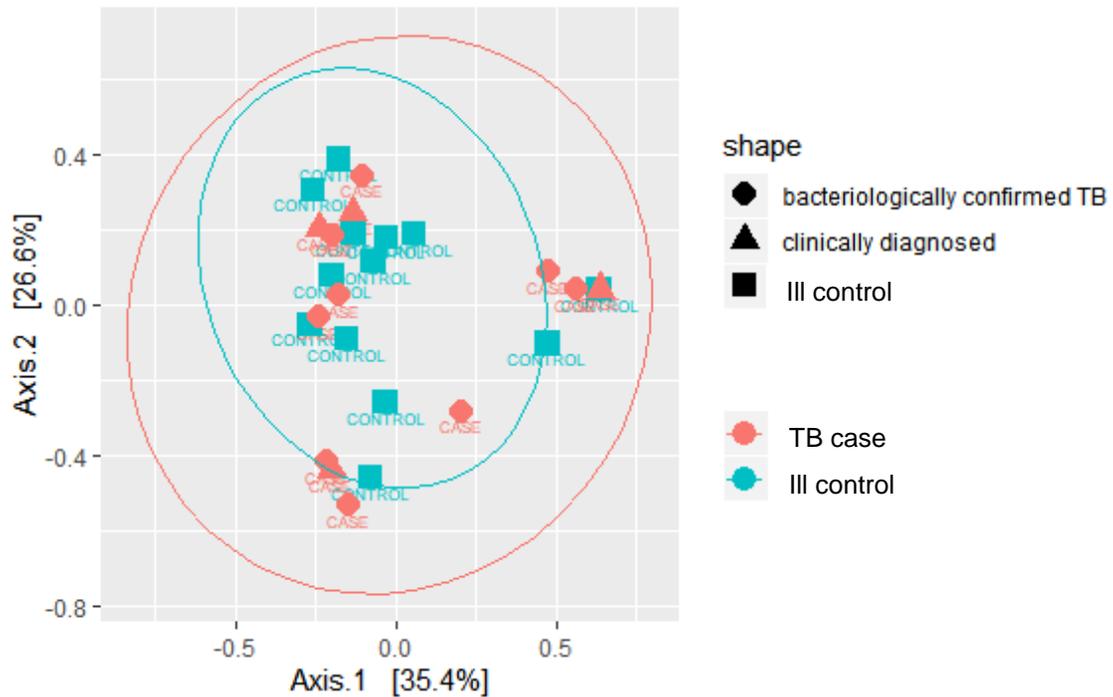
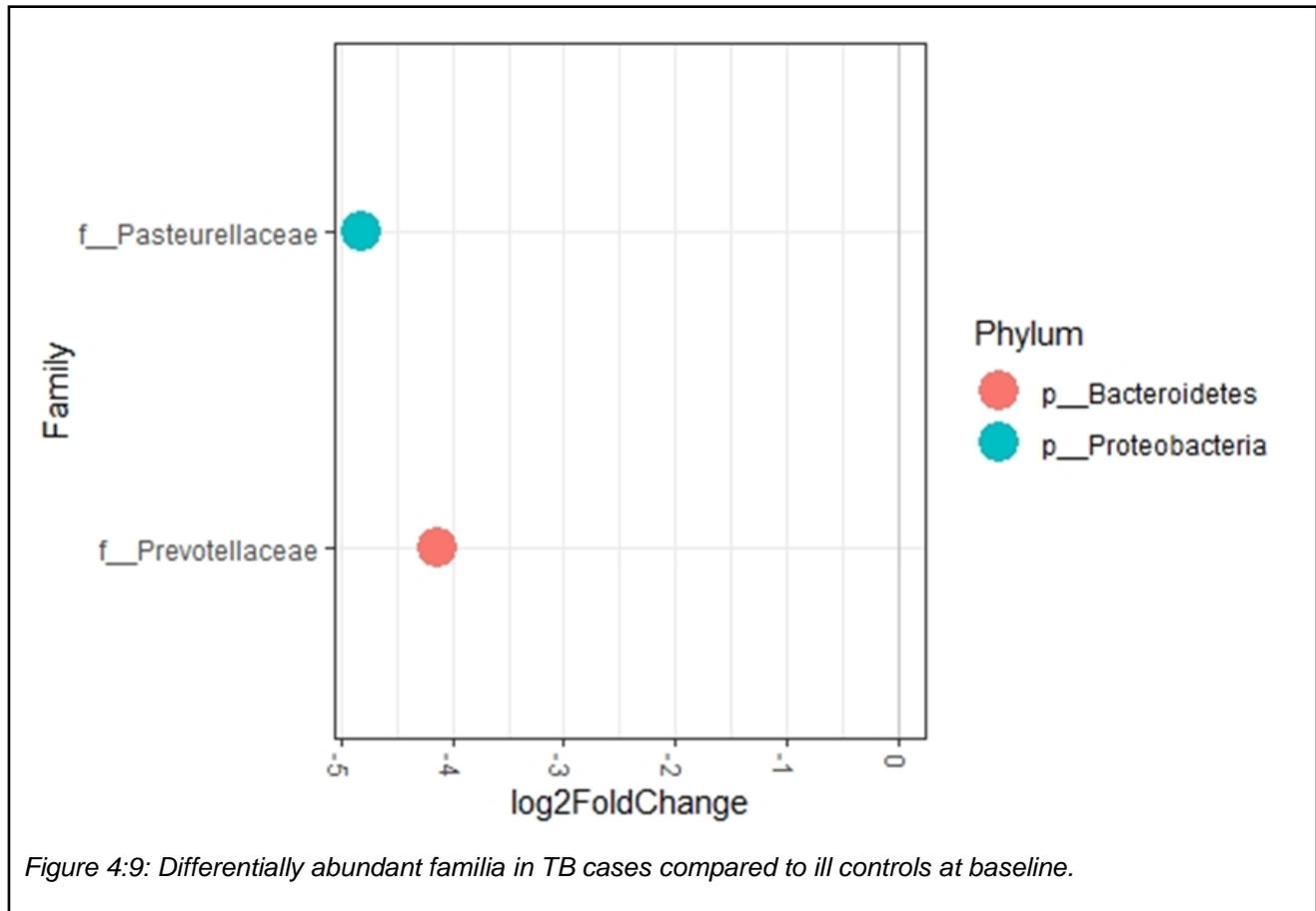


Figure 4:8: Principal coordinate analysis (Bray-Curtis) showing the difference between respiratory samples obtained from the TB case (bacteriologically confirmed TB or clinically diagnosed) and ill-control groups at baseline. The PCOA plots were based on OTUs classified at the family level.



### 4.3.2 The respiratory microbiota during TB treatment

Follow up samples were obtained at two time points, month 2 and month 6, to investigate the microbial profiles of participants during TB treatment. One participant (from the unconfirmed TB category) did not have a month 2 sample collected and another participant had a sample collected at month 1 (Sample 72) instead of month 2. Month 6 follow up samples were obtained from all participants. All but three month 2 samples (Sample 31, Sample 60 and Sample 83) and 2 month 6 samples (Sample 41 and Sample 84) were NPA samples. IS Samples 83 and 84 were obtained from the same participant.

#### 4.3.2.1 Taxonomic classification of TB cases at month 2 and month 6

No distinct microbial profiles were observed at month 2 or month 6 for the TB case group. Twenty-three familia were identified in the TB case samples at month 2 (Figure 4.10). Similarly to the baseline TB case samples, the most common familia observed were *Streptococcaceae* (83.3%, n= 10/12), *Veillonellaceae* (66.7%, n= 8/12), *Fusobacteriaceae* (58.3%, n= 7/12), *Prevotellaceae* (50%, n= 6/12), *Pasteurellaceae* (50%, n= 6/12) and *Neisseriaceae* (50%, n= 6/12). Less commonly observed familia included *Paraprevotellaceae*, *Mycoplasmataceae*, *Campylobacteriaceae* and *Gemellaceae*. In comparison to the baseline, there seemed to be a decrease in the presence of *Moraxellaceae*. The microbial profiles of Sample 46 and Sample 72 were completely different to the other month 2 samples. Sample 72 mainly consisted of *Carnobacteriaceae*, *Lachnospiraceae*, *Ruminococcaceae* and *Bacteroidaceae*, which were not present in any other the other samples. This could be indicative of the variability of the nasopharyngeal microbiota at different time points since this sample was a month 1 sample. Sample 46 was completely dominated by *Enterobacteriaceae*, which was not expected since it is usually present in low abundance in nasopharyngeal samples (Allen *et al.*, 2014; Chonmaitree *et al.*, 2017). This sample had a similar microbial profile to the PCR positive control and was thought to have been mistakenly spiked with the DNA from the PCR positive control. However, the possibility of it being the positive control was ruled out after other sequencing reads were identified in the sample that were not identified in the PCR positive control; though not observed in the figures as a result of the threshold (2%) that was used.

Eighteen familia were identified in the TB case samples at month 6 (Figure 4.11). The month 6 samples were similar to the month 2 samples with the exception of an increase in *Paraprevotellaceae* (38.5%, n= 5/13). The most common familia were the same; *Streptococcaceae* (92.3%, n= 12/13), *Pasteurellaceae* (92.3%, n= 12/13),

*Veillonellaceae*, *Neisseriaceae* (76.9%, n=10/13), *Prevotellaceae* (61.5%, n= 8/13) and *Fusobacteriaceae* (46.2%, n= 6/13). There was a decrease in the relative abundance of *Prevotellaceae* in the month 6 samples in comparison to month 2. Sample 78 was completely dominated by *Pasteurellaceae*.

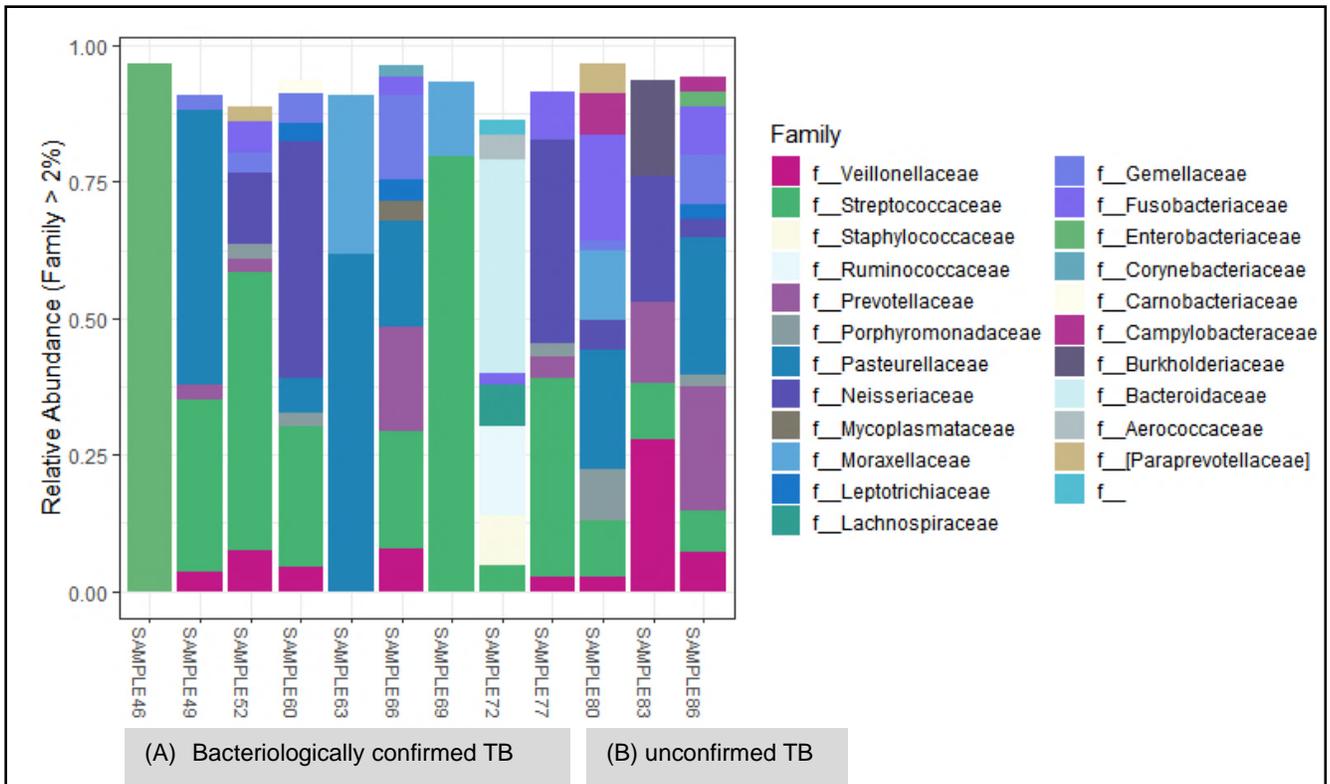


Figure 4:10: The relative abundance of familia observed in samples obtained at month 2 from the TB case group, categorized into (A) Bacteriologically confirmed TB and (B) Clinically diagnosed TB Only families with a minimum of 2 % relative abundance are represented.

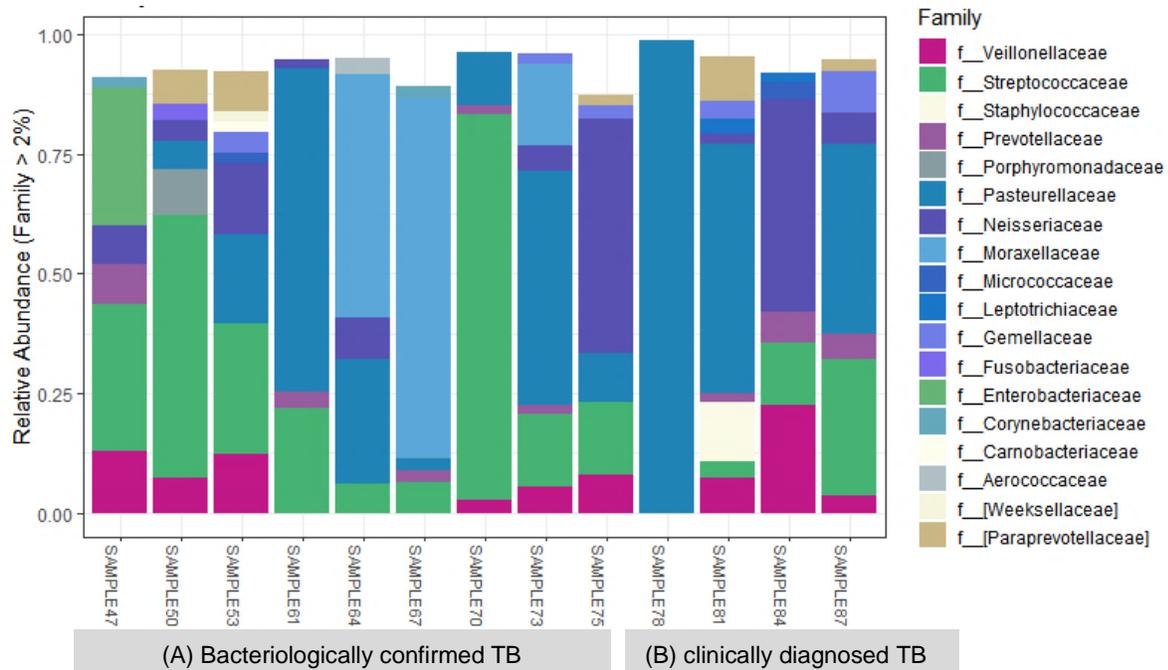


Figure 4:11: The relative abundance of familia observed in samples obtained at month 6 from the TB case group, categorized into (A) Bacteriologically confirmed TB and (B) clinically diagnosed TB Only families with a minimum of 2 % relative abundance are represented.

#### 4.3.2.2 Taxonomic classification of well-defined ill controls at Month 2 and month 6

Month 2 and month 6 samples from the ill control group, who were not on TB treatment, were also evaluated, to control for normal microbiota variation over time (Figure 4.12 and Figure 4.13). The microbial profiles appeared to be similar to the baseline samples with regard to the common familia identified: *Veillonellaceae*, *Streptococcaceae*, *Neisseriaceae*, *Pasteurellaceae*, *Moraxellaceae* and *Prevotellaceae*. There were fewer familia observed at month 2 in comparison to month 6 where 22 familia were identified, with the addition of *Cardiobacteriaceae*, *Weeksellaceae* and *Tisseriellaceae*. There was an increase in the relative abundance of *Neisseriaceae* from baseline to month 2, and an increase in the presence of *Moraxellaceae* and *Paraprevotellaceae* and a decrease in the presence of *Prevotellaceae* from month 2 to month 6. familia identified in fewer samples at month 2 and month 6 and at a lower relative abundance were *Corynebacteriaceae*, *Carnobacteriaceae*, and *Aerococcaceae*, with the addition of *Lachnospiraceae* and *Enterobacteriaceae* in the month 6 samples.

#### 4.3.2.3 Alpha and beta diversity in response to TB treatment

The alpha diversity appeared lower at baseline compared to month 2 and month 6 in the TB case samples, although this was not statistically significantly for either the Shannon (Baseline vs month 2,  $p=0.51$  and Baseline vs month 6,  $p=0.83$ ) or Simpson (Baseline vs month 2,  $p=0.32$  and Baseline vs month 6,  $p=0.83$ ) indices (Figure 4.14). The interquartile ranges were also more condensed at the later time points which could suggest that after treatment the number of taxa was more evenly spread. In the ill control group, the alpha diversity at month 2 was slightly decreased in comparison to baseline and month 6 for both the Shannon (Baseline vs month 2,  $p=0.82$  and Baseline vs month 6,  $p=0.78$ ) and Simpson (Baseline vs month 2,  $p=0.82$  and Baseline vs month 6,  $p=0.55$ ) indices, although this was not statistically significant (Figure 4.14).

No distinct clustering was observed for the TB case and ill control groups over time based on the Bray-Curtis heatmap (Figure 4.6) or Bray-Curtis and weighted Unifrac PCOA analyses (Adonis test,  $R^2=0.123$  and Adonis test  $R^2=0.039$ ,  $PR>F=0.121$ , respectively) (Figure 4.15 and Addendum 10), suggesting similar taxa composition in samples obtained at the different time points, and between TB cases and ill controls.

On the other hand, differentially abundant familia were observed in month 6 TB cases with fold changes of 4 to 8 across the familia; *Veillonellaceae*, *Staphylococcaceae*, *Prevotellaceae*, *Neisseriaceae*, *Enterobacteriaceae* and *Aerococcaceae* (Figure 4.16).

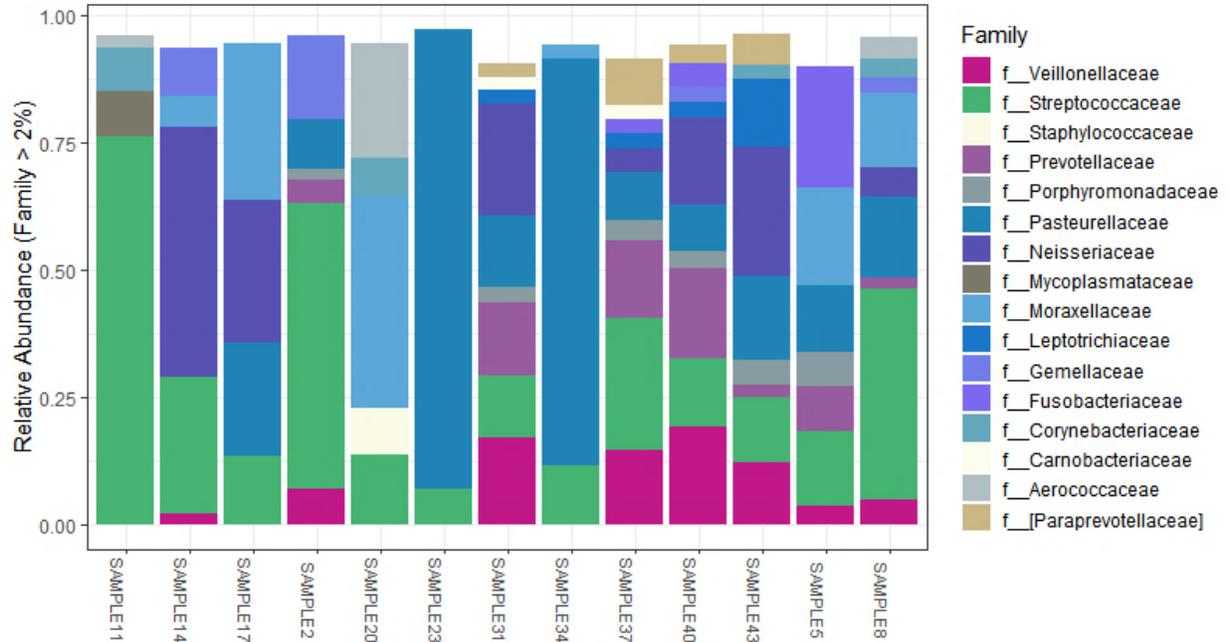


Figure 4:12: The relative abundance of familia observed in the unlikely TB control group (well-defined ill controls) at month 2. Only families with a minimum of 2 % relative abundance are represented.

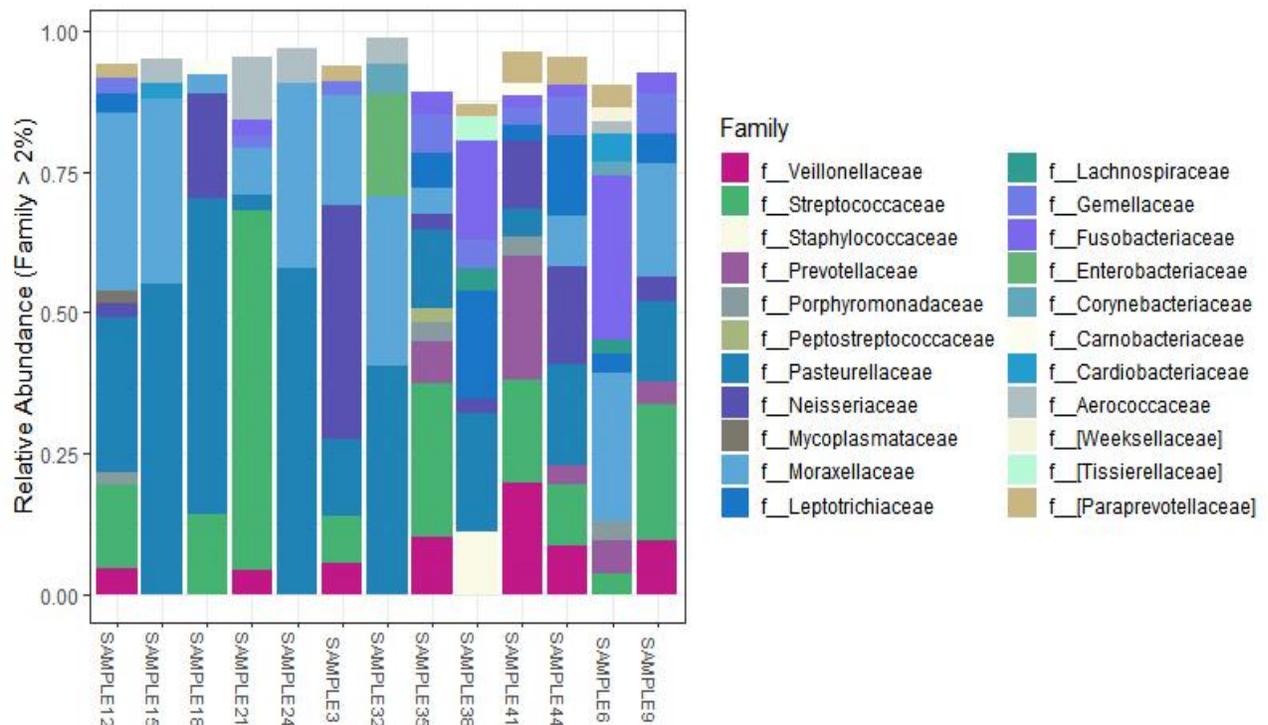


Figure 4:13: The relative abundance of familia observed in the unlikely TB control group (well-defined ill controls) at month 6. Only families with a minimum of 2 % relative abundance are represented.

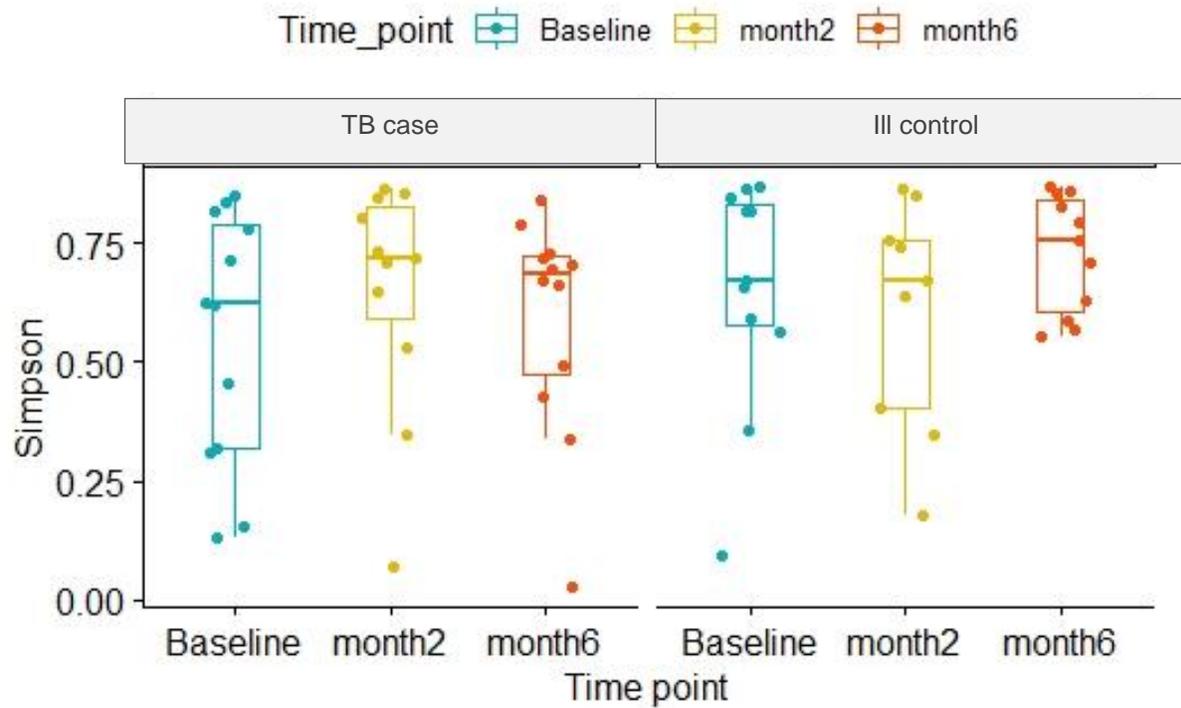
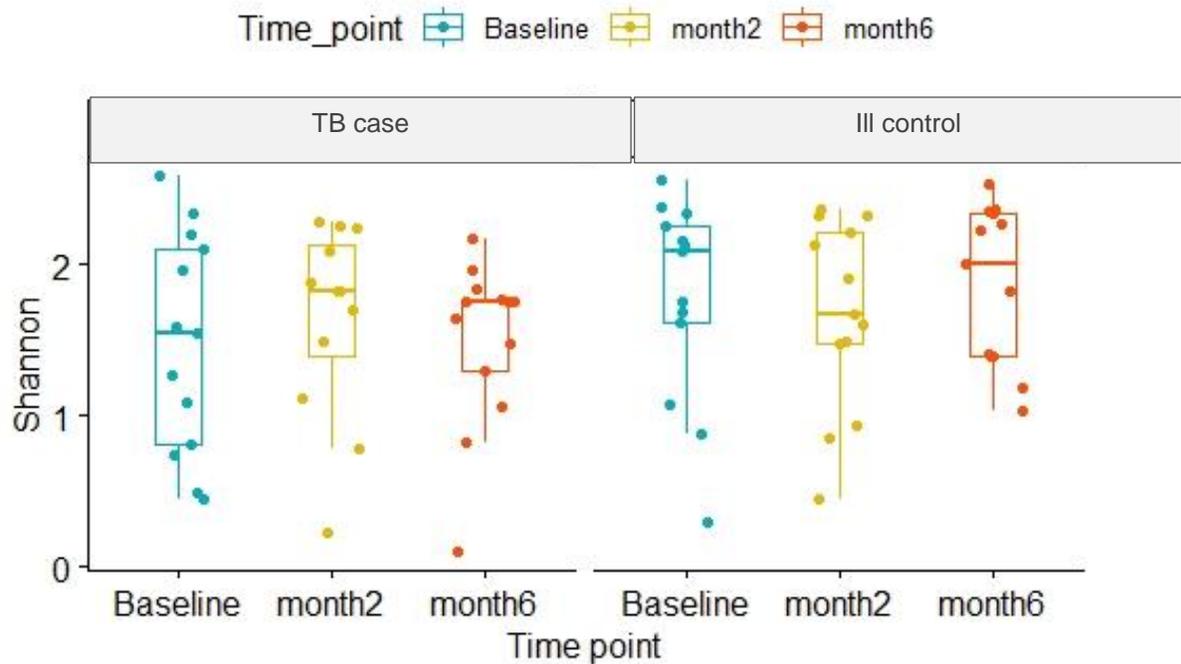


Figure 4:14: Alpha Diversity measures in TB case and ill-control groups during treatment based on the (A) Shannon and (B) Simpson Indices.

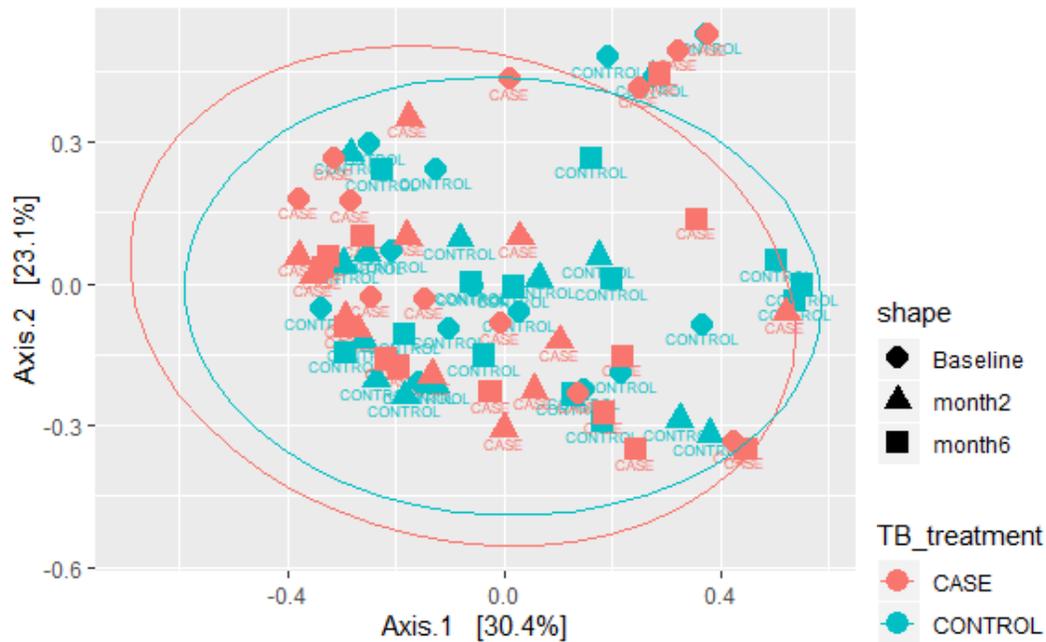


Figure 4:15: Principal coordinate analysis (Bray-Curtis) showing no differences between respiratory samples obtained from the TB case and ill-control group during TB treatment. The PCOA plots were based on OTUs classified at the family level.

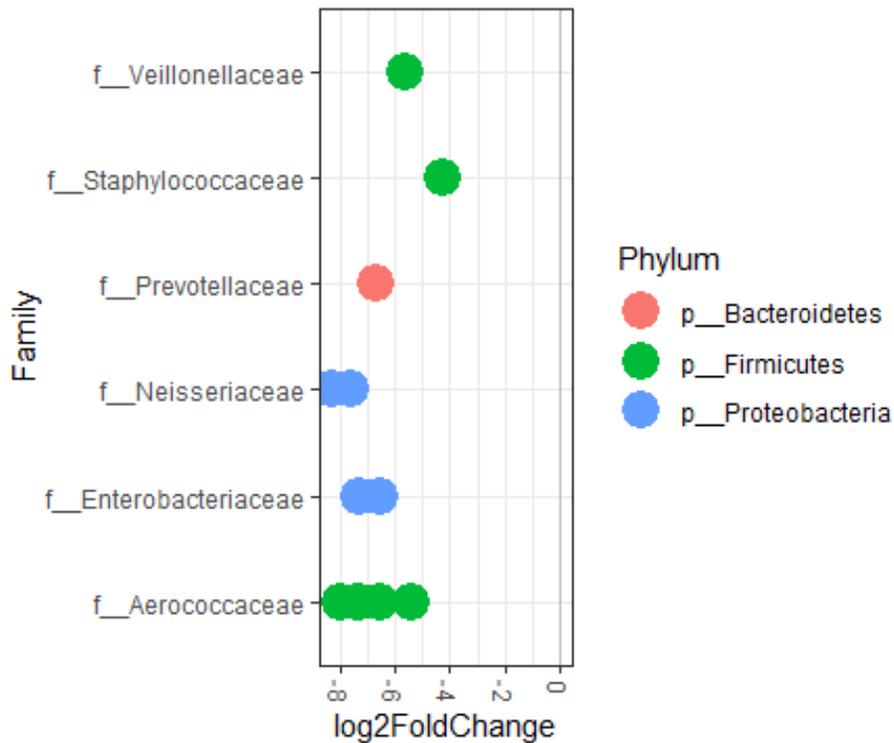


Figure 4.16: Differentially abundant taxa in month 6 TB cases

### 4.3.3 Baseline TB cases compared to month 6 ill controls as a proxy for healthy microbiome (microbiota)

All participants in the study were ill at recruitment (baseline) and were exposed to antibiotics. However, at month 6 in the ill control group (unlikely TB group) participants could be considered “healthy” since their antibiotic treatment would have been completed and their “microbiomes” possibly restored. Therefore, the baseline TB cases were compared to the month 6 “recovered” ill control samples to identify any potential differences between the microbiota of TB cases at baseline and a “healthy” microbiota (Figure 4.3 and Figure 4.13).

Almost all of the month 6 ill controls (84.6%, n=11/13) had *Moraxellaceae* present in comparison to only 61.5% (n=8/13) in the baseline TB cases (Figure 4.13 and Figure 4.3, respectively). *Neisseriaceae*, *Paraprevotellaceae*, *Fusobacteriaceae* were more commonly observed in the month 6 ill controls than in the baseline TB cases, while *Streptococcaceae*, *Aerococcaceae*, *Enterococcaceae* and *Prevotellaceae* were more commonly observed in the baseline TB case (Table 4.2).

Table 4.2: Common familia present in the baseline TB cases and month 6 ill controls.

Family	Baseline TB cases (n=13)	Month 6 ill controls (n=13)
<i>Streptococcaceae</i>	84.6% (11)	69.2% (9)
<i>Neisseriaceae</i> ,	46.1% (6)	61.5% (8)
<i>Fusobacteriaceae</i> ,	15.4% (2)	53.8% (7)
<i>Veillonellaceae</i>	53.8% (7)	53.8% (7)
<i>Paraprevotellaceae</i>	15.4% (2)	46.1% (6)
<i>Prevotellaceae</i>	46.1% (6)	38.5% (5)
<i>Aerococcaceae</i>	38.5% (5)	30.8% (4)
<i>Enterococcaceae</i>	30.8% (4)	7.7% (1)
<i>Corynebacteriaceae</i>	7.7% (1)	7.7% (1)

The alpha diversity appeared higher in the month 6 ill controls in comparison to the baseline TB cases, although this was not statistically significant for either the Shannon ( $p=0.15$ ) or Simpson indices ( $p=0.11$ ) (Figure 4.17). For both indices, the alpha diversity of the baseline case samples appeared to be more variable than that of the month 6 ill control samples. No distinct separation between the groups was observed using either beta diversity measure (Bray-Curtis and weighted UniFrac) (Figures 4.6 and Addendum 10).

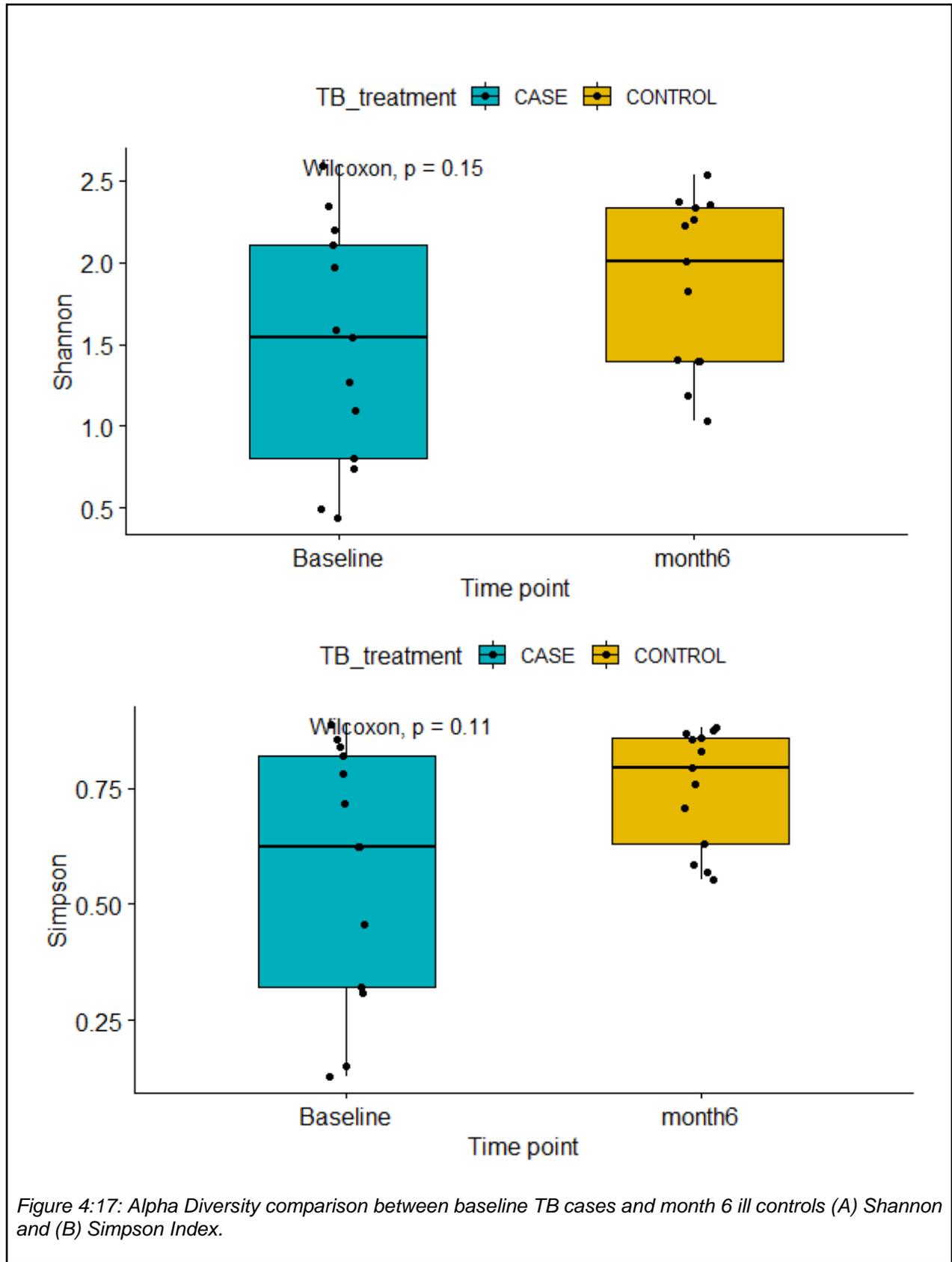


Figure 4:17: Alpha Diversity comparison between baseline TB cases and month 6 ill controls (A) Shannon and (B) Simpson Index.

## 4.4 Discussion

In Chapter 4, we presented the results of a pilot study which aimed to describe the differences in the respiratory microbiota in children with and without TB, and the impact of TB treatment on the microbiota. The nasopharyngeal microbiota was investigated in children with suspected PTB who were categorized into bacteriologically confirmed PTB, clinically diagnosed PTB and unlikely PTB groups. Although the sequencing quality was suboptimal for accurately identifying bacteria at the genus level, we were able to identify bacterial phyla and familia in all samples based on the clustering of forward read sequences as discussed in Chapter 3. We did not identify a distinct microbial profile within the TB case group (bacteriologically confirmed/ clinically diagnosed PTB) or between the TB case and ill control group at baseline, month 2 and month 6. In addition, there was no distinct microbial profile between TB cases after 6 months of TB treatment and ill controls that did not receive TB treatment.

The most abundant phyla identified in all samples and diagnostic categories were *Proteobacteria*, *Firmicutes*, *Bacteroidetes*, *Fusobacteria* and *Actinobacteria*, to a lesser extent *Tenericutes* (only identified in a few samples). The above-mentioned phyla (with the exception of *Tenericutes*) are commonly identified in human microbiome studies, including those studies investigating the nasopharynx. These phyla have been reported in different relative proportions in both infants and children during healthy and diseased states (Bogaert *et al.*, 2011; Teo, Mok, Pham, Kusel, Serralha, Troy, Barbara J. Holt, *et al.*, 2015). On the other hand, the presence of *Tenericutes* is less common, but similar to another study (Stearns *et al.*, 2015) which identified it in nasopharyngeal samples from healthy participants (<19 months of age). We identified it in ill participants of the same age (median age 19 months for participants included in the microbiome section) which may suggest that its presence in the samples is age related.

### The respiratory microbiota in TB cases and ill controls

The most commonly identified familia in both the baseline TB case and ill control groups were *Streptococcaceae*, *Veillonellaceae*, *Prevotellaceae*, *Moraxellaceae*, *Pasteurellaceae* and *Neisseriaceae*. There were no statistically significant differences in alpha or beta diversity (Bray Curtis or weighted UniFrac) between the TB cases and ill controls at baseline or between the bacteriologically confirmed and clinically diagnosed TB groups. This may be due to the fact that all of the participants suffered from a respiratory illness and had likely been exposed to antibiotics prior to recruitment for the study; therefore, they may have had similar microbial profiles. The results may also be influenced by the higher taxonomic rank being used for comparisons, which is less specific as multiple genera and species can belong within a higher rank, or by the small

sample size which may not be truly representative of the different populations. Conversely, differential abundance testing using DeSeq2 suggests that there are differences between the TB case and ill control groups at baseline, namely as a result of a lower abundance of *Pasteurellaceae* and *Prevotellaceae* in TB cases. Differential abundance is useful for determining overall shifts in abundance of microbial compositions between two conditions e.g. healthy vs diseased/treated vs untreated (Banerjee *et al.*, 2019). This method was particularly useful because at a taxonomic level (microbial profile) and using standard diversity measures: alpha (Shannon and Simpson indices) and beta analysis (Bray Curtis and weighted uniFrac) suggested that “no differences” were apparent among these groups, which may have brought about an inaccurate interpretation of the outcome of this study. The use of this method also highlights the importance of including quantitative analysis to assist in the understanding of microbiome studies when comparing different conditions.

The familia identified include genera that are typically found in the nasopharynx, but can exist in different patterns or associations with each other, depending on age, the presence of specific viruses (Allen *et al.*, 2014; Li *et al.*, 2017), early succession patterns (Biesbroek, Tsivtsivadze, Elisabeth A. M. Sanders, *et al.*, 2014) and antibiotic use (Teo, Mok, Pham, Kusel, Serralha, Troy, Barbara J. Holt, *et al.*, 2015). In healthy individuals, certain genera have been described to be dominant depending on age; for example in infants ( $\leq 6$  months) the nasopharynx is composed of *Streptococcus*, *Corynebacterium*, *Staphylococcus* and *Dolosigranulum* (Rosas-Salazar *et al.*, 2016), while in children (18 months of age) a different microbial pattern is typically seen: *Moraxella*, *Haemophilus*, *Streptococcus* and *Flavobacteria* (Bogaert *et al.*, 2011). During acute respiratory rhinovirus infection a decrease in the relative abundance of genera such as *Streptococcus*, *Moraxella*, *Corynebacterium* and *Haemophilus*, with the addition of *Dolosigranulum* was observed (Rosas-Salazar *et al.*, 2016). Contrary to this, another study found that the enrichment of *Moraxella* and a decrease in the relative abundance of taxa such as *Faecabacterium* and *Alkkermansia* spp could be associated with a history of acute sinusitis (Santee *et al.*, 2016). We found a high abundance of the family *Streptococcaceae* to which the genera *Streptococcus* belongs; it is likely that the main genus in the samples is *Streptococcus* since it is usually identified in nasopharyngeal samples. Assuming that this is the case in our study, the presence of *Streptococcus* could be associated with age; Biesbroek *et al* (2014) found that it has a high abundance between the ages of 12 and 24 months. The presence of this genus is also associated with an unstable respiratory microbiome (Biesbroek, Tsivtsivadze, Elisabeth A. M. Sanders, *et al.*, 2014) and since all of the participants were ill and likely on antibiotic treatment, this could be the case in our study. More so since those genera that confer or are associated with

a stable respiratory microbiome were not observed in high abundance in our study (*Corynebacteriaceae*-genus *Corynebacterium* and *Carnobacteriaceae*-genus *Dolosigranulum*). *Staphylococcaceae*, *Enterobacteriaceae*, *Mycoplasmataceae* and *Mycobacteriaceae* were identified in a few baseline samples in the bacteriologically confirmed category, but not seen in the clinically diagnosed category. The significance of these bacteria was not evaluated because they were only identified in a few samples. However, it is interesting in that of those that were bacteriologically confirmed to have TB one participant had *Mycobacteriaceae* detected in the NPA sample which could suggest recent exposure to *Mycobacteriaceae* and could be further investigated in future studies. Also, the same participant also had *Mycoplasmataceae* and *Enterobacteriaceae* detected which may suggest susceptibility to the acquisition of other pathogenic bacteria, however this would have to be evaluated on a larger group to fully understand the presence of these pathogens in the samples. Another reason for the presence of unexpected microorganisms in the samples (or controls) could be an indication of (1) index hopping which is the misassignment of reads in multiplexed libraries that occurs when the DNA index of one sample is switched with the index of another sample. (2) Contamination is likely in low biomass samples as even a small contaminant signal from laboratory reagents or environmental sources such as air within the facility in which the samples are prepared (in this study samples were prepared in a general laboratory) can over power the intrinsic signal in a sample. (3) Spillover and micro droplets can also be contributing factors to the presence of unexpected OTU's in samples (or control) during library preparation.

Only one study has described the presence of other respiratory microbiota in children with and without TB in children, and was limited to the detection of specific microorganisms, including viruses (and fungi). The study showed that there was no clear separation between those with TB and those without, but that the microbial profile associated with the TB group was mainly composed of viruses (Dube *et al.*, 2016), which is consistent with our finding that the bacterial microbiota did not vary significantly between TB cases and ill controls. Other studies that have described the respiratory microbiome in relation to TB have investigated the adult lung; and although a distinct microbial profile was not identified, differences in diversity (Cui *et al.*, 2012; Botero *et al.*, 2014) and relative abundance of phyla and genera between the healthy and PTB cases were observed (Eshetie and Van Soolingen, 2019).

A healthy control group was not included in this study, therefore the baseline microbiota of TB cases was compared to the asymptomatic 6 month samples from participants in the ill control group; which were considered to represent “healthy” microbiota, as the original respiratory illness

would have been treated, and the microbiota could have recovered by this point. This was merely done to describe potential taxonomic differences between the two groups. The month 6 ill controls mainly differed from the TB cases by being dominated by the presence of *Pasteurellaceae*, *Streptococcaceae* and *Moraxellaceae*, where the TB cases had more *Streptococcaceae* present, which could support the hypothesis that those with TB had an unstable microbiome in comparison to those without TB. Comparing the alpha diversity (Shannon and Simpson) indices between the TB cases and ill control group at baseline to the TB cases and month 6 ill control group, it was observed that the p values for the Shannon ( $p= 0.24$  versus  $0.15$ ) and Simpson ( $p= 0.31$  versus  $0.11$ ) indices decreased. This suggests that there may be a difference in diversity between the microbiota of TB cases at diagnosis and healthy controls. This warrants further investigation of the respiratory microbiota in children but with the inclusion of a healthy control group; and a larger population size.

#### The respiratory microbiota during TB treatment

The microbial profiles of the TB cases at baseline and during TB treatment were similar with regard to the common familia observed in the samples, *Streptococcaceae*, *Veillonellaceae*, *Pasteurellaceae*, *Prevotellaceae* and *Neisseriaceae*. However, at a taxonomic level it appeared that at month 2 and month 6 there was increase in the presence of *Fusobacteriaceae* which was identified in only a few samples and in low abundance at baseline. From baseline to month 2 there was a decrease in the presence and abundance of *Moraxellaceae* and *Pasteurellaceae* which then seemed to increase in abundance at month 6. On the other hand, in the month 6 samples there seemed to be an increase in the presence of *Paraprevotellaceae* and a decrease in the relative abundance of *Prevotellaceae* in comparison to the month 2 samples. Namasivayam *et al* (2018) reviewed studies that investigated the effects of first line TB treatment on the intestinal microbiome in humans and mouse models. A decrease in *Paraprevotella* and *Pasteurella* in mouse models and *Prevotella* in humans, in addition to an increase in *Fusobacterium* and *Prevotella* in humans (Namasivayam *et al.*, 2018) was noted. Our findings are similar in that familia associated with the above-mentioned genera increased (*Fusobacteriaceae*) and decreased (*Pasteurellaceae*) during TB treatment. We also observed an increase in the presence of familia *Paraprevotellaceae* and *Prevotellaceae* where they observed a decrease in the associated genera, *Paraprevotellilla* and *Prevotella*. This could be attributed to the different effects of TB treatment on different microbial niches or the variability of the nasopharyngeal microbiome in itself.

In this study, it appears that TB treatment did not seem to influence the microbial composition, or alpha and beta diversity in the TB case group over time, since no major differences were observed. This might not be surprising since most of the first line TB drugs (INH, PZA and EMB) specifically target *M.tb*; although RIF has a broad spectrum of activity. Also, perhaps the concentration of the drugs in the infants'/childrens' body was not high enough to have a substantial effect on the overall diversity. Some studies have previously reported lower TB drug concentrations in children, like a lower concentration of PZA in children under the age of 5 years (Graham *et al.*, 2006) and lower concentrations of rifampin (RIF), PZA and EMB when co-infected with HIV (Ramachandran *et al.*, 2016; Antwi *et al.*, 2017). This should also be accounted for in future studies since treatment is adjusted throughout therapy as doses are adjusted according to weight (South African National Department of Health, 2013). Nonetheless, our findings are similar to those of Wiperman *et al* (2017) and Namasivayam *et al* (2017) who showed that TB treatment had a minimal effect on diversity in intestinal microbiota of adults and mice.

However, differentially abundant familia were observed when comparing baseline TB cases to month 6 TB cases (pre and post TB treatment) which indicates that TB treatment may not have an overall effect on the diversity of the niche being investigated but may influence the richness and or abundance of the identified taxa in the niche. This was similar to what was found by Wiperman *et al* (2017) in an intestinal microbiota study where the overall diversity was not affected but certain bacterial taxa was reduced. Determining what drove the reduction (combined broad and narrow spectrum activity of TB treatment or solely broad spectrum activity) of specific taxa in this study was not investigated but would be an interesting concept to investigate in future especially since the findings of this study may have been influenced by the limitations of the study.

Factors to consider when interpreting effects of treatment on the microbiota should be factored in. In this study it is recognized that the nasopharyngeal microbiome is constantly being exposed to and is easily accessible for the acquisition of other pathogens, therefore a more accurate measure of the effect of TB treatment should rule out other factors that could contribute to changes, such as age, the acquisition of other pathogens (bacterial, viral or fungal), temporal variation or immune response (García-Rodríguez and Fresnadillo Martínez, 2002), especially if significant differences in diversity are found prior to treatment. Further research should be conducted on a larger scale while taking into account these factors, to fully assess the effect of TB treatment in the respiratory microbiome in children. Additional sampling after completion of treatment would also be of value, as alterations to the microbiome can persist after cessation of TB treatment (Namasivayam *et al.*, 2017; Wiperman *et al.*, 2017).

As previously mentioned, all participants were ill and even after 6 months the participants in the ill control group still had a microbial profile that resembled an unstable microbiome (i.e. high abundance of Streptococcaceae observed). This could imply that at this point their microbiomes had not yet fully improved since those organisms that are associated with a more stable microbiome were not commonly observed (Corynebacteriaceae: genus *Corynebacterium* and Carnobacteriaceae: genus *Dolosigranulum*). This could allow for interventional therapy to be implemented, such as probiotics to assist in the restoration of microbiota in both groups. However, a challenge with the implementation of probiotics include determining the best approach i.e. intranasal application or oral supplements and determining the bacteria that would be most effective for restoring the microbiota while counteracting pathogens. This warrants the further investigation into developing appropriate probiotics that are potentially niche specific.

Although the alpha and beta diversity measures showed no significant differences in the control group at the different time points, a number of changes in the familia were observed over time, including an apparent increase in the relative abundance of *Neisseriaceae* at month 2, and an increase in the presence of *Moraxellaceae* and *Paraprevotellaceae* and a decrease in the presence of *Prevotellaceae* from month 2 to month 6. *Cardiobacteriaceae*, *Weeksellaceae* and *Tisseriellaceae* were also unique to the month 6 ill control samples. The reasons for changes are not clear, but some of these might be attributed to the initial treatment prescribed for the respiratory illness at baseline, and the recovery of the microbiota following treatment. However, this also demonstrates that the nasopharyngeal microbiome can change over time, in the absence of continuous antibiotic therapy. Also, it would be important to determine whether certain bacteria, viruses or other microbes could have contributed to the increase and decrease of certain groups of bacteria.

### Limitations

The family level was used to describe differences in the microbial profiles between the TB case and ill control groups and to measure the diversity in and between samples in the case and ill control group at baseline and during TB treatment. Multiple bacterial genera and species can be clustered within a higher taxonomic rank and therefore clustering at family level does not provide a high level of detail of the range of bacteria present in the respiratory samples. Therefore, although more accurate, this level of classification was less specific and may have prevented the identification of significant differences between groups. Comparing the findings in this study to previously reported studies is difficult as taxonomic profiles and diversity measures can be influenced by sequencing quality, reference databases and library preparations or target regions

of the 16S rRNA gene and different sample types. Furthermore, there is not much applicable data available to compare it to for this age group.

All participants were ill at the time of recruitment and may have received other antibiotics prior to baseline sample collection, as only participants that had received TB treatment for > 2 days in the previous two weeks were excluded. This could be a confounding factor, as antibiotics have previously been identified as an external interference that can contribute to dysbiosis. The ill control group may also have been exposed to additional antibiotics following recruitment, to treat their respiratory infections. Since a healthy control group was not included in the study it cannot be determined whether the microbial profile is specific to the TB case or ill control group or more related to respiratory disease in general. It is therefore necessary to consider the inclusion of age matched healthy participants for future studies.

We mainly focused on describing taxa that were present in children with suspected PTB and no associations were made between the presence of these taxa in relation to TB disease. In future studies it would be useful to try and associate certain taxa with PTB to determine whether it plays a role in disease pathogenesis and severity. Factors that could contribute changes in the microbiome such as mode of delivery, seasonal changes, viruses, siblings and breastfeeding were not taken into account during analysis and therefore limits the understanding of the presence of the familia identified at different time points or even at baseline. It would therefore be crucial to incorporate this in future studies since it is known to impact the nasopharyngeal microbiome. However, this would also require increased sample size, to enable statistical evaluation.

Lastly, a small proportion (8/77) of samples were IS samples. This did not appear to influence the results as the IS samples only differed from NPA samples by the relative abundance of familia and the presence of *Leptotrichiaceae*. The genus *Leptotrichia* has previously been identified in sputum samples obtained from adult TB cases and controls (Cheung *et al.*, 2013) suggesting that it may also be sputum-associated in our study.

## 4.5 Conclusion

TB remains a global health concern especially in the pediatric population group and there are gaps in our understanding of this disease in this group. The improvement of sequencing technologies and the discovery of the human microbiome provided insight into healthy and diseased states, which could assist in our understanding of disease pathogenesis and the effects of antibiotics on the microbiome.

This study aimed to describe the nasopharyngeal microbiota in children with suspected PTB and the effects of TB treatment on the nasopharyngeal microbiota. The findings suggest that there are no significant differences in the microbiota profile or diversity between the TB cases and ill controls at baseline based on alpha and beta diversity (Bray Curtis and weighted UniFrac), which may have been influenced by the fact that both groups investigated were ill and probably received antibiotic therapy prior to recruitment into the study. Similarly, no significant differences were apparent in the alpha diversity after 2 and 6 months of TB treatment in the TB case group which may suggest that TB treatment did not have a major influence on the diversity but could have contributed to the changes observed to familia in the nasopharyngeal samples. Without the addition of differential abundance testing the results would suggest that no differences existed between the TB case and ill control group or that TB treatment had no effect on the microbiota. This was found to be an imprecise description of the results of the study as at the family level the TB cases did differ to the ill controls and changes to the microbiota was observed after TB treatment. The inclusion of this test therefore allowed for a more informed interpretation of the results for this study and also showed the importance of including quantitative data alongside qualitative data analysis. However, these analyses were limited by the small sample size, and in particular the poor sequencing quality which limited taxonomic classification to the family level. Therefore, further research into these questions are warranted.

## CHAPTER 5: Concluding remarks

Tuberculosis (TB) is an ancient scourge that persists. The initial contributions to understanding the pathophysiology of the disease dates to the 1800's. To date, the major etiological agent has been identified with an understanding of disease pathogenesis, and measures to control the disease have been implemented through diagnosis and TB treatment. However, in children TB disease is somewhat obscured by the unclear clinical presentation and the paucibacillary nature of the disease. Through scientific advancements, microbial identification has become easier and microbial interplay has become increasingly evident. More so, that human health is established through our "microbiome". Respiratory microbiome research is still in its infancy, but studies have found associations between the upper respiratory tract and lower respiratory tract infections such as pneumonia (Sakwinska *et al.*, 2014; Kelly *et al.*, 2017), asthma (Teo, Mok, Pham, Kusel, Serralha, Troy, Barbara J Holt, *et al.*, 2015; Teo, Mok, Pham, Kusel, Serralha, Troy, Barbara J. Holt, *et al.*, 2015) and cystic fibrosis (Prevaes *et al.*, 2016). However, with regard to TB disease, respiratory microbiome research has focused on the adult lung, limiting the studies conducted in children.

This study aimed to compare the respiratory microbiota of children with bacteriologically confirmed, clinically diagnosed and unlikely TB (well-defined ill controls) and to describe the effect of TB treatment on the respiratory microbiota in children with PTB. The aims were achieved by completing three main objectives, firstly to determine the presence of various bacterial pathogens and the fungal pathogen (*Pneumocystis jirovecii*) in respiratory samples from children with suspected PTB at baseline (study entry), using the Seegene Allplex Respiratory Panel 4 and an in-house real-time PCR assay, respectively. Secondly, to describe the respiratory microbiota in children with PTB (bacteriologically confirmed/ clinically diagnosed) and without PTB (well-defined ill controls) and thirdly to describe the effect of TB treatment on the respiratory microbiota in children with PTB after 2 and 6 months in comparison to those without PTB. The second and third objectives were completed by sequencing the v4 region of the 16S ribosomal RNA (rRNA) gene on the Illumina Miseq platform.

The study population included children aged 0-13 years with suspected PTB that were enrolled from secondary and tertiary provincial hospitals (Karl Bremer and Tygerberg Hospital, respectively) in Cape Town. Participants were classified as having "bacteriologically confirmed TB", "clinically diagnosed TB (unconfirmed)", or "unlikely PTB" after enrollment and intensive

investigation and clinical follow-up (Chapter 1, Section 1.9). A healthy control group was not included in this study.

The first section of this study showed no association between the presence or absence of certain pathogens (and pathobionts) and the different disease categories evaluated (bacteriologically diagnosed/clinically diagnosed and unlikely PTB). This may have been a result of the limited sample size, the lack of a healthy control group or being limited to the detection of certain bacteria and fungi based on the assays used. Nonetheless, it was observed that *Streptococcus pneumoniae* and *Haemophilus influenzae* were the most prevalent bacterial species identified in participants with suspected PTB, which was not surprising since both pathobionts are among the common and transient pathobionts identified in the nasopharynx (Bogaert *et al.*, 2011). However, these pathobionts are also associated with unstable colonization patterns in children, which has been suggested to increase the risk of development of respiratory infections (Biesbroek, Tsvitvadze, Elisabeth A M Sanders, *et al.*, 2014). This could be relevant to our study since all participants that were included in this study suffered from respiratory disease. However, this would have to be explored further in relation to a healthy control group to evaluate the significance of these pathobionts in our study.

Atypical pathogens such as *Chlamydiae pneumoniae* and *Mycoplasma pneumoniae* were detected in both the unlikely TB and the TB case groups, respectively. The presence of atypical respiratory pathogens in the unlikely TB group could have provided insight into the cause of the clinical disease in these participants, but based on their detection amongst TB cases, the findings are difficult to interpret. The limited atypical pathogens detected (bacteria and fungal) may suggest that viruses contribute to disease in the unlikely TB controls. Like bacteria, viruses can also exist in the nasopharynx without causing disease, i.e. asymptomatic carriage. However, viruses are also most likely to contribute to acute respiratory infections in young children (Brealey *et al.*, 2015) which supports the likelihood of it contributing to disease in the unlikely TB group. This is being explored in the same population but has yet to be combined with our data set. The importance of including viruses is that viral and bacterial interactions can exist which can contribute to disease severity or acquisition of other pathogenic bacteria. This was previously reported in adults with suspected PTB, where being co-infected with both viruses and bacteria was associated with more severe TB disease (Mhimbira *et al.*, 2018). Viruses were also seen to dominate the microbial profile in children with TB, but the significance of co-detection with viruses and bacteria could not be determined (Dube *et al.*, 2016). It was suggested that the host becomes susceptible to other respiratory infections as a consequence of relative immune suppression or

the lung pathology associated with PTB or becomes predisposed to an accelerated clinical course or presentation of symptoms due to intercurrent infections that caused immunosuppression (Dube *et al.*, 2016). The detection of certain microorganisms in our study provided insight into what was present in the nasopharynx in children with suspected PTB but to better understand the significance and association of these pathogens (and pathobionts), they should be evaluated alongside viruses (and fungi) to contribute to our understanding of disease pathogenesis, severity and clinical presentation of disease in those with and without PTB and in healthy controls.

Certain risk factors for the presence of bacterial pathogens were evaluated of which only mode of delivery and breastfeeding were found to be statistically significantly associated. Nonetheless, some studies found that these particular risk factors are associated with the carriage of microorganisms as seen for mode of delivery (Esposito and Principi, 2018) or offer potential protective benefits against respiratory infections (Biesbroek, Tsvitshivadze, Elisabeth A. M. Sanders, *et al.*, 2014), where other studies did not see an association with breastfeeding (Kaleida *et al.*, 1991; Bakhshae *et al.*, 2015). In future, this would need to be evaluated alongside a healthy control group and on a larger scale to determine the relevance in our population.

The second part of this study focused on describing the nasopharyngeal microbiota in children with suspected PTB as well as to describe the effect of TB treatment on nasopharyngeal microbiota. Although the term microbiota describes bacteria, viruses and fungi etc., the focus of this study was on the bacterial assemblage of organisms in the nasopharynx. This section of the study focused on a subset of participants (TB cases (drug susceptible TB cases) and well-defined ill-controls) and established and assessed the utility of the 16S rRNA library sequencing in our setting.

No distinct microbial profile was identified for the TB cases (bacteriologically confirmed/clinically diagnosed TB) or ill control group at baseline or at follow up time points. The alpha and beta diversity (Bray-Curtis and weighted UniFrac) measures also showed no significant differences between the TB cases or ill control group at baseline or at follow up time points. This may have been due to the fact that all participants were ill and that healthy participants were not included or due to the fact that the majority of participants were exposed to antibiotics prior to and after recruitment. First-line TB treatment did not seem to largely influence the microbial composition and diversity in this population. Contrary to the standard alpha and beta diversity measures used in this study, the addition of differential abundance testing indicated that shifts in abundance of certain Familia between the TB case group at baseline and month 6 occurred. Which indicates that at a “superficial” level it may seem that TB treatment did not have an overall effect on the

composition of taxa but could influence the abundance of certain taxa. However, evaluating this on a larger scale and with a healthy control group may provide better insight into the microbial composition and diversity changes after TB treatment. The use of 16S rRNA sequencing informs us of what pathogens (or commensals) are present (specifically bacteria) in a particular niche or sample, however, it does not provide insight into what other (non-bacterial) pathogens or pathobionts are present or what they are doing. Shotgun metagenomics may be more appropriate for this as it provides more information regarding the microbial functionality and biological processes that are taking place (Segata *et al.*, 2013) but has additional financial, expertise and analytical challenges. As far as it is known no study has investigated the role of the nasopharyngeal microbiome in relation to TB disease which makes it difficult to compare to other studies. The studies that have investigated the role of the respiratory microbiome in relation to TB disease either investigated the lung microbiome (microbiota) in adults (Cui *et al.*, 2012; Botero *et al.*, 2014; Eshetie and Van Soolingen, 2019) or identified nasopharyngeal microorganisms using other PCR-based methods (Dube *et al.*, 2016), which allows us to compare the presence of certain pathogens but not the role these pathogens play in these children specifically. This signifies the importance of investigating these findings further.

#### Limitations of microbiome sequencing in our study

This pilot study was limited by technical issues related to sample collection and storage, DNA quality, 16S library preparation and sequencing and data analysis. Limited sample volume limited our ability to re-extract samples with poor quantity or quality DNA, and suboptimal storage conditions (long term at -20°C) may have compromised sample integrity. All respiratory samples collected at baseline and follow up visits (month 2 and month 6) and selected for microbiome analysis, were included, irrespective of DNA quality, sample type (NPA versus IS), time point (month 1 versus month 2) or whether the sample DNA was amplifiable by PCR. Under ideal circumstances, samples would be excluded on the basis of these factors to exclude confounders, however, due to the limited number of participants that were selected for this section of the work, all samples were included. Poorer DNA quality is often seen in low biomass respiratory samples which are also easily contaminated (Faner *et al.*, 2017). This should be kept in mind when analyzing microbiome results as contamination may provide an inaccurate representation of what is present in the samples. However, in our study contamination did not seem to majorly influence the results, but this highlights the importance of including adequate controls in the study.

Paired-end sequencing (2x300 bp) of the v4 region of the 16S rRNA gene was done on the Illumina Miseq platform at the Institute for Microbial Biotechnology and Metagenomics,

Department of Biotechnology, University of the Western Cape. This introduced additional challenges as the library preparation was performed at two different facilities, which was not ideal especially because of the consistency that is required to conduct thorough microbiome research. Furthermore, due to the infrastructural challenges in South Africa, load shedding was experienced during the sequencing run and due to failure of the back-up generator, sequencing was terminated prematurely, resulting in shorter reverse reads. These challenges highlight limitations in conducting research in resource limited settings. Due to these challenges, only the forward reads were used for the microbiome analysis which may be a reason why the sequences could only be classified to a higher taxonomic level (family). The database (Greengenes) used for the sequencing alignment is the default database in the QIIME pipeline, and is the most frequently used; but is also the least updated of the available databases for 16S taxonomic assignment (Pollock *et al.*, 2018). Taxonomic assignment is a primary component of microbial community analysis, so choosing a suitable database is essential as it may influence post analysis and interpretation of the community composition (Park and Won, 2018). Park and Won (2018) evaluated bacterial reference databases (Greenegenes, SILVA and EzBioCloud) with a mock community and found that Greengenes predicted fewer genera than expected. This was thought to be due to the fact that the database was not updated (since 2013) and therefore would not contain newly identified novel bacterial sequences. However, in our study some common bacteria (*Escherichia* and *Enterococcus* sp) could not be identified at the genus level, which may be due to the sequences not being specific enough, due to suboptimal sequencing quality, to accurately identify taxa at the genus level and the short read length which influences taxonomic assignment; as a result the family level was used.

This could also be a disadvantage of using OTUs for taxonomic clustering as sequences are clustered based on similarity to the database, therefore any errors in the database caused by sequencing/PCR error or by incorrectly labelled sequences may lead to misclassification or identification (Pollock *et al.*, 2018). In recent years an alternative approach to using OTUs has been introduced, namely amplicon sequence variants (ASVs) or exact sequence variants (ESVs) (Callahan, McMurdie and Holmes, 2017). This has also been incorporated into QIIME 2 which has since succeeded QIIME 1. ASVs are inferred by a *de novo* process where sequences are differentiated from errors, partly based on the expectation that biological sequences are more likely to be repeatedly observed than are error sequences. Consequently, ASV inference is not performed separately on each read but by sample. This could potentially be beneficial in low biomass samples, such as respiratory samples, that are prone to contamination and could be considered for future studies. However, using new methods to analyze high throughput

sequencing data makes it difficult to compare results from one study to another. Therefore, the selection of pipeline, database or clustering method used should be guided by what is commonly used for the biological site being investigated and what is available at hand in the research setting where the analysis will be conducted.

The relative abundance of OTUs was used in our study to normalize the data, however this approach is not always recommended as it has been reported that it can lead to samples clustering by sequencing depth (Goodrich *et al.*, 2014). This approach was used as the alternative method, rarefaction, was considered to be a more stringent approach which would have substantially reduced the amount of sequences to work with. Rarefaction requires equal number of sequences to be selected from each sample, which can be selected based on the sample with the least amount of sequences. A disadvantage to this approach is that data from high sequence count samples will be discarded, but may at the same time be a conservative approach to view the abundances of rarer taxa across the samples (Goodrich *et al.*, 2014). Identifying rarer taxa in our study was not particularly pertinent as we aimed to describe the nasopharyngeal microbiota as whole in children with suspected PTB and therefore we considered using the relative abundance of OTUs to be sufficient for this pilot study.

In a general sense, microbiome research is a fairly new concept and much of the focus of this research has been on gastrointestinal microbiome research in comparison to other microbial habits, such as the respiratory microbiome which is still in its infancy. Therefore, various methods are being implemented to conduct research and the best practice has yet to be determined. Goodrich *et al* (2014) and Pollock *et al* (2018) reviewed and proposed various (general) approaches that should be considered during microbiome research this ranged from sample collection to contamination issues. A standardized protocol could be established based on this reviewed literature and should be considered for future microbiome projects. The main suggestion is that all data should be recorded throughout the project and that consistency is key. In future, a single database could be generated prior to starting the study which would include participant information and different factors that could influence the overall result, e.g. antibiotic use and the duration it was used and distributed to the respective researchers. For example, including the sample type and the way in which the sample was stored, i.e. stored in a cryoprotectant and temperature at which it was stored after collection. Controls could also be obtained at this point which could account for possible contamination during sample collection. As for library preparation and sequencing this would depend on the availability of appropriate equipment required and an appropriate laboratory to conduct the research. This is not always feasible in

resource limited settings, but each step during library preparation should be recorded and accounted for.

### Future recommendations

Not many studies highlight the challenges that are experienced alongside research outputs. In doing this it assists researchers who are experiencing similar challenges and informs them of possible ways in which it can be overcome. Other challenges such as lack of support can also be a factor in microbiome research, especially when this type of research is a fairly new research concept that is being explored in the researcher's department. However, this can be addressed through collaborations with other researchers and departments (as was seen in this study). For example, Stellenbosch University recently established the African Microbiome Institute, it is a fairly new establishment that has the potential to offer the necessary support that is not always readily available for departments that are trying to broaden their microbiome research. This could be a facility where students interact with one another about challenges that is experienced and ways in which it could possibly be approached alongside research outputs. Additionally, identifying researchers and students with similar research interests, such as exploring the gut microbiome or the respiratory microbiome will allow for the ideal opportunity to discuss approaches from sample collection to sequencing analysis which could possibly assist in establishing standardized protocols that could be used depending on the specific research project. These protocols would be a useful guide for departments while establishing some sort of consistency for microbiome projects.

## Conclusion

In conclusion, this study observed no significant differences in terms of alpha or beta diversity between the respiratory microbiotas in children with bacteriologically confirmed, clinically diagnosed and unlikely PTB categories or before and after TB treatment. However, differential abundance testing indicated that TB cases and ill controls differed and that TB treatment contributed to shifts in certain Familia after TB treatment. The Seegene Allplex respiratory panel data was comparable to the microbiota data as bacteria identified with the Seegene respiratory panel belonged to familia identified in the samples that were sequenced. This provides confidence in the sequencing information that was obtained irrespective of the challenges experienced and the lack of stringent quality filtering steps after sequencing. Additionally, the microbiotas of both TB cases and ill controls appeared unstable which should be investigated further. The instability of the microbiome in the ill control groups may be related to other factors such as poor nutrition or environmental exposures like smoking. This warrants further research as possible interventions like probiotics, nutritional support could potentially prevent future respiratory disease in this vulnerable population. This study contributed to the data available with regard to respiratory microbiome (microbiota) in children with suspected PTB in a TB endemic setting and also highlighted the challenges of conducting research in resource limited settings and suggests ways to approach these challenges in future.

## References

- Adams, L. V and Starke, J. (2019) *Tuberculosis disease in children - UpToDate*. Available at: <https://www.uptodate.com/contents/tuberculosis-disease-in-children> (Accessed: 4 October 2018).
- adonis function | R Documentation* (no date). Available at: <https://www.rdocumentation.org/packages/vegan/versions/2.4-2/topics/adonis> (Accessed: 8 December 2019).
- Allen, E. K. *et al.* (2014) 'Characterization of the nasopharyngeal microbiota in health and during rhinovirus challenge.', *Microbiome*, 2(1), p. 22. doi: 10.1186/2049-2618-2-22.
- Anderson, M. J. (2001) 'A new method for non-parametric multivariate analysis of variance', *Austral Ecology*, 26(1), pp. 32–46. doi: 10.1111/j.1442-9993.2001.01070.pp.x.
- Antwi, S. *et al.* (2017) 'Pharmacokinetics of the first-line antituberculosis drugs in Ghanaian children with tuberculosis with or without HIV coinfection', *Antimicrobial Agents and Chemotherapy*. American Society for Microbiology, 61(2). doi: 10.1128/AAC.01701-16.
- Bakhshaei, M. *et al.* (2015) 'Breastfeeding and Nasopharyngeal Colonization With Common Respiratory Pathogens Among Children', *Shiraz E-Medical Journal*. Kowsar, 16(8). doi: 10.17795/semj20295.
- Balajee, R. and Dhana Rajan, M. S. (2011) 'Molecular docking studies of Pantothenate synthase bound with novel molecules for Mycobacterium Tuberculosis', *Journal of Pharmaceutical Sciences and Research*, 3(6), pp. 1276–1279. Available at: <http://www.scfbioitd.res.in/utility/LipinskiFilters> (Accessed: 1 October 2018).
- Banerjee, K. *et al.* (2019) 'An Adaptive Multivariate Two-Sample Test With Application to Microbiome Differential Abundance Analysis', *Frontiers in Genetics*. Frontiers Media S.A., 10(APR), p. 350. doi: 10.3389/fgene.2019.00350.
- van den Bergh, M. R. *et al.* (2012) 'Associations between Pathogens in the Upper Respiratory Tract of Young Children: Interplay between Viruses and Bacteria', *PLoS ONE*, 7(10). doi: 10.1371/journal.pone.0047711.
- Biesbroek, G., Tsivtsivadze, E., Sanders, Elisabeth A M, *et al.* (2014) 'Early respiratory microbiota composition determines bacterial succession patterns and respiratory health in children', *American Journal of Respiratory and Critical Care Medicine*. American Thoracic Society, 190(11), pp. 1283–1292. doi: 10.1164/rccm.201407-1240OC.

Biesbroek, G., Tsvitvadze, E., Sanders, Elisabeth A. M., *et al.* (2014) 'Early Respiratory Microbiota Composition Determines Bacterial Succession Patterns and Respiratory Health in Children', *American Journal of Respiratory and Critical Care Medicine*. American Thoracic Society, 190(11), pp. 1283–1292. doi: 10.1164/rccm.201407-1240OC.

Bogaert, D. *et al.* (2011) 'Variability and diversity of nasopharyngeal microbiota in children: A metagenomic analysis', *PLoS ONE*. Edited by M. Semple. Public Library of Science, 6(2), pp. 1–8. doi: 10.1371/journal.pone.0017035.

Borg, I. *et al.* (2003) 'Evaluation of a quantitative real-time PCR for the detection of respiratory syncytial virus in pulmonary diseases.', *The European respiratory journal*. European Respiratory Society, 21(6), pp. 944–51. doi: 10.1183/09031936.03.00088102.

Bosch, A. A. T. M. *et al.* (2013) 'Viral and Bacterial Interactions in the Upper Respiratory Tract', *PLoS Pathogens*, 9(1). doi: 10.1371/journal.ppat.1003057.

Bosch, A. A. T. M. *et al.* (2016) 'Development of Upper Respiratory Tract Microbiota in Infancy is Affected by Mode of Delivery', *EBioMedicine*. Elsevier, 9, pp. 336–345. doi: 10.1016/j.ebiom.2016.05.031.

Botero, L. *et al.* (2014) 'Respiratory tract clinical sample selection for microbiota analysis in patients with pulmonary tuberculosis', *Microbiome*. BioMed Central Ltd., 2(1), pp. 1–7. doi: 10.1186/2049-2618-2-29.

Brealey, J. C. *et al.* (2015) 'Viral bacterial co-infection of the respiratory tract during early childhood', *FEMS Microbiology Letters*, 362. doi: 10.1093/femsle/fnv062.

Callahan, B. J., McMurdie, P. J. and Holmes, S. P. (2017) 'Exact sequence variants should replace operational taxonomic units in marker-gene data analysis', *ISME Journal*. Nature Publishing Group, 11, pp. 2639–2643. doi: 10.1038/ismej.2017.119.

Caporaso, J. G. *et al.* (2010) 'QIIME allows analysis of high-throughput community sequencing data', *Nature Methods*, pp. 335–336. doi: 10.1038/nmeth.f.303.

Caporaso, J. G. *et al.* (2011) 'Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 108 Suppl 1(Supplement 1), pp. 4516–22. doi: 10.1073/pnas.1000080107.

Case, R. J. *et al.* (2007) 'Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies', *Applied and Environmental Microbiology*, 73(1), pp. 278–288. doi:

10.1128/AEM.01177-06.

CDC (2013) *TB Elimination Extensively Drug-Resistant Tuberculosis (XDR TB) What is Extensively Drug-Resistant Tuberculosis (XDR TB)? How is XDR TB spread? Who is at risk for getting XDR TB? How can I prevent myself from getting TB? Can the TB vaccine (BCG) help prev.* Available at: <http://www.cdc.gov/tb> (Accessed: 20 September 2018).

Cheung, M. K. *et al.* (2013) 'Sputum Microbiota in Tuberculosis as Revealed by 16S rRNA Pyrosequencing', *PLoS ONE*, 8(1). doi: 10.1371/journal.pone.0054574.

Chiang, S. S., Swanson, D. S. and Starke, J. R. (2015) 'New Diagnostics for Childhood Tuberculosis', *Infect Dis Clin N Am*, 29, pp. 477–502. doi: 10.1016/j.idc.2015.05.011.

Chonmaitree, T. *et al.* (2017) 'Nasopharyngeal microbiota in infants and changes during viral upper respiratory tract infection and acute otitis media', *PLOS ONE*. Edited by E. N. Miyaji. Public Library of Science, 12(7), pp. 1–15. doi: 10.1371/journal.pone.0180630.

Churchyard, G. J. *et al.* (2014) 'Tuberculosis control in South Africa: Successes, challenges and recommendations', *South African Medical Journal*, 104(3), pp. 244–248. doi: 10.7196/SAMJ.7689.

Cillóniz, C. *et al.* (2011) 'Microbial aetiology of community-acquired pneumonia and its relation to severity.', *Thorax*. BMJ Publishing Group Ltd, 66(4), pp. 340–6. doi: 10.1136/thx.2010.143982.

Claassen-Weitz, S. *et al.* (2018) 'HIV-exposure, early life feeding practices and delivery mode impacts on faecal bacterial profiles in a South African birth cohort', *Scientific Reports*. Springer US, 8(1), pp. 1–15. doi: 10.1038/s41598-018-22244-6.

Cole, J. R. *et al.* (2014) 'Ribosomal Database Project: Data and tools for high throughput rRNA analysis', *Nucleic Acids Research*, 42(D1). doi: 10.1093/nar/gkt1244.

Cruz, A. T. and Starke, J. R. (2007) 'Clinical manifestations of tuberculosis in children', *Paediatric Respiratory Reviews*, 8(2), pp. 107–117. doi: 10.1016/j.prrv.2007.04.008.

Cui, Z. *et al.* (2012) 'Complex sputum microbial composition in patients with pulmonary tuberculosis', *BMC Microbiology*, 12, pp. 1–8. doi: 10.1186/1471-2180-12-276.

Davis, N. M. *et al.* (2018) 'Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data', *Microbiome*, 6(1), p. 226. doi: 10.1186/s40168-018-0605-2.

Dominguez-Bello, M. G. *et al.* (2010) 'Delivery mode shapes the acquisition and structure of the

initial microbiota across multiple body habitats in newborns', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 107(26), pp. 11971–11975. doi: 10.1073/pnas.1002601107.

Donald, P. R., Marais, B. J. and Barry, C. E. (2010) 'Age and the epidemiology and pathogenesis of tuberculosis', *The Lancet*, 375(9729), pp. 1852–1854. doi: 10.1016/S0140-6736(10)60580-6.

Dube, F. S. *et al.* (2016) 'Respiratory microbes present in the nasopharynx of children hospitalised with suspected pulmonary tuberculosis in Cape Town, South Africa.', *BMC infectious diseases*. BMC Infectious Diseases, 16(1), p. 597. doi: 10.1186/s12879-016-1934-z.

Dumbrell, A. J., Ferguson, R. M. . and Clark, D. R. (2016) 'Microbial Community Analysis by Single-Amplicon High-Throughput Next Generation Sequencing: Data Analysis – From Raw Output to Ecology', *Hydrocarbon and lipid Microbiology Protocols, Springer Protocols Handbooks*. Springer-Verlag Berlin Heidelberg, pp. 155–206. doi: 10.1007/8623\_2016\_228.

Dunn, J. J., Starke, J. R. and Revell, P. A. (2016) 'Laboratory Diagnosis of Mycobacterium tuberculosis Infection and Disease in Children.', *Journal of clinical microbiology*. American Society for Microbiology, 54(6), pp. 1434–41. doi: 10.1128/JCM.03043-15.

Edgar, R. C. *et al.* (2011) 'UCHIME improves sensitivity and speed of chimera detection', *Bioinformatics*. Oxford University Press, 27(16), pp. 2194–2200. doi: 10.1093/bioinformatics/btr381.

Ehlers, S. and Schaible, U. E. (2012) 'The granuloma in tuberculosis: dynamics of a host-pathogen collusion.', *Frontiers in immunology*. Frontiers Media SA, 3, p. 411. doi: 10.3389/fimmu.2012.00411.

Elahi, S. *et al.* (2008) 'Infection with *Bordetella parapertussis* but Not *Bordetella pertussis* Causes Pertussis-Like Disease in Older Pigs', *The Journal of Infectious Diseases*. Oxford University Press, 198(3), pp. 384–392. doi: 10.1086/589713.

Eshetie, S. and Van Soolingen, D. (2019) 'The respiratory microbiota: New insights into pulmonary tuberculosis', *BMC Infectious Diseases*. BioMed Central Ltd., 19(1), pp. 1–7. doi: 10.1186/s12879-019-3712-1.

Esposito, S. and Principi, N. (2018) 'Impact of nasopharyngeal microbiota on the development of respiratory tract diseases', *European Journal of Clinical Microbiology & Infectious Diseases*. Springer Berlin Heidelberg, 37(1), pp. 1–7. doi: 10.1007/s10096-017-3076-7.

Faden, H. *et al.* (1997) 'Relationship between Nasopharyngeal Colonization and the Development

of Otitis Media in Children', pp. 1440–1445.

Fadrosh, D. W. *et al.* (2014) 'An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform', *Microbiome*. BioMed Central, 2(1), p. 6. doi: 10.1186/2049-2618-2-6.

Faner, R. *et al.* (2017) 'The microbiome in respiratory medicine: Current challenges and future perspectives', *European Respiratory Journal*, 49(4), pp. 1–12. doi: 10.1183/13993003.02086-2016.

Fogel, N. (2015) 'Tuberculosis: A disease without boundaries', *Tuberculosis*. Churchill Livingstone, 95(5), pp. 527–531. doi: 10.1016/J.TUBE.2015.05.017.

Gadsby, N. J. *et al.* (2015) 'Development of two real-time multiplex PCR assays for the detection and quantification of eight key bacterial pathogens in lower respiratory tract infections.', *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*. Elsevier, 21(8), pp. 788.e1-788.e13. doi: 10.1016/j.cmi.2015.05.004.

Gallacher, D. J. and Kotecha, S. (2016) 'Respiratory Microbiome of New-Born Infants.', *Frontiers in pediatrics*. Frontiers Media S.A., 4(February), p. 10. doi: 10.3389/fped.2016.00010.

García-Rodríguez, J. Á. and Fresnadillo Martínez, M. J. (2002) 'Dynamics of nasopharyngeal colonization by potential respiratory pathogens', *J.Antimicrob.Chemother.*, 50 Suppl S(0305-7453 (Print)), pp. 59–73. doi: 10.1093/jac/dkf506.

Gerber, G. K. (2014) 'The dynamic microbiome', *FEBS Letters*. Federation of European Biochemical Societies, 588(22), pp. 4131–4139. doi: 10.1016/j.febslet.2014.02.037.

Goodrich, J. K. *et al.* (2014) 'Conducting a Microbiome Study', *Cell*. Elsevier Inc., 158(2), pp. 250–262. doi: 10.1016/J.CELL.2014.06.037.

Gradmann, C. (2006) 'Robert Koch and the white death : from tuberculosis to tuberculin', 8, pp. 294–301. doi: 10.1016/j.micinf.2005.06.004.

Graham, S. M. *et al.* (2006) 'Low levels of pyrazinamide and ethambutol in children with tuberculosis and impact of age, nutritional status, and human immunodeficiency virus infection', *Antimicrobial Agents and Chemotherapy*, 50(2), pp. 407–413. doi: 10.1128/AAC.50.2.407-413.2006.

Graham, S. M. (2011) 'Treatment of paediatric TB: Revised WHO guidelines', *Paediatric Respiratory Reviews*. Elsevier Ltd, 12(1), pp. 22–26. doi: 10.1016/j.prrv.2010.09.005.

Graham, S. M. *et al.* (2012) 'Evaluation of tuberculosis diagnostics in children: 1. Proposed clinical case definitions for classification of intrathoracic tuberculosis disease. Consensus from an expert panel.', *The Journal of infectious diseases*. Oxford University Press, 205 Suppl(Suppl 2), pp. S199-208. doi: 10.1093/infdis/jis008.

Gupta, R. *et al.* (2009) 'Diagnostic significance of nested polymerase chain reaction for sensitive detection of *Pneumocystis jirovecii* in respiratory clinical specimens', *Diagnostic Microbiology and Infectious Disease*. Elsevier, 64(4), pp. 381–388. doi: 10.1016/J.DIAGMICROBIO.2009.04.008.

He, Q. (1998) 'Whooping Cough Caused by *Bordetella pertussis* and *Bordetella parapertussis* in an Immunized Population', *JAMA*. American Medical Association, 280(7), p. 635. doi: 10.1001/jama.280.7.635.

Hong, B.-Y. *et al.* (2016) 'Microbiome Changes during Tuberculosis and Antituberculous Therapy.', *Clinical microbiology reviews*. American Society for Microbiology Journals, 29(4), pp. 915–26. doi: 10.1128/CMR.00096-15.

Hornung, B. V. H., Zwittink, R. D. and Kuijper, E. J. (2019) 'Issues and current standards of controls in microbiome research', *FEMS Microbiology Ecology*, 95, p. 45. doi: 10.1093/femsec/fiz045.

Huang, S. N. *et al.* (1999) 'Development of a PCR assay for diagnosis of *Pneumocystis carinii* pneumonia based on amplification of the multicopy major surface glycoprotein gene family.', *Diagnostic microbiology and infectious disease*, 35(1), pp. 27–32. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/10529878>.

Illumina (2019) *Cluster Optimization Overview Guide (1000000071511)*. Available at: [www.illumina.com/company/legal.html](http://www.illumina.com/company/legal.html). (Accessed: 26 June 2020).

Jaakkola, J. J. *et al.* (2006) 'Prenatal and postnatal tobacco smoke exposure and respiratory health in Russian children', *Respiratory Research*. BioMed Central, 7(1), p. 48. doi: 10.1186/1465-9921-7-48.

Janda, J. M. and Abbott, S. L. (2007) '16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls.', *Journal of clinical microbiology*. American Society for Microbiology (ASM), 45(9), pp. 2761–4. doi: 10.1128/JCM.01228-07.

Jnawali, H. N. and Ryoo, S. (2013) 'First – and Second – Line Drugs and Drug Resistance', *Tuberculosis-Current issues in diagnosis and management*, pp. 163–180. doi: 10.5772/54960.

Kaleida, P. H. *et al.* (1991) 'Prevalence of Bacterial Respiratory Pathogens in the Nasopharynx in

Breast-Fed versus Formula-Fed Infants', *JOURNAL OF CLINICAL MICROBIOLOGY*, 31(10), pp. 2674–2678. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC265971/pdf/jcm00022-0136.pdf> (Accessed: 28 April 2019).

Kelly, M. S. *et al.* (2017) 'The Nasopharyngeal Microbiota of Children With Respiratory Infections in Botswana', *The Pediatric Infectious Disease Journal*. NIH Public Access, 36(9), pp. e211–e218. doi: 10.1097/INF.0000000000001607.

Kim, M.-J. *et al.* (2010) 'Caseation of human tuberculosis granulomas correlates with elevated host lipid metabolism.', *EMBO molecular medicine*. Wiley-Blackwell, 2(7), pp. 258–74. doi: 10.1002/emmm.201000079.

Klindworth, A. *et al.* (2013) 'Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies.', *Nucleic acids research*. Oxford University Press, 41(1), p. e1. doi: 10.1093/nar/gks808.

Knechel, N. A. (2009) 'Tuberculosis: pathophysiology, clinical features, and diagnosis.', *Critical care nurse*. American Association of Critical Care Nurses, 29(2), pp. 34–43; quiz 44. doi: 10.4037/ccn2009968.

Lawn, S. D. and Zumla, A. I. (2011) 'Tuberculosis'. doi: 10.1016/S0140-6736(10)62173-3.

Lee, Y.-J. *et al.* (2015) 'Single-channel multiplexing without melting curve analysis in real-time PCR', *Scientific Reports*. Nature Publishing Group, 4(1), p. 7439. doi: 10.1038/srep07439.

Lewinsohn, D. M. A. M. A., Gennaro, M. L. and Scholvinck, L. (2004) 'Tuberculosis immunology in children : diagnostic and therapeutic challenges and opportunities', *World Health*, 8(5), pp. 658–674.

Li, Y. *et al.* (2017) '16S rDNA sequencing analysis of upper respiratory tract flora in patients with influenza H1N1 virus infection', *Frontiers in Laboratory Medicine*. Elsevier, 1(1), pp. 16–26. doi: 10.1016/J.FLM.2017.02.005.

Locht, C. (2016) 'Live pertussis vaccines: will they protect against carriage and spread of pertussis?', *Clinical Microbiology and Infection*. Elsevier, 22, pp. S96–S102. doi: 10.1016/J.CMI.2016.05.029.

Lockman, S. *et al.* (2003) 'Etiology of pulmonary infections in predominantly HIV-infected adults with suspected tuberculosis, Botswana', *International Journal Tuberculosis lung disease*. International Union Against Tuberculosis and Lung Disease, 7(8), pp. 714–723. Available at: <http://www.ingentaconnect.com/content/iuatld/ijtld/2003/00000007/00000008/art00003;jsessionid>

d=ffdaj1h5n41q.x-ic-live-01 (Accessed: 12 July 2018).

Love, M. I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*. BioMed Central Ltd., 15(12), p. 550. doi: 10.1186/s13059-014-0550-8.

Lozupone, C. and Knight, R. (2005) 'UniFrac: A new phylogenetic method for comparing microbial communities', *Applied and Environmental Microbiology*. American Society for Microbiology (ASM), 71(12), pp. 8228–8235. doi: 10.1128/AEM.71.12.8228-8235.2005.

Lucena-Aguilar, G. *et al.* (2016) 'DNA Source Selection for Downstream Applications Based on DNA Quality Indicators Analysis', in *Biopreservation and Biobanking*. Mary Ann Liebert Inc., pp. 264–270. doi: 10.1089/bio.2015.0064.

Man, W. H., de Steenhuijsen Piters, W. A. A. and Bogaert, D. (2017) 'The microbiota of the respiratory tract: gatekeeper to respiratory health', *Nature Reviews Microbiology*. Nature Publishing Group, 15(5), pp. 259–270. doi: 10.1038/nrmicro.2017.14.

Marais, B. J. *et al.* (2004) 'A proposed radiological classification of childhood intra-thoracic tuberculosis', *Pediatric Radiology*. doi: 10.1007/s00247-004-1238-0.

Marais, B. J. *et al.* (2006) 'Childhood Pulmonary Tuberculosis Old Wisdom and New Challenges', *Am J Respir Crit Care Med*, 173, pp. 1078–1090. doi: 10.1164/rccm.200511-1809SO.

Marais, B. J. and Schaaf, S. H. (2010) 'Tuberculosis in Children', *Cold Spring Harbor perspectives in medicine*, (January 2001), pp. 5–30. doi: 10.1101/cshperspect.a017855.

Martín, R. *et al.* (2014) 'The role of metagenomics in understanding the human microbiome in health and disease.', *Virulence*. Taylor & Francis, 5(3), pp. 413–23. doi: 10.4161/viru.27864.

McDonald, D. *et al.* (2012) 'An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea', *ISME Journal*, 6(3), pp. 610–618. doi: 10.1038/ismej.2011.139.

McNally, L. M. *et al.* (2007) 'Effect of age, polymicrobial disease, and maternal HIV status on treatment response and cause of severe pneumonia in South African children: a prospective descriptive study', *The Lancet*. Elsevier, 369(9571), pp. 1440–1451. doi: 10.1016/S0140-6736(07)60670-9.

Mhimbira, F. *et al.* (2018) 'Prevalence and clinical significance of respiratory viruses and bacteria detected in tuberculosis patients compared to household contact controls in Tanzania: a cohort study', *Clinical Microbiology and Infection*. Elsevier. doi: 10.1016/J.CMI.2018.03.019.

Mizgerd, J. P. (2014) 'The infant nose. Introducing the respiratory tract to the world.', *American journal of respiratory and critical care medicine*. American Thoracic Society, 190(11), pp. 1206–7. doi: 10.1164/rccm.201410-1919ED.

Morgan, X. C., Segata, N. and Huttenhower, C. (2013) 'Biodiversity and functional genomics in the human microbiome', *Trends in Genetics*, pp. 51–58. doi: 10.1016/j.tig.2012.09.005.

Morris, A. *et al.* (2008) 'Epidemiology and Clinical Significance of *Pneumocystis* Colonization', *The Journal of Infectious Diseases*. Oxford University Press, 197(1), pp. 10–17. doi: 10.1086/523814.

Morrow, B. M. *et al.* (2010) 'Pneumocystis Pneumonia in South African Children With and Without Human Immunodeficiency Virus Infection in the Era of Highly Active Antiretroviral Therapy', *The Pediatric Infectious Disease Journal*, 29(6), p. 1. doi: 10.1097/INF.0b013e3181ce871e.

Myer, P. R. *et al.* (2016) 'Evaluation of 16S rRNA amplicon sequencing using two next-generation sequencing technologies for phylogenetic analysis of the rumen bacterial community in steers', *Journal of Microbiological Methods*. Elsevier, 127, pp. 132–140. doi: 10.1016/J.MIMET.2016.06.004.

Naidoo, C. C. *et al.* (2019) 'The microbiome and tuberculosis: state of the art, potential applications, and defining the clinical research agenda', *The Lancet Respiratory Medicine*. Elsevier Ltd, 2600(18). doi: 10.1016/S2213-2600(18)30501-0.

Namasivayam, S. *et al.* (2017) 'Longitudinal profiling reveals a persistent intestinal dysbiosis triggered by conventional anti-tuberculosis therapy.', *Microbiome*. BioMed Central, 5(1), p. 71. doi: 10.1186/s40168-017-0286-2.

Namasivayam, S. *et al.* (2018) 'The Microbiome and Tuberculosis: Early Evidence for Cross Talk.', *mBio*. American Society for Microbiology (ASM), 9(5). doi: 10.1128/mBio.01420-18.

NICD (2016) *Legionnaires' disease Frequently Asked Questions*. Available at: [http://www.nicd.ac.za/assets/files/Legionella\\_FAQ\\_Jan\\_2016\\_final.pdf](http://www.nicd.ac.za/assets/files/Legionella_FAQ_Jan_2016_final.pdf) (Accessed: 4 August 2018).

NIH Human Microbiome Portfolio (2019) 'A review of 10 years of human microbiome research activities at the US National Institutes of Health, Fiscal Years 2007-2016', *Microbiome*, 7(1), p. 31. doi: 10.1186/s40168-019-0620-y.

O'Brien, K. L. *et al.* (2019) 'Causes of severe pneumonia requiring hospital admission in children without HIV infection from Africa and Asia: the PERCH multi-country case-control study', *The*

*Lancet*. Lancet Publishing Group, 394(10200), pp. 757–779. doi: 10.1016/S0140-6736(19)30721-4.

Oksanen, J. (2019) *Vegan: ecological diversity*.

Olsson, M. *et al.* (1993) *Detection of Pneumocystis carinii DNA in Sputum and Bronchoalveolar Lavage Samples by Polymerase Chain Reaction*, *JOURNAL OF CLINICAL MICROBIOLOGY*. Available at: <http://jcm.asm.org/> (Accessed: 13 November 2018).

Park, S.-C. and Won, S. (2018) 'Evaluation of 16S rRNA Databases for Taxonomic Assignments Using a Mock Community', *Genomics & Informatics*. Korea Genome Organization, 16(4), p. e24. doi: 10.5808/gi.2018.16.4.e24.

Piccini, P. *et al.* (2014) 'Clinical peculiarities of tuberculosis.', *BMC infectious diseases*. BioMed Central, 14 Suppl 1(Suppl 1), p. S4. doi: 10.1186/1471-2334-14-S1-S4.

Pollock, J. *et al.* (2018) 'The Madness of Microbiome: Attempting To Find Consensus &quot;Best Practice&quot; for 16S Microbiome Studies.', *Applied and environmental microbiology*. American Society for Microbiology, 84(7), pp. e02627-17. doi: 10.1128/AEM.02627-17.

Prevaes, S. M. P. J. *et al.* (2016) 'Development of the Nasopharyngeal Microbiota in Infants with Cystic Fibrosis', *American Journal of Respiratory and Critical Care Medicine*, 193(5), pp. 504–515. doi: 10.1164/rccm.201509-1759OC.

Principi, N. *et al.* (2001) 'Role of Mycoplasma pneumoniae and Chlamydia pneumoniae in Children with Community-Acquired Lower Respiratory Tract Infections', *Clinical Infectious Diseases*. Oxford University Press, 32(9), pp. 1281–1289. doi: 10.1086/319981.

Ramachandran, G. *et al.* (2016) 'Low Serum Concentrations of Rifampicin and Pyrazinamide Associated with Poor Treatment Outcomes in Children with Tuberculosis Related to HIV Status', *Pediatric Infectious Disease Journal*. Lippincott Williams and Wilkins, 35(5), pp. 530–534. doi: 10.1097/INF.0000000000001069.

Ravi, R. K., Walton, K. and Khosroheidari, M. (2018) 'Miseq: A next generation sequencing platform for genomic analysis', in *Methods in Molecular Biology*. Humana Press Inc., pp. 223–232. doi: 10.1007/978-1-4939-7471-9\_12.

Rosas-Salazar, C. *et al.* (2016) 'Nasopharyngeal Microbiome in Respiratory Syncytial Virus Resembles Profile Associated with Increased Childhood Asthma Risk.', *American journal of respiratory and critical care medicine*. American Thoracic Society, 193(10), pp. 1180–3. doi: 10.1164/rccm.201512-2350LE.

- Rudzani Muloiwa; Felix S. Dube; Mark P. Nicol; Heather J. Zar; Gregory D. Hussey (2016) 'Incidence and Diagnosis of Pertussis in South African children Hospitalized with lower respiratory tract infection', *The Pediatric Infectious Disease Journal @BULLET*, 35(6). doi: 10.1097/INF.0000000000001132.
- Ruiz, M. *et al.* (1999) 'Etiology of Community-Acquired Pneumonia', *American Journal of Respiratory and Critical Care Medicine*. American Thoracic Society New York, NY, 160(2), pp. 397–405. doi: 10.1164/ajrccm.160.2.9808045.
- Ryu, Y. J. (2015) 'Diagnosis of pulmonary tuberculosis: recent advances and diagnostic algorithms.', *Tuberculosis and respiratory diseases*. The Korean Academy of Tuberculosis and Respiratory Diseases, 78(2), pp. 64–71. doi: 10.4046/trd.2015.78.2.64.
- Sakwinska, O. *et al.* (2014) 'Nasopharyngeal microbiota in healthy children and pneumonia patients.', *Journal of clinical microbiology*. American Society for Microbiology (ASM), 52(5), pp. 1590–1594. doi: 10.1128/JCM.03280-13.
- Saleem, A. and Azher, M. (2013) 'The Next Pandemic -Tuberculosis: The Oldest Disease of Mankind Rising One More Time', *British Journal of Medical Practitioners BJMP*, 66(21). Available at: <http://www.bjmp.org/files/2013-6-2/bjmp-2013-6-2-a615.pdf> (Accessed: 23 February 2018).
- Salter, S. J. *et al.* (2014) 'Reagent and laboratory contamination can critically impact sequence-based microbiome analyses', *BMC Biology*. BioMed Central, 12(1), pp. 1–12. doi: 10.1186/s12915-014-0087-z.
- Samuel, Catherine M. *et al.* (2011) 'Improved detection of *Pneumocystis jirovecii* in upper and lower respiratory tract specimens from children with suspected pneumocystis pneumonia using real-time PCR: a prospective study', *BMC Infectious Diseases*. BioMed Central, 11(1), p. 329. doi: 10.1186/1471-2334-11-329.
- Santee, C. A. *et al.* (2016) 'Nasopharyngeal microbiota composition of children is related to the frequency of upper respiratory infection and acute sinusitis.', *Microbiome*. Microbiome, 4(1), p. 34. doi: 10.1186/s40168-016-0179-9.
- Schaaf, H. S. *et al.* (1995) 'Respiratory tuberculosis in childhood: the diagnostic value of clinical features and special investigations.', *The Pediatric infectious disease journal*, 14(3), pp. 189–94. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/7761183> (Accessed: 16 May 2019).

Schirmer, M. *et al.* (2015) 'Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform', *Nucleic Acids Research*. Oxford University Press, 43(6). doi: 10.1093/nar/gku1341.

Schloss, P. D. *et al.* (2009) 'Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities', *Applied and Environmental Microbiology*, 75(23), pp. 7537–7541. doi: 10.1128/AEM.01541-09.

Seddon, J. A. *et al.* (2018) 'The wonder years: What can primary school children teach us about immunity to mycobacterium tuberculosis?', *Frontiers in Immunology*. Frontiers Media S.A. doi: 10.3389/fimmu.2018.02946.

Segata, N. *et al.* (2013) 'Computational meta'omics for microbial community studies', *Molecular Systems Biology*. Blackwell Publishing Ltd. doi: 10.1038/msb.2013.22.

Shaler, C. R. *et al.* (2013) 'Within the Enemy's Camp: contribution of the granuloma to the dissemination, persistence and transmission of Mycobacterium tuberculosis.', *Frontiers in Immunology*. Frontiers Media SA, 4, p. 30. doi: 10.3389/fimmu.2013.00030.

Shilts, M. H. *et al.* (2016) 'Minimally Invasive Sampling Method Identifies Differences in Taxonomic Richness of Nasal Microbiomes in Young Infants Associated with Mode of Delivery.', *Microbial ecology*. NIH Public Access, 71(1), pp. 233–42. doi: 10.1007/s00248-015-0663-y.

Shukla, S. D. *et al.* (2017) 'Microbiome effects on immunity, health and disease in the lung.', *Clinical & translational immunology*. Nature Publishing Group, 6(3), p. e133. doi: 10.1038/cti.2017.6.

Sotgiu, G. *et al.* (2015) 'Tuberculosis treatment and drug regimens.', *Cold Spring Harbor perspectives in medicine*. Cold Spring Harbor Laboratory Press, 5(5), p. a017822. doi: 10.1101/cshperspect.a017822.

South African National Department of Health (2013) *GUIDELINES FOR THE MANAGEMENT OF TUBERCULOSIS IN CHILDREN 2013*. Available at: [www.doh.gov.za](http://www.doh.gov.za) (Accessed: 7 December 2019).

Stearns, J. C. *et al.* (2015) 'Culture and molecular-based profiles show shifts in bacterial communities of the upper respiratory tract that occur with age.', *The ISME journal*. Nature Publishing Group, 9, pp. 1246–59. doi: 10.1038/ismej.2014.250.

Swaminathan, S. (2004) 'Tuberculosis in HIV-infected children', *Paediatric Respiratory Reviews*, 5(3), pp. 225–230. doi: 10.1016/j.prrv.2004.04.006.

Teo, S. M., Mok, D., Pham, K., Kusel, M., Serralha, M., Troy, N., Holt, Barbara J, *et al.* (2015) 'The infant airway microbiome in health and disease impacts later asthma development', *Cell Host Microbe*, 17(5), pp. 704–715. doi: 10.1016/j.chom.2015.03.008.

Teo, S. M., Mok, D., Pham, K., Kusel, M., Serralha, M., Troy, N., Holt, Barbara J., *et al.* (2015) 'The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development', *Cell Host and Microbe*. Elsevier Inc., 17(5), pp. 704–715. doi: 10.1016/j.chom.2015.03.008.

The NIH HMP Working Group (2009) 'The NIH Human Microbiome Project', *Genome Research*, 19(12), pp. 2317–2323. doi: 10.1101/gr.096651.109.

Tromp, I. *et al.* (2017) 'Breastfeeding and the risk of respiratory tract infections after infancy: The Generation R Study.', *PloS one*. Public Library of Science, 12(2), p. e0172763. doi: 10.1371/journal.pone.0172763.

Truong, J. and Ashurst, J. V. (2019) *Pneumocystis (Carinii) Jiroveci Pneumonia*, StatPearls. StatPearls Publishing. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/29493992> (Accessed: 2 May 2019).

Tsai, K.-S. *et al.* (2013) 'Childhood Tuberculosis: Epidemiology, Diagnosis, Treatment, and Vaccination', *Pediatrics & Neonatology*. Elsevier Taiwan LLC, 54(5), pp. 295–302. doi: 10.1016/j.pedneo.2013.01.019.

Del Valle-Mendoza, J. *et al.* (2017) 'High Prevalence of Mycoplasma pneumoniae and Chlamydia pneumoniae in Children with Acute Respiratory Infections from Lima, Peru.', *PloS one*. Public Library of Science, 12(1). doi: 10.1371/journal.pone.0170787.

Wagner, B. D. *et al.* (2018) 'On the Use of Diversity Measures in Longitudinal Sequencing Studies of Microbial Communities', *Frontiers in Microbiology*, 9, pp. 1–11. doi: 10.3389/fmicb.2018.01037.

Wang, J. *et al.* (2017) 'Breastfeeding and respiratory tract infections during the first 2 years of life.', *ERJ open research*. European Respiratory Society, 3(2). doi: 10.1183/23120541.00143-2016.

'WHO | Tuberculosis and HIV' (2019) WHO. World Health Organization. Available at: <https://www.who.int/hiv/topics/tb/en/> (Accessed: 13 May 2019).

Wipperman, M. F. *et al.* (2017) 'Antibiotic treatment for Tuberculosis induces a profound dysbiosis of the microbiome that persists long after therapy is completed', *Scientific Reports*. Nature Publishing Group, 7(1), p. 10767. doi: 10.1038/S41598-017-10346-6.

World Health Organization (2013) 'WHO News Release', *Saudi Med J*, 34(11), pp. 1205–1207.

World Health Organization (2017) *Global Tuberculosis Report*.

World Health Organization (2018) *Global Health TB Report*.

World Health Organization (2019) *Global Tuberculosis Report, Global Tuberculosis Report 2012*. doi: 978 92 4 156450 2.

Yang, B., Wang, Y. and Qian, P.-Y. (2016) 'Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis.', *BMC bioinformatics*. BioMed Central, 17, p. 135. doi: 10.1186/s12859-016-0992-y.

Yilmaz, P. *et al.* (2014) 'The SILVA and "all-species Living Tree Project (LTP)" taxonomic frameworks', *Nucleic Acids Research*, 42(D1). doi: 10.1093/nar/gkt1209.

Zar, H. J. (2010) 'Pneumocystis pneumonia in HIV-infected children: recent advances and future hurdles', *Pediatric Health*. Future Medicine Ltd London, UK , 4(3), pp. 243–245. doi: 10.2217/phe.10.25.

Zar, H. J. *et al.* (2016) 'Aetiology of childhood pneumonia in a well vaccinated South African birth cohort: a nested case-control study of the Drakenstein Child Health Study', *The Lancet Respiratory Medicine*, 4(6), pp. 463–472. doi: 10.1016/S2213-2600(16)00096-5.

## Addendum 1

**Iib. Clinically significant fluid specimens** are extracted as per adaptation of the Body Fluids Protocol (Pg 27) in the Qiagen QIAamp DNA Mini and Blood Mini Handbook.

1. Pipette 20µl of proteinase K into a 1.5ml microcentrifuge tube.
  2. Add 200µl of sample to the microcentrifuge tube.
  3. Add 200µl of Buffer AL to the sample, mix by pulse-vortexing for 15 seconds.
  4. Place in a heating block at 56°C for 10 minutes.
  5. Add 200µl ethanol (96-100%), pulse vortex for 15s.
  6. Transfer solution (including any precipitate) to the QIAamp Mini Spin column (in 2ml collection tube) without wetting the rim.
  7. Centrifuge at 6000xg (8000 rpm) for 1 minute.
  8. Place the QIAamp Mini Spin column in a clean 2ml collection tube.
  9. Add 500µl Buffer AW1 to column.
  10. Centrifuge at 6000xg (8000 rpm) for 1 minute.
  11. Place the QIAamp Mini Spin column in a clean 2ml collection tube.
  12. Add 500µl Buffer AW2 to column.
  13. Centrifuge at full speed for 3 minutes.
  14. Place the QIAamp Mini Spin column in a clean 2ml collection tube.
  15. Centrifuge at full speed for 1 minute.
  16. Place the QIAamp Mini Spin column in a clean sterile 1.5ml Eppendorf tube.
- 
17. Add 50µl of buffer AE, incubate at room temperature for 2 minutes.
  18. Centrifuge at 6000xg (8000 rpm) for 1 minute to elute the DNA.

Qiagen QIAamp protocol for DNA isolation from samples

## Addendum 2

# 16S Metagenomic Sequencing Library Preparation

## *Preparing 16S Ribosomal RNA Gene Amplicons for the Illumina MiSeq System*

Introduction	2
16S Library Preparation Workflow	5
Amplicon PCR	6
PCR Clean-Up	8
Index PCR	10
PCR Clean-Up 2	13
[Optional] Validate Library	15
Library Quantification, Normalization, and Pooling	16
Library Denaturing and MiSeq Sample Loading	17
MiSeq Reporter Metagenomics Workflow	20
Supporting Information	21

### IMPORTANT NOTICE

This document provides information for an application for Illumina technology that has been demonstrated internally and may be of interest to customers. This information is provided as-is and is not an Illumina product and is not accompanied by any rights or warranties. Customers using or adapting this information should obtain any licenses required and materials from authorized vendors. Illumina products mentioned herein are for research use only unless marked otherwise. While customer feedback is welcomed, this application is not supported by Illumina Technical Support and Field Application Scientists.

## Introduction

Page 146

# Introduction

Metagenomic studies are commonly performed by analyzing the prokaryotic 16S ribosomal RNA gene (16S rRNA), which is approximately 1,500 bp long and contains nine variable regions interspersed between conserved regions. Variable regions of 16S rRNA are frequently used in phylogenetic classifications such as genus or species in diverse microbial populations.

Which 16S rRNA region to sequence is an area of debate, and your region of interest might vary depending on things such as experimental objectives, design, and sample type. This protocol describes a method for preparing samples for sequencing the variable V3 and V4 regions of the 16S rRNA gene. This protocol can also be used for sequencing other regions with different region-specific primers. This protocol combined with a benchtop sequencing system, on-board primary analysis, and secondary analysis using MiSeq Reporter or BaseSpace, provides a comprehensive workflow for 16S rRNA amplicon sequencing.

Workflow Summary:

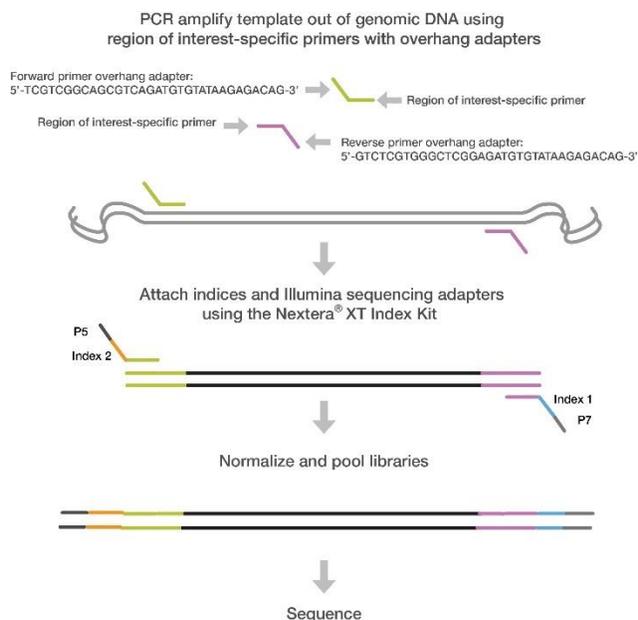
- 1** Order amplicon primers—The protocol includes the primer pair sequences for the V3 and V4 region that create a single amplicon of approximately ~460 bp. The protocol also includes overhang adapter sequences that must be appended to the primer pair sequences for compatibility with Illumina index and sequencing adapters. Illumina does not sell these primers. They must be ordered from a third party. See [Amplicon Primers](#), on page 3 for more information on amplicon primers.
- 2** Prepare library—The protocol describes the steps to amplify the V3 and V4 region and using a limited cycle PCR, add Illumina sequencing adapters and dual-index barcodes to the amplicon target. Using the full complement of Nextera XT indices, up to 96 libraries can be pooled together for sequencing.
- 3** Sequence on MiSeq—Using paired 300-bp reads, and MiSeq v3 reagents, the ends of each read are overlapped to generate high-quality, full-length reads of the V3 and V4 region in a single 65-hour run. The MiSeq run output is approximately > 20 million reads and, assuming 96 indexed samples, can generate > 100,000 reads per sample, commonly recognized as sufficient for metagenomic surveys.
- 4** Analyze on MSR or BaseSpace—The Metagenomics workflow is a secondary analysis option built into the MiSeq Reporter (on-system software) or available on BaseSpace (cloud-based software). The Metagenomics Workflow performs a taxonomic classification using the Greengenes database showing genus or species level classification in a graphical format.

This protocol can be used to sequence alternative regions of the 16S rRNA gene and for other targeted amplicon sequences of interest. When using this protocol for amplicon sequencing other than 16S rRNA, use the [Generate FASTQ Workflow](#) (secondary analysis option). For more information, see [MiSeq Reporter Metagenomics Workflow](#), on page 20.



### DISCLAIMER

The information in this Illumina Demonstrated Protocol is being provided as a courtesy; in some cases reagents are required to be purchased from non-authorized third-party suppliers. Illumina does not guarantee nor promises technical support for the performance of our products used with reagents purchased from a non-authorized third-party supplier.

**Figure 1** 16S V3 and V4 Amplicon Workflow

User-defined forward and reverse primers that are complementary upstream and downstream of the region of interest are designed with overhang adapters, and used to amplify templates from genomic DNA. A subsequent limited-cycle amplification step is performed to add multiplexing indices and Illumina sequencing adapters. Libraries are normalized and pooled, and sequenced on the MiSeq system using v3 reagents.

## Amplicon Primers

- The gene-specific sequences used in this protocol target the 16S V3 and V4 region. They are selected from the Klindworth et al. publication (Klindworth A, Pruesse E, Schweer T, Peplles J, Quast C, et al. (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* 41(1).) as the most promising bacterial primer pair. Illumina adapter

overhang nucleotide sequences are added to the gene-specific sequences. The full length primer sequences, using standard IUPAC nucleotide nomenclature, to follow the protocol targeting this region are:

16S Amplicon PCR Forward Primer = 5'  
TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG

16S Amplicon PCR Reverse Primer = 5'  
GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC

- This method can also be utilized to target other regions on the genome (either for 16S with other sets of primer pairs, or non-16S regions throughout the genome; ie any amplicon). The overhang adapter sequence must be added to the locus-specific primer for the region to be targeted (Figure 1). The Illumina overhang adapter sequences to be added to locus-specific sequences are:

Forward overhang: 5' TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-[locus-specific sequence]

Reverse overhang: 5' GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-[locus-specific sequence]

- The following considerations are recommended for designing other locus-specific primers:
  - a** Illumina recommends targeting regions that result in an amplicon that when sequenced with paired-end reads has at least ~50 bp of overlapping sequence in the middle. For example, if running 2x300 bp paired-end reads Illumina recommends having an insert size of 550 bp or smaller so that the bases sequenced at the end of each read overlap.
  - b** The locus-specific portion of primer (not including overhang sequence) must have a melting temperature ( $T_m$ ) of 60°–65°C. You can use online PCR primer sequence analysis tools (e.g. <http://www.idtdna.com/analyzer/Applications/OligoAnalyzer/>) to check the properties of primer designs. For the  $T_m$  calculation only, the gene-specific portion must be used in calculation. For hairpin and dimer calculations, the fully- assembled primer sequence (including the overhang) should be used.
  - c** Illumina recommends using standard desalting purification when ordering oligo primer sets.

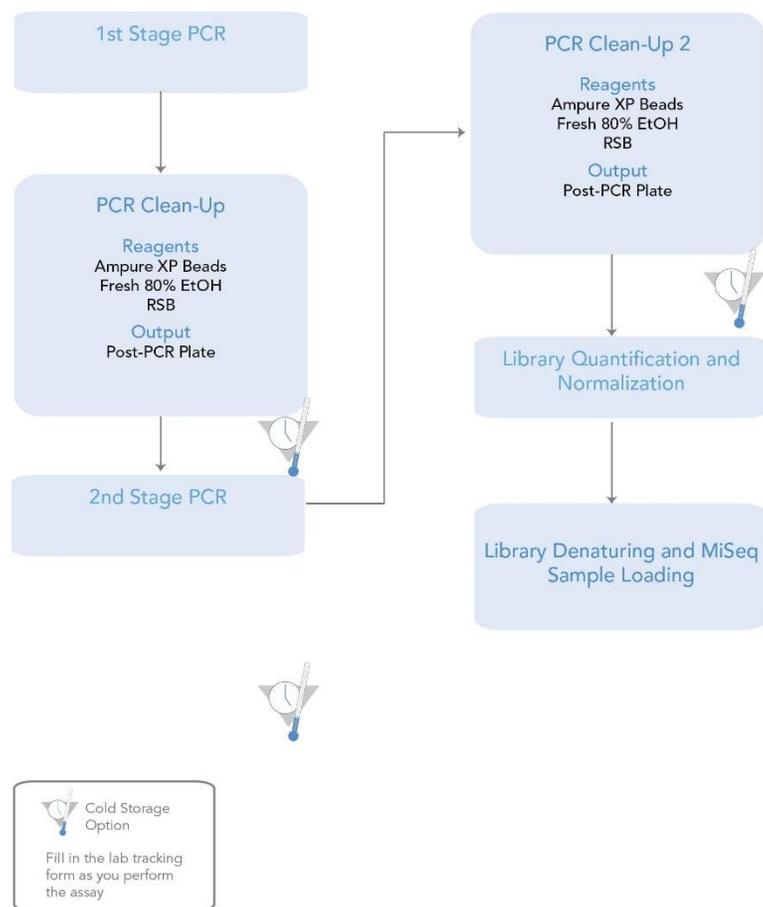
**NOTE**

For more information on reagents used in the protocol, see Consumables and Equipment, on page 21.

## 16S Library Preparation Workflow

The following diagram illustrates the workflow using the 16S Library Preparation Protocol. Safe stopping points are marked between steps.

**Figure 2** 16S Library Preparation Workflow



**Amplicon PCR**

Page 150

## Amplicon PCR

This step uses PCR to amplify template out of a DNA sample using region of interest- specific primers with overhang adapters attached. For more information on primer sequences, see Amplicon Primers, on page 3.

### Consumables

**NOTE**

For more information on consumables and equipment for this protocol see Consumables and Equipment, on page 21.

Item	Quantity	Storage
Microbial Genomic DNA (5 ng/ $\mu$ l in 10 mM Tris pH 8.5)	2.5 $\mu$ l per sample	-15° to -25°C
Amplicon PCR Reverse Primer (1 $\mu$ M)	5 $\mu$ l per sample	-15° to -25°C
Amplicon PCR Forward Primer (1 $\mu$ M)	5 $\mu$ l per sample	-15° to -25°C
2x KAPA HiFi HotStart ReadyMix	12.5 $\mu$ l per sample	-15° to -25°C
Microseal 'A' film		
96-well 0.2 ml PCR plate	1 plate	

[Optional] Bioanalyzer chip (Agilent DNA 1000 kit catalog # 5067-1504)

**Amplicon PCR**

Page 151

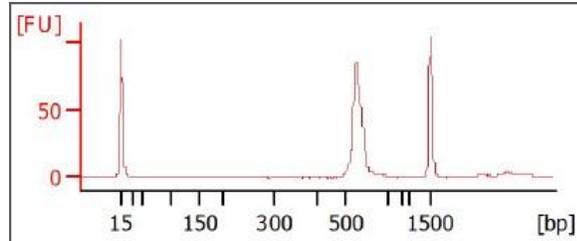
**Procedure**

- 1 Set up the following reaction of DNA, 2x KAPA HiFi HotStart ReadyMix, and primers:

	Volume
Microbial DNA (5 ng/ $\mu$ l)	2.5 $\mu$ l
Amplicon PCR Forward Primer 1 $\mu$ M	5 $\mu$ l
Amplicon PCR Reverse Primer 1 $\mu$ M	5 $\mu$ l
2x KAPA HiFi HotStart ReadyMix	12.5 $\mu$ l
<b>Total</b>	<b>25 <math>\mu</math>l</b>

- 2 Seal plate and perform PCR in a thermal cycler using the following program:
  - 95°C for 3 minutes
  - 25 cycles of:
    - 95°C for 30 seconds
    - 55°C for 30 seconds
    - 72°C for 30 seconds
  - 72°C for 5 minutes
  - Hold at 4°C
  
- 3 **[Optional]** Run 1  $\mu$ l of the PCR product on a Bioanalyzer DNA 1000 chip to verify the size. Using the V3 and V4 primer pairs in the protocol, the expected size on a Bioanalyzer trace after the Amplicon PCR step is ~550 bp.

**Figure 3** Example Bioanalyzer Trace after Amplicon PCR Step



**PCR Clean-Up**

Page 153

## PCR Clean-Up

This step uses AMPure XP beads to purify the 16S V3 and V4 amplicon away from free primers and primer dimer species.

### Consumables

Item	Quantity	Storage
10 mM Tris pH 8.5	52.5 µl per sample	-15° to -25°C
AMPure XP beads	20 µl per sample	2° to 8°C
Freshly Prepared 80% Ethanol (EtOH)	400 µl per sample	
96-well 0.2 ml PCR plate	1 plate	
[Optional] Microseal 'B' film		
[Optional] 96-well MIDI plate	1 plate	

### Preparation

- Bring the AMPure XP beads to room temperature.

### Procedure

- 1 Centrifuge the Amplicon PCR plate at 1,000 × g at 20°C for 1 minute to collect condensation, carefully remove seal.
- 2 **[Optional - for use with shaker for mixing]** Using a multichannel pipette set to 25 µl, transfer the entire Amplicon PCR product from the PCR plate to the MIDI plate. Change tips between samples.

**NOTE**

Transfer the sample to a 96-well MIDI plate if planning to use a shaker for mixing. If mixing by pipette, the sample can remain in the 96-well PCR plate.

## PCR Clean-Up

Page 154

- 3** Vortex the AMPure XP beads for 30 seconds to make sure that the beads are evenly dispersed. Add an appropriate volume of beads to a trough depending on the number of samples processing.
- 4** Using a multichannel pipette, add 20  $\mu$ l of AMPure XP beads to each well of the Amplicon PCR plate. Change tips between columns.
- 5** Gently pipette entire volume up and down 10 times if using a 96-well PCR plate or seal plate and shake at 1800 rpm for 2 minutes if using a MIDI plate.
- 6** Incubate at room temperature without shaking for 5 minutes.
- 7** Place the plate on a magnetic stand for 2 minutes or until the supernatant has cleared.
- 8** With the Amplicon PCR plate on the magnetic stand, use a multichannel pipette to remove and discard the supernatant. Change tips between samples.

- 9 With the Amplicon PCR plate on the magnetic stand, wash the beads with freshly prepared 80% ethanol as follows:
  - a Using a multichannel pipette, add 200  $\mu$ l of freshly prepared 80% ethanol to each sample well.
  - b Incubate the plate on the magnetic stand for 30 seconds.
  - c Carefully remove and discard the supernatant.
  
- 10 With the Amplicon PCR plate on the magnetic stand, perform a second ethanol wash as follows:
  - a Using a multichannel pipette, add 200  $\mu$ l of freshly prepared 80% ethanol to each sample well.
  - b Incubate the plate on the magnetic stand for 30 seconds.
  - c Carefully remove and discard the supernatant.
  - d Use a P20 multichannel pipette with fine pipette tips to remove excess ethanol.
  
- 11 With the Amplicon PCR plate still on the magnetic stand, allow the beads to air-dry for 10 minutes.
  
- 12 Remove the Amplicon PCR plate from the magnetic stand. Using a multichannel pipette, add 52.5  $\mu$ l of 10 mM Tris pH 8.5 to each well of the Amplicon PCR plate.
  
- 13 Gently pipette mix up and down 10 times, changing tips after each column (or seal plate and shake at 1800 rpm for 2 minutes). Make sure that beads are fully resuspended.
  
- 14 Incubate at room temperature for 2 minutes.
  
- 15 Place the plate on the magnetic stand for 2 minutes or until the supernatant has cleared.
  
- 16 Using a multichannel pipette, carefully transfer 50  $\mu$ l of the supernatant from the Amplicon PCR plate to a new 96-well PCR plate. Change tips between samples to avoid cross-contamination.

**SAFE STOPPING POINT**

If you do not immediately proceed to *Index PCR*, seal plate with Microseal “B” adhesive seal and store it at -15° to -25°C for up to a week.

**Index PCR**

Page 156

**Index PCR**

This step attaches dual indices and Illumina sequencing adapters using the Nextera XT Index Kit.

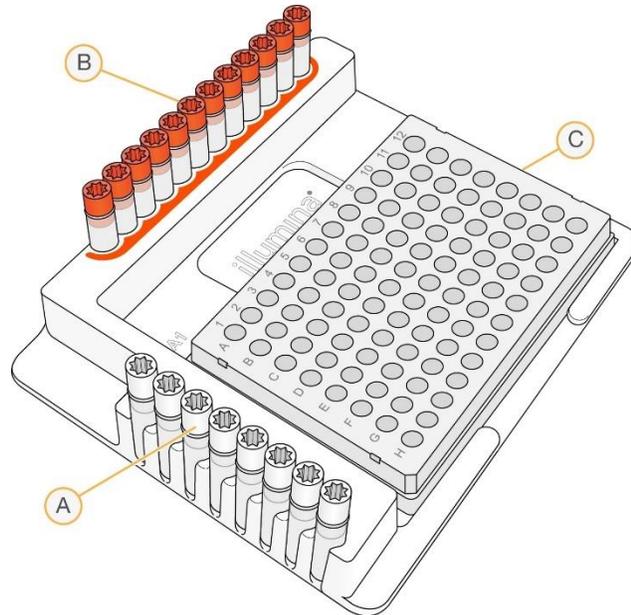
**Consumables**

<b>Item</b>	<b>Quantity</b>	<b>Storage</b>
2x KAPA HiFi HotStart ReadyMix	25 µl per sample	-15° to -25°C
Nextera XT Index 1 Primers (N7XX) from the Nextera XT Index kit (FC-131-1001 or FC-131-1002)	5 µl per sample	-15° to -25°C
Nextera XT Index 2 Primers (S5XX) from the Nextera XT Index kit (FC-131-1001 or FC-131- 1002)	5 µl per sample	-15° to -25°C
PCR Grade Water	10 µl per sample	
TruSeq Index Plate Fixture (FC-130-1005)	1	
96-well 0.2 ml PCR plate	1 plate	
Microseal 'A' film	1	

**Procedure**

- 1** Using a multichannel pipette, transfer 5 µl from each well to a new 96-well plate. The remaining 45 µl is not used in the protocol and can be stored for other uses.
- 2** Arrange the Index 1 and 2 primers in a rack (i.e. the TruSeq Index Plate Fixture) using the following arrangements as needed:
  - a** Arrange Index 2 primer tubes (white caps, clear solution) vertically, aligned with rows A through H.
  - b** Arrange Index 1 primer tubes (orange caps, yellow solution) horizontally, aligned with columns 1 through 12.

For more information on index selection, see Dual Indexing Principle, on page 23.

**Figure 4** TruSeq Index Plate Fixture

- A** Index 2 primers (white caps)
- B** Index 1 primers (orange caps)
- C** 96-well plate

- 3** Place the 96-well PCR plate with the 5  $\mu$ l of resuspended PCR product DNA in the TruSeq Index Plate Fixture.
- 4** Set up the following reaction of DNA, Index 1 and 2 primers, 2x KAPA HiFi HotStart ReadyMix, and PCR Grade water:

	Volume
DNA	5 $\mu$ l
Nextera XT Index Primer 1 (N7xx)	5 $\mu$ l
Nextera XT Index Primer 2 (S5xx)	5 $\mu$ l
2x KAPA HiFi HotStart ReadyMix	25 $\mu$ l
PCR Grade water	10 $\mu$ l
<b>Total</b>	<b>50 <math>\mu</math>l</b>

- 5 Gently pipette up and down 10 times to mix.
- 6 Cover the plate with Microseal 'A'.
- 7 Centrifuge the plate at 1,000  $\times$  g at 20°C for 1 minute.

**Index PCR**

*Page 159*

- 8** Perform PCR on a thermal cycler using the following program:
  - 95°C for 3 minutes
  - 8 cycles of:
    - 95°C for 30 seconds
    - 55°C for 30 seconds
    - 72°C for 30 seconds
  - 72°C for 5 minutes
  - Hold at 4°C

## PCR Clean-Up 2

This step uses AMPure XP beads to clean up the final library before quantification.

### Consumables

Item	Quantity	Storage
10 mM Tris pH 8.5	27.5 $\mu$ l per sample	-15° to -25°C
AMPure XP beads	56 $\mu$ l per sample	2° to 8°C
Freshly Prepared 80% Ethanol (EtOH)	400 $\mu$ l per sample	
96-well 0.2 ml PCR plate	1 plate	
[Optional] Microseal 'B' film		
[Optional] 96-well MIDI plate	1 plate	

### Procedure

- 1 Centrifuge the Index PCR plate at 280 × g at 20°C for 1 minute to collect condensation.
- 2 **[Optional - for use with shaker for mixing]** Using a multichannel pipette set to 50  $\mu$ l, transfer the entire Index PCR product from the PCR plate to the MIDI plate. Change tips between samples.



#### NOTE

Transfer the sample to a 96-well MIDI plate if planning to use a shaker for mixing. If mixing by pipette, the sample can remain in the 96-well PCR plate.

- 3 Vortex the AMPure XP beads for 30 seconds to make sure that the beads are evenly dispersed. Add an appropriate volume of beads to a trough.
- 4 Using a multichannel pipette, add 56  $\mu$ l of AMPure XP beads to each well of the Index PCR plate.

- 5** Gently pipette mix up and down 10 times if using a 96-well PCR plate or seal plate and shake at 1800 rpm for 2 minutes if using a MIDI plate.
- 6** Incubate at room temperature without shaking for 5 minutes.
- 7** Place the plate on a magnetic stand for 2 minutes or until the supernatant has cleared.
- 8** With the Index PCR plate on the magnetic stand, use a multichannel pipette to remove and discard the supernatant. Change tips between samples.
- 9** With the Index PCR plate on the magnetic stand, wash the beads with freshly prepared 80% ethanol as follows:
  - a** Using a multichannel pipette, add 200  $\mu$ l of freshly prepared 80% ethanol to each sample well.
  - b** Incubate the plate on the magnetic stand for 30 seconds.
  - c** Carefully remove and discard the supernatant.

**PCR Clean-Up 2**

Page 162

- 10** With the Index PCR plate on the magnetic stand, perform a second ethanol wash as follows:
  - a** Using a multichannel pipette, add 200  $\mu$ l of freshly prepared 80% ethanol to each sample well.
  - b** Incubate the plate on the magnetic stand for 30 seconds.
  - c** Carefully remove and discard the supernatant.
  - d** Use a P20 multichannel pipette with fine pipette tips to remove excess ethanol.
- 11** With the Index PCR plate still on the magnetic stand, allow the beads to air-dry for 10 minutes.
- 12** Remove the Index PCR plate from the magnetic stand. Using a multichannel pipette, add 27.5  $\mu$ l of 10 mM Tris pH 8.5 to each well of the Index PCR plate.
- 13** If using a 96-well PCR plate, gently pipette mix up and down 10 times until beads are fully resuspended, changing tips after each column. If using a MIDI plate, seal plate and shake at 1800 rpm for 2 minutes.
- 14** Incubate at room temperature for 2 minutes.
- 15** Place the plate on the magnetic stand for 2 minutes or until the supernatant has cleared.
- 16** Using a multichannel pipette, carefully transfer 25  $\mu$ l of the supernatant from the Index PCR plate to a new 96-well PCR plate. Change tips between samples to avoid cross-contamination.

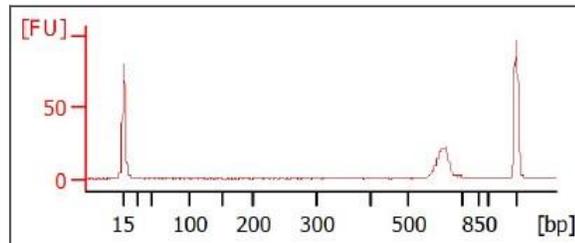
**SAFE STOPPING POINT**

If you do not plan to proceed to *Library Quantification, Normalization, and Pooling*, on page [16](#), seal the plate with Microseal "B" adhesive seal. Store the plate at -15° to -25°C for up to a week.

## [Optional] Validate Library

Run 1  $\mu\text{l}$  of a 1:50 dilution of the final library on a Bioanalyzer DNA 1000 chip to verify the size. Using the V3 and V4 primer pairs in the protocol, the expected size on a Bioanalyzer trace of the final library is  $\sim 630$  bp.

**Figure 5** Example Bioanalyzer Trace of Final Library



## Library Quantification, Normalization, and Pooling

Illumina recommends quantifying your libraries using a fluorometric quantification method that uses dsDNA binding dyes.

Calculate DNA concentration in nM, based on the size of DNA amplicons as determined by an Agilent Technologies 2100 Bioanalyzer trace:

$$\frac{(\text{concentration in ng/}\mu\text{l})}{(660 \text{ g/mol} \times \text{average library size})} \times 10^6 = \text{concentration in nM}$$

For example:

$$\frac{15 \text{ ng/}\mu\text{l}}{(660 \text{ g/mol} \times 500)} \times 10^6 = 45 \text{ nM}$$

Dilute concentrated final library using Resuspension Buffer (RSB) or 10 mM Tris pH 8.5 to 4 nM. Aliquot 5  $\mu\text{l}$  of diluted DNA from each library and mix aliquots for pooling libraries with unique indices. Depending on coverage needs, up to 96 libraries can be pooled for one MiSeq run.

For metagenomics samples, >100,000 reads per sample is sufficient to fully survey the bacterial composition. This number of reads allows for sample pooling to the maximum level of 96 libraries, given the MiSeq output of > 20 million reads.

## Library Denaturing and MiSeq Sample Loading

In preparation for cluster generation and sequencing, pooled libraries are denatured with NaOH, diluted with hybridization buffer, and then heat denatured before MiSeq sequencing. Each run must include a minimum of 5% PhiX to serve as an internal control for these low- diversity libraries. Illumina recommends using MiSeq v3 reagent kits for improved run metrics.

### Consumables

Item	Quantity	Storage
10 mM Tris pH 8.5 or RSB (Resuspension Buffer)	6 $\mu$ l	-15° to -25°C
HT1 (Hybridization Buffer)	1540 $\mu$ l	-15° to -25°C
0.2 N NaOH (less than a week old)	10 $\mu$ l	-15° to -25°C
PhiX Control Kit v3 (FC-110-3001)	4 $\mu$ l	-15° to -25°C
MiSeq reagent cartridge	1 cartridge	-15° to -25°C
1.7 ml microcentrifuge tubes (screw cap recommended)	3 tubes	

2.5 L ice bucket

## Preparation

- 1 Set a heat block suitable for 1.7 ml microcentrifuge tubes to 96°C
- 2 Remove a MiSeq reagent cartridge from -15°C to -25°C storage and thaw at room temperature.
- 3 In an ice bucket, prepare an ice-water bath by combining 3 parts ice and 1 part water.

## Denature DNA

- 1 Combine the following volumes of pooled final DNA library and freshly diluted 0.2 N NaOH in a microcentrifuge tube:
  - 4 nM pooled library (5  $\mu$ l)
  - 0.2 N NaOH (5  $\mu$ l)
- 2 Set aside the remaining dilution of 0.2 N NaOH to prepare a PhiX control within the next 12 hours.
- 3 Vortex briefly to mix the sample solution, and then centrifuge the sample solution at 280  $\times$  g at 20°C for 1 minute.
- 4 Incubate for 5 minutes at room temperature to denature the DNA into single strands.
- 5 Add the following volume of pre-chilled HT1 to the tube containing denatured DNA:
  - Denatured DNA (10  $\mu$ l)

- Pre-chilled HT1 (990  $\mu$ l)

Adding the HT1 results in a 20 pM denatured library in 1 mM NaOH.

- 6 Place the denatured DNA on ice until you are ready to proceed to final dilution.

### Dilute Denatured DNA

- 1 Dilute the denatured DNA to the desired concentration using the following example.

#### NOTE

Illumina recommends targeting 800–1000 K/mm<sup>2</sup> raw cluster densities using MiSeq v3 reagents. It is suggested to start your first run using a 4 pM loading concentration and adjust subsequent runs appropriately.

Final Concentration	2 pM	4 pM	6 pM	8 pM	10 pM
20 pM denatured library	60 $\mu$ l	120 $\mu$ l	180 $\mu$ l	240 $\mu$ l	300 $\mu$ l
Pre-chilled HT1	540 $\mu$ l	480 $\mu$ l	420 $\mu$ l	360 $\mu$ l	300 $\mu$ l

- 2 Invert several times to mix and then pulse centrifuge the DNA solution.
- 3 Place the denatured and diluted DNA on ice.

### Denature and Dilution of PhiX Control

Use the following instructions to denature and dilute the 10 nM PhiX library to the same loading concentration as the Amplicon library. The final library mixture must contain at least 5% PhiX.

- 1 Combine the following volumes to dilute the PhiX library to 4 nM:
  - 10 nM PhiX library (2  $\mu$ l)
  - 10 mM Tris pH 8.5 (3  $\mu$ l)

- 2** Combine the following volumes of 4 nM PhiX and 0.2 N NaOH in a microcentrifuge tube:
  - 4 nM PhiX library (5  $\mu$ l)
  - 0.2 N NaOH (5  $\mu$ l)
  
- 3** Vortex briefly to mix the 2 nM PhiX library solution.
  
- 4** Incubate for 5 minutes at room temperature to denature the PhiX library into single strands.
  
- 5** Add the following volumes of pre-chilled HT1 to the tube containing denatured PhiX library to result in a 20 pM PhiX library:
  - Denatured PhiX library (10  $\mu$ l)
  - Pre-chilled HT1 (990  $\mu$ l)
  
- 6** Dilute the denatured 20 pM PhiX library to the same loading concentration as the Amplicon library as follows:

Final Concentration	2 pM	4 pM	6 pM	8 pM	10 pM
20 pM denatured library	60 µl	120 µl	180 µl	240 µl	300 µl
Pre-chilled HT1	540 µl	480 µl	420 µl	360 µl	300 µl

- 7 Invert several times to mix and then pulse centrifuge the DNA solution.
- 8 Place the denatured and diluted PhiX on ice.

## Combine Amplicon Library and PhiX Control



### NOTE

The recommended PhiX control spike-in of  $\geq 5\%$  for low diversity libraries is possible with RTA v1.17.28 or later, which is bundled with MCS v2.2. For optimal performance, update to v3 software (MCS 2.3). If you are using an older version of the MiSeq software or sequencing these libraries on the GA or HiSeq, Illumina recommends using  $\geq 25\%$  PhiX control spike-in.

- 1 Combine the following volumes of denatured PhiX control library and your denatured amplicon library in a microcentrifuge tube:
  - Denatured and diluted PhiX control (30 µl)
  - Denatured and diluted amplicon library (570 µl)
- 2 Set the combined sample library and PhiX control aside on ice until you are ready to heat denature the mixture immediately before loading it onto the MiSeq v3 reagent cartridge.
- 3 Using a heat block, incubate the combined library and PhiX control tube at 96°C for 2 minutes.
- 4 After the incubation, invert the tube 1–2 times to mix and immediately place in the ice-water bath.
- 5 Keep the tube in the ice-water bath for 5 minutes.



**NOTE**

Perform the heat denaturation step immediately before loading the library into the MiSeq reagent cartridge to ensure efficient template loading on the MiSeq flow cell.

## MiSeq Reporter Metagenomics Workflow

After samples are loaded, the MiSeq system provides on-instrument secondary analysis using the MiSeq Reporter software (MSR). MSR provides several options for analyzing MiSeq sequencing data. For this demonstrated 16S protocol, select the Metagenomics workflow.

By following this 16S Metagenomics protocol, the Metagenomics workflow classifies organisms from your V3 and V4 amplicon using a database of 16S rRNA data. The classification is based on the Greengenes database (<http://greengenes.lbl.gov/>). The output of this workflow is a classification of reads at several taxonomic levels: kingdom, phylum, class, order, family, genus, and species. The analysis output includes:

- Clusters Graph – shows numbers of raw cluster, clusters passing filter, clusters that did not align, clusters not associated with an index, and duplicates.
- Sample Table – summarizes the sequencing results for each sample.
- Cluster Pie Chart – a graphical representation of the classification breakdown for each sample.

See the *MiSeq Reporter Metagenomics Workflow – Reference Guide* (Part # 15042317) for detailed instructions and guidance.

The method described in this 16S Metagenomics protocol can be used for any targeted amplicon sequencing, relevant to virus research, mutation detection, or other microbiology- related studies. If you use the protocol for other targeted amplicon sequencing studies, select the MiSeq Reporter Generate FASTQ Workflow for on-instrument generation of FASTQ files for downstream analysis. For specific guidance on the Generate FASTQ Workflow, see the *MiSeq Reporter Generate FASTQ Workflow – Reference Guide* (Part # 15042322).

## Supporting Information

The protocols described in this guide assume that you are familiar with the contents of this section and have obtained all of the requisite equipment and consumables.

### Acronyms

**Table 1** Acronyms

Acronym	Definition
HT1	Hybridization Buffer
IEM	Illumina Experiment Manager
MSR	MiSeq Reporter
PCR	Polymerase Chain Reaction
rRNA	Ribosomal RNA
RSB	Resuspension Buffer

### Consumables and Equipment

Check to make sure that you have all of the necessary user-supplied consumables and equipment before proceeding to sample preparation.

**Table 2** User-Supplied Consumables

Consumable	Supplier
1.7 ml microcentrifuge tubes	General lab supplier
10 µl barrier pipette tips	General lab supplier

**Support Information: Genomics Workflow***Page 173*

10 µl multichannel pipettes	General lab supplier
10 µl single channel pipettes	General lab supplier
20 µl barrier pipette tips	General lab supplier
20 µl multichannel pipettes	General lab supplier
20 µl single channel pipettes	General lab supplier
200 µl barrier pipette tips	General lab supplier
200 µl multichannel pipettes	General lab supplier
200 µl single channel pipettes	General lab supplier
1000 µl barrier pipette tips	General lab supplier

Consumable	Supplier
1000 µl multichannel pipettes	General lab supplier
1000 µl single channel pipettes	General lab supplier
96-well 0.2 ml skirtless PCR plates or Twin.Tec 96-well PCR plates	Bio-Rad, part # MSP-9601
Agencourt AMPure XP 60 ml kit	Beckman Coulter Genomics, part # A63881
Ethanol 200 proof (absolute) for molecular biology (500 ml)	Sigma-Aldrich, part # E7023
Amplicon PCR Forward Primer (Standard desalting)	
Amplicon PCR Reverse Primer (Standard desalting)	
KAPA HiFi HotStart ReadyMix (2X)	KAPA Biosystems, part # KK2601
Microseal 'A' adhesive seals	Bio-Rad, part # MSA-5001
Microseal 'B' adhesive seals	Bio-Rad, part # MSB-1001
MiSeq Reagent Kit v3 (600 cycle)	Illumina, catalog # MS-102-3003
Nextera XT Index Kit	Illumina, catalog # FC-131-1001 or Illumina, catalog # FC-131-1002
PhiX Control Kit v3	Illumina, catalog # FC-110-3001
PCR grade water	General lab supplier
Fluorometric quantitation with dsDNA binding dye reagents	General lab supplier

RNase/DNase-free 8-well PCR strip tubes and caps      General lab supplier

RNase/DNase-free multichannel reagent reservoirs, disposable      VWR, part # 89094-658

Tris-HCl 10 mM, pH 8.5      General lab supplier

[Optional] 96-well storage plates, round well, 0.8 ml      Fisher Scientific, part # AB-0859

("MIDI" plate)

**Table 3** User-Supplied Equipment

Equipment	Supplier
2.5 L ice bucket	General lab supplier
96-well thermal cycler (with heated lid)	General lab supplier

Equipment	Supplier
Fluorometer for quantitation with  dsDNA binding dyes	General lab supplier
Magnetic stand-96	Life Technologies, catalog # AM10027
Microplate centrifuge	General lab supplier
TruSeq Index Plate Fixture Kit (reusable)	Illumina, catalog # FC-130-1005
[Optional] 2100 Bioanalyzer Desktop System	Agilent, part # G2940CA
[Optional] Agilent DNA 1000 Kit	Agilent, part # 5067-1504
[Optional] High Speed Micro Plate Shaker	VWR, catalog # 13500-890 (110V/120V)  or  VWR, catalog # 14216-214 (230V)

## Dual Indexing Principle

The dual indexing strategy uses two 8 base indices, Index 1 (i7) adjacent to the P7 sequence, and Index 2 (i5) adjacent to the P5 sequence. Dual indexing is enabled by adding a unique Index 1 (i7) and Index 2 (i5) to each sample. The 96 sample Nextera XT Index Kit (FC-131–

1002) use 12 different Index 1 (i7) adapters (N701–N712) and 8 different Index 2 (i5) adapters (S501–S508). The 24 sample Nextera XT Index Kit (FC-131–1001) uses 6 different Index 1 (i7) adapters (N701–N706) and 4 different Index 2 (i5) adapters (S501–S504). In the Index adapter name, the N or S refers to Nextera XT sample preparation, 7 or 5 refers to Index 1 (i7) or Index 2 (i5), respectively. The 01–12 refers to the Index number. A list of index sequences is provided for generating sample sheets to demultiplex the samples:

Index 1 (i7)	Sequence	Index 2 (i5)	Sequence N701
TAAGGCGA	S501	TAGATCGC N702	CGTACTAG
S502	CTCTCTAT N703	AGGCAGAA	S503
TATCCTCT N704	TCCTGAGC	S504	AGAGTAGA
N705	GGACTCCT	S505	GTAAGGAG N706
TAGGCATG	S506	ACTGCATA N707	CTCTCTAC
S507	AAGGAGTA N708	CAGAGAGG	S508
CTAAGCCT N709	GCTACGCT		
N710	CGAGGCTG N711		
AAGAGGCA	N712		
GTAGAGGA			

## Low Plexity Pooling Guidelines

Illumina uses a green laser or LED to sequence G/T and a red laser or LED to sequence A/C. At each cycle, at least one of two nucleotides for each color channel are read to ensure proper registration. It is important to maintain color balance for each base of the index read being sequenced, otherwise index read sequencing could fail due to registration failure. If you choose the dual-indexed sequencing workflow, always use at least two unique and

compatible barcodes for each index (index 1 and index 2). The following tables illustrate possible pooling strategies:

**Table 4** Libraries Pooled: 6 or fewer; Sequencing Workflow: Single Index

Plex	Index 1 (i7) Selection	Index 2 (i5) Selection
1-plex (no pooling)	Any Index 1 adapter	Any Index 2 adapter
2-plex	<ul style="list-style-type: none"> <li>• [option 1] N702 and N701</li> <li>• [option 2] N702 and N704</li> </ul>	
3-plex	<ul style="list-style-type: none"> <li>• [option 1] N701, N702, and N704</li> <li>• [option 2] N703, N705, and N706</li> </ul>	
4- or 5-plex	<ul style="list-style-type: none"> <li>• [option 1] N701, N702, N704, and any other Index 1 adapter</li> <li>• [option 2] N703, N705, N706, and any other Index 1 adapter</li> </ul>	
6-plex	N701, N702, N703, N704, N705, and N706	

**Table 5** Sequencing Workflow: Single or Dual Index

Plex	Index 1 (i7) Selection	Index 2 (i5) Selection
7–12 plex, Dual Index	<ul style="list-style-type: none"> <li>• [option 1] N701, N702, N704, and any other Index 1 adapter (as needed)</li> <li>• [option 2] N703, N705, N706, and any other Index 1 adapter (as needed)</li> </ul>	<ul style="list-style-type: none"> <li>• [option 1] S501 and S502</li> <li>• [option 2] S503 and S504</li> <li>• [option 3] S505 and S506</li> </ul>
7–12 plex, Single Index (96 sample Nextera Index adapter kit)	• N701–N706 and any other Index 1 adapter (as needed)	• Any Index 2 (i5) adapter
Greater than 12-plex		

- N701, N702, N703, N704, N705, N706, and any other Index 1 adapter
- [option 1] S501, S502, and any other Index 2 adapter (as needed)
  - [option 2] S503, S504, and any other Index 2 adapter (as needed)
  - [option 3] S505, S506, and any other Index 2 adapter (as needed)

These strategies represent only some of the acceptable combinations. Alternatively, check the real sequences of each index in the tables to make sure that each base position has a signal in both color channels for the index read:

Good				Bad			
Index 1		Index 2		Index 1		Index 2	
705	GGACTCCT	503	TATCCTCT	705	GGACTCCT	502	CTCTCTAT
706	TAGGCATG	503	TATCCTCT	706	TAGGCATG	502	CTCTCTAT
701	TAAGGCGA	504	AGAGTAGA	701	TAAGGCGA	503	TATCCTCT
702	CGTACTAG	504	AGAGTAGA	702	CGTACTAG	503	TATCCTCT
v v v v v v v v		v v v v v v v v		v v v v v v v v		v v v v xxxx	

v=signal in both color

x=signal missing in one color channel

## Prevent PCR Product Contamination

The PCR process is commonly used in the laboratory to amplify specific DNA sequences. Unless proper laboratory hygiene is used, PCR products can contaminate reagents, instrumentation, and genomic DNA samples, causing inaccurate and unreliable results. PCR product contamination can shut down lab processes and significantly delay normal operations.

Make sure that the lab is set up appropriately to reduce the risk of PCR product contamination:

- **Physically Separate Pre-PCR and Post-PCR Areas**
  - Physically separate laboratory space where pre-PCR processes are performed (DNA extraction, quantification, and normalization) from the laboratory space where PCR products are made and processed (post-PCR processes).
  - Never use the same sink to wash pre-PCR and post-PCR troughs.
  - Never share water purification systems for pre-PCR and post-PCR processes.
  - Store all supplies used in the protocols in the pre-PCR area, and transfer to the post-PCR area as needed.
- **Use Dedicated Equipment and Supplies**
  - Dedicate separate full sets of equipment and supplies (pipettes, centrifuges, oven, heat block, etc.) to pre-PCR and post-PCR lab processes, and never share between processes.

- Dedicate separate storage areas (freezers and refrigerators) to pre-PCR and post-PCR consumables.

Because the pre- and post-amplification reagents are shipped together, it is important to unpack the reagents in the pre-PCR lab area. After unpacking the reagents, move the post- amplification reagents to the proper post-PCR storage area.

## Pre-PCR and Post-PCR Lab Procedures

To prevent PCR product contamination, it is important to establish lab procedures and follow best practices. Illumina recommends daily and weekly cleaning of lab areas using

0.5% Sodium Hypochlorite (10% Bleach).



#### CAUTION

To prevent sample or reagent degradation, make sure that all vapors from the cleaning solution have fully dissipated before beginning any processes.

## Daily Cleaning of Pre-PCR Area

A daily cleaning of the pre-PCR area using a 0.5% Sodium Hypochlorite (10% Bleach)

solution helps to eliminate PCR product that has entered the pre-PCR area.

Identify pre-PCR areas that pose the highest risk of contamination, and clean these areas with a 0.5% Sodium Hypochlorite (10% Bleach) solution before beginning any pre-PCR processes. High-risk areas might include, but are not limited to, the following items:

- Benchtops
- Door handles
- Refrigerator/freezer door handles
- Computer mouse
- Keyboards

## Daily Cleaning of Post-PCR Area

Reducing the amount of PCR product in the post-PCR area helps reduce the risk of contamination in the pre-PCR area. Daily cleaning of the post-PCR area using a 0.5% Sodium Hypochlorite (10% Bleach) solution helps reduce the risk of contamination.

Identify post-PCR areas that pose the highest risk of contamination, and clean these areas with a 0.5% Sodium Hypochlorite (10% Bleach) solution daily. High-risk areas might include, but are not limited to, the following items:

- Thermal cyclers
- Bench space used to process amplified DNA
- Door handles
- Refrigerator/freezer door handles
- Computer mouse
- Keyboards

## Weekly Cleaning of All Lab Areas

One time a week, perform a thorough cleaning of the pre-PCR and post-PCR areas using

0.5% Sodium Hypochlorite (10% Bleach).

- Clean all benchtops and laboratory surfaces.
- Clean all instruments that are not cleaned daily.
- Thoroughly mop lab floors.
- Make sure that personnel responsible for weekly cleaning are properly trained on prevention of PCR product contamination.

### Items Fallen to the Floor

The floor is contaminated with PCR product transferred on the shoes of individuals coming from the post-PCR area; therefore, anything falling to the floor must be treated as contaminated.

- Disposable items that have fallen to the floor, such as empty tubes, pipette tips, gloves, lab coat hangers, must be discarded.

- Non-disposable items that have fallen to the floor, such as a pipette or an important sample container, must be immediately and thoroughly cleaned. Use a 0.5% Sodium Hypochlorite (10% Bleach) solution to remove PCR product contamination.
- Clean any lab surface that has come in contact with the contaminated item. Individuals handling anything that has fallen to the floor, disposable or non-disposable, must discard their lab gloves and put on a new pair.

## Best Practices

When preparing libraries for sequencing, always adhere to good molecular biology practices. Read through the entire protocol before starting to make sure that all of the required materials are available and your equipment is programmed and ready to use.

## Handling Liquids

Good liquid handling measures are essential, particularly when quantifying libraries or diluting concentrated libraries for making clusters.

- Small differences in volumes ( $\pm 0.5 \mu\text{l}$ ) can sometimes cause large differences in cluster numbers ( $\sim 100,000$ ).
- Small volume pipetting can be a source of potential error in protocols requiring the generation of standard curves, such as qPCR, or small but precise volumes, such as the Agilent Bioanalyzer.
- If small volumes are unavoidable, use due diligence to make sure that pipettes are correctly calibrated.
- Make sure that pipettes are not used at the volume extremes of their performance specifications.
- Prepare the reagents for multiple samples simultaneously, to minimize pipetting errors, especially with small volume enzyme additions. As a result, pipette one time from the reagent tubes with a larger volume, rather than many times with small volumes. Aliquot to individual samples in a single pipetting movement to allow for standardization across multiple samples.

## Handling Magnetic Beads



### NOTE

Cleanup procedures have only been validated using the 96-well plates and the magnetic stand specified in the *Consumables and Equipment* list. Comparable performance is not guaranteed when using a microcentrifuge tube or other formats, or other magnets.

- Before use, allow the beads to come to room temperature.
- Do not reuse beads. Always add fresh beads when performing these procedures.

- Immediately before use, vortex the beads until they are well dispersed and the color of the liquid is homogeneous.
- When pipetting the beads, pipette slowly and dispense slowly due to the viscosity of the solution.
- Take care to minimize bead loss, which can affect final yields.
- Change the tips for each sample, unless specified otherwise.
- Let the mixed samples incubate at room temperature for the time indicated in the protocol for maximum recovery.

- When removing and discarding supernatant from the wells, use a single channel or multichannel pipette and take care not to disturb the beads.
- When aspirating the cleared solution from the reaction plate and wash step, it is important to keep the plate on the magnetic stand and not disturb the separated magnetic beads. Aspirate slowly to prevent the beads from sliding down the sides of the wells and into the pipette tips.
- To prevent the carryover of beads after elution, approximately 2.5  $\mu\text{l}$  of supernatant is left when the eluates are removed from the bead pellet.
- Be sure to remove all of the ethanol from the bottom of the wells, as it can contain residual contaminants.
- Keep the reaction plate on the magnetic stand and let it air-dry at room temperature to prevent potential bead loss due to electrostatic forces. Allow for the complete evaporation of residual ethanol, because the presence of ethanol affects the performance of the subsequent reactions. Illumina recommends at least minutes drying time, but a longer drying time can be required. Remaining ethanol can be removed with a 10  $\mu\text{l}$  pipette.
- Avoid over drying the beads, which can impact final yields.
- Do not scrape the beads from the edge of the well using the pipette tip.
- To maximize sample recovery during elution, incubate the sample/bead mix for  
2 minutes at room temperature before placing the samples onto the magnet.

## Avoiding Cross-Contamination

Practice the following to avoid cross-contamination:

- Open only one adapter tube at a time.
- Change the tips for each sample, unless specified otherwise.
- Pipette carefully to avoid spillage.
- Clean pipettes and change gloves between handling different adapter stocks.
- Clean work surfaces thoroughly before and after the procedure.

## Potential DNA Contaminants

When handling and processing samples using this protocol, use best practices to avoid PCR contamination, as you would when preparing PCR amplicons.

## Temperature Considerations

Temperature is an important consideration for making libraries:

- Keep libraries at temperatures  $\leq 37^{\circ}\text{C}$ , except where specifically noted.
- Place reagents on ice after thawing at room temperature.

## Equipment

- Review the programming instructions for your thermal cycler user guide to make sure that it is programmed appropriately using the heated lid function.
- It is acceptable to use the thermal cycler tracked heating lid function.

## Addendum 3

Table1 Nextera indexing sequences

<b>Index 1 (i7)</b>	<b>Sequence</b>	<b>Index 2 (i5)</b>	<b>Sequence</b>
<b>N701</b>	TAAGGCGA	S517	AGATCGC
<b>N702</b>	CGTACTAG	S502	CTCTCTAT
<b>N703</b>	AGGCAGAA	S503	TATCCTCT
<b>N704</b>	TCCTGAGC	S504	AGAGTAGA
<b>N705</b>	GGA CTCCT	S505	GTAAGGAG
<b>N706</b>	TAGGCATG	S506	ACTGCATA
<b>N707</b>	CTCTCTAC	S507	AAGGAGTA
<b>N708</b>	CAGAGAGG	S508	CTAAGCCT
<b>N709</b>	GCTACGCT		
<b>N710</b>	CGAGGCTG		
<b>N711</b>	AAGAGGCA		
<b>N712</b>	GTAGAGGA		

## Addendum 4

Index design in a 96 well plate format

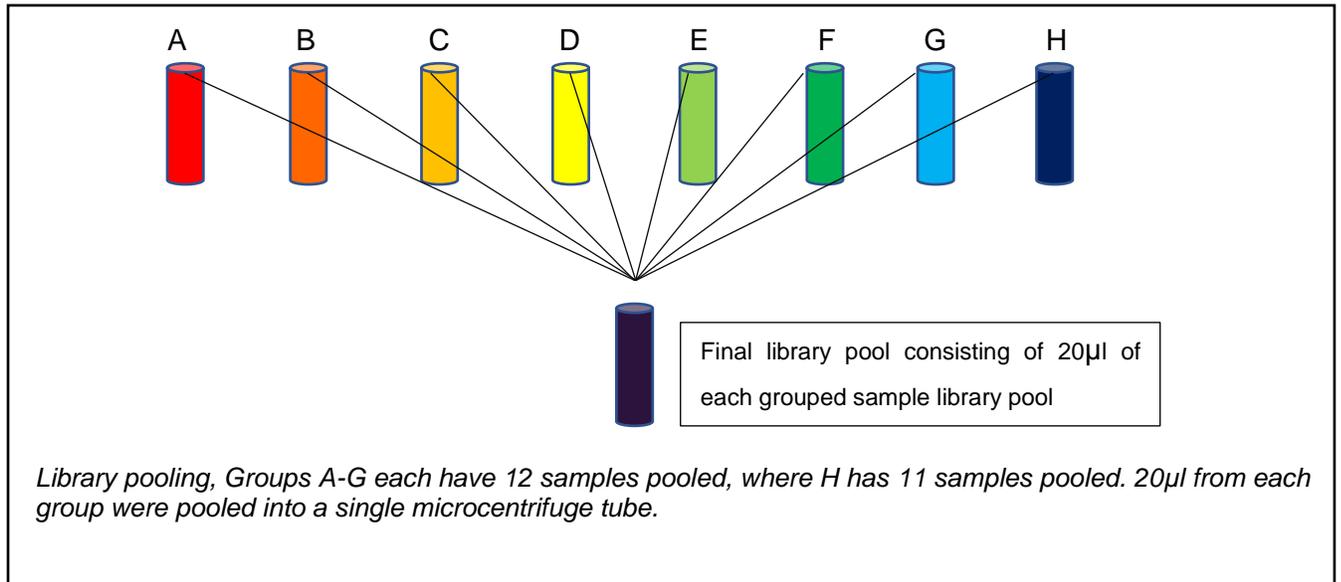
Index 1 (N7)

→

Index 2 (S5-S517-508)

	1	2	3	4	5	6	7	8	9	10	11	12
A	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10	Sample 11	Sample 12
B	Sample 13	Sample 14	Sample 15	Sample 16	Sample 17	Sample 18	Sample 19	Sample 20	Sample 21	Sample 22	Sample 23	Sample 24
C	Control 1 EB1	control 2 NEC1	Control 3 NTC	Control 4 VTN	control 5 H <sub>2</sub> O	Sample 30	Sample 31	Sample 32	Sample 33	Sample 34	Sample 35	Sample 36
D	Sample 37	Sample 38	Sample 39	WRR	Sample 41	Sample 42	Sample 43	Sample 44	Sample 45	Sample 46	Sample 47	Sample 48
E	Sample 49	Sample 50	Sample 51	Sample 52	Sample 53	Control 6EB2	Control 7 NEC2	Control 8 NTC2	Control 9 WRR	Control 10 TRIS	Sample 59	Sample 60
F	Sample 61	Sample 62	Sample 63	Sample 64	Sample 65	Sample 66	Sample 67	Sample 68	Sample 69	Sample 70	Sample 71	Sample 72
G	Sample 73	Sample 74	Sample 75	Sample 76	Sample 77	Sample 78	Sample 79	Sample 80	sample 81	sample 82	sample 83	sample 84
H	sample 85	Sample 86	Sample 87	Control 11 EB3	Control 12 NEC3	Control 13 EB4	Control 14 NEC4	Control 15 NTC3	Control 16 MCKEV	Control 17 MCKEC8	Control 18PC	Blank

## Addendum 5



## Addendum 6

15/03/2019 Generate FASTQ from an Incomplete Run in MiSeq Reporter

## Generate FASTQ from an incomplete run in MiSeq Reporter

Back

Jun 7, 2018

It is possible to generate FASTQ files, even if a MiSeq run does not complete. To do so, determine the last cycle with a complete set of \*.bcl files, modify run files, and requeue analysis. This bulletin describes how to generate FASTQ files from an incomplete run using MiSeq Reporter.

Before you start:

- If you plan to do secondary analysis on the MiSeq, make sure that a sequencing run is not in progress and that MiSeq Reporter is ready for analysis.
- If you opted for BaseSpace, it is not possible to generate FASTQ files from a partial run and save the analysis in BaseSpace. The analysis must be performed locally.
- If the run stopped before completing the index reads (ie, in Read 1), it is not possible to demultiplex your samples.
- Make sure that the **RTACComplete.txt** file is **not** present in the run folder.

Steps to take:

1. Go to `D:\Illumina\MiSeqAnalysis\runfolder\Data\Intensities\BaseCalls\L001` and check which cycle was the last to generate all \*.bcl files. If v3 reagents were used, there are 38 \*.bcl files per cycle. If v2 reagents were used, there are 28 \*.bcl files per cycle.
 

*\*If .bcl files are not present, it might be possible to recover them by manually launching RTA. To do so, contact Illumina Technical Support.*

Calculate how many cycles have been successfully completed and extracted.  
*For example:*  
 I stopped at cycle 207 in a  $2 \times 150 + 6$  bp index run. If the last cycle with complete \*.bcl files was cycle 206, it is a good rule of thumb not to push against the last cycle with complete \*.bcl files. Select the previous complete cycle from which to do the analysis. In this case, it is cycle 205. The run breakdown is (R1) 150 + (R2 Index) 6 + (R3) 49 = 205 cycles.

$$300 + 8 + 8 + 274 = 590$$

$$590$$
2. Backup the original **RunInfo.xml** and **SampleSheet.csv** files by renaming them to something like **RunInfo.xml.BAK** and **SampleSheet.csv.BAK**. Using Notepad, edit these documents and save the edited documents as **RunInfo.xml** and **SampleSheet.csv** in the root run folder in MiSeqAnalysis. If the reverse read of a paired-end run – Read 3 in the following example – has not been performed, delete the corresponding line in **RunInfo.xml** and **SampleSheet.csv**.
 

*For Example:*

**RunInfo.xml:**

```
<Reads>
  <Read Number="1" IsIndexedRead="N" NumCycles="150"/>
  <Read Number="2" IsIndexedRead="Y" NumCycles="6"/>
  <Read Number="3" IsIndexedRead="N" NumCycles="49"/>
</Reads>
```

**SampleSheet.csv:**

	(reads)
11	150
12	49
13	
3. Copy an **RTACComplete.txt** file from a successful run folder and paste it into your run folder.
4. You can expect to see the run queued in MiSeq Reporter and a **QueuedForAnalysis.txt** file created in the run folder in MiSeqAnalysis. To make sure that the run has been requeued in MiSeq Reporter, point a web browser window to <http://localhost:8042> and look for your run in the list opened with the **Analyses** button.

**Innovative technologies**

At Illumina, our goal is to apply innovative technologies to the analysis of genetic variation and function, making studies possible that were not even imaginable just a few years ago. It is mission critical for us to deliver innovative, flexible, and scalable solutions to meet the needs of our customers. As a global company that places high value on collaborative interactions, rapid delivery of solutions, and providing the highest level of quality, we strive to meet this challenge. Illumina innovative sequencing and array technologies are fueling groundbreaking advancements in life science research, translational and consumer genomics, and molecular diagnostics.

<https://support.illumina.com/bulletins/2016/10/generate-fastq-from-an-incomplete-run-in-miseq-reporter.html> 1/2

## Addendum 7

### Commands: Qiime and R

#QIIME input commands

#files demultiplexed at UWC

#Demultiplexing single read illumina Amplicon libraries in QIIME

#commands are boxed

```
$qiime>multiple_split_libraries_fastq.py-i"/filelocation"/single_read-  
o"/filelocation"/sampleid_by_file--read_indicator_R1
```

#This script runs split\_libraries\_fastq.py on data that are already demultiplexed (split up according to sample, with one sample per file). The script supports the following types of input: a directory containing many files, where each file is named on a per-sample basis (with different prefixes before the read number), a directory containing many directories, where each directory is named on a per-sample basis

```
$qiime > pick_open_reference_otus.py -i seqs.fna -r ./greengenes/rep_set/97_otus.fasta -o  
./openpick_otu
```

#This script is broken down into 4 possible OTU picking steps, and 2 steps involving the creation of OTU tables and trees. The commands for each step are described below, including what the input and resulting output files are. Additionally, the optional specified parameters of this script that can be passed are referenced. #Step 1) Prefiltering and picking closed reference OTUs The first step is an optional prefiltering of the input fasta file to remove sequences that do not hit the reference database with a given sequence identity (PREFILTER\_PERCENT\_ID). This step can take a very long time, so is disabled by default. The prefilter parameters can be changed with the options: `-prefilter_refseqs_fp` `-prefilter_percent_id` This filtering is accomplished by picking closed reference OTUs at the specified prefilter percent id to produce: `prefilter_otus/seqs_otus.log` `prefilter_otus/seqs_otus.txt` `prefilter_otus/seqs_failures.txt` `prefilter_otus/seqs_clusters.uc` Next, the `seqs_failures.txt` file is used to remove these failed sequences from the original input fasta file to produce: `prefilter_otus/prefiltered_seqs.fna` This `prefiltered_seqs.fna` file is then considered to contain the reads of the marker gene of interest, rather than spurious reads such as host genomic sequence or sequencing artifacts. #If prefiltering is applied, this step progresses with the `prefiltered_seqs.fna`. Otherwise it progresses with the input file. The Step 1 closed reference OTU picking is done against the supplied reference database. This command produces: `step1_otus/_clusters.uc` `step1_otus/_failures.txt` `step1_otus/_otus.log` `step1_otus/_otus.txt`

#The representative sequence for each of the Step 1 picked OTUs are selected to produce: `step1_otus/step1_rep_set.fna`



#If `--min_otu_size 10` was passed, it would produce: `final_otu_map_mc10.txt`

#The `final_otu_map_mc2.txt` is used to build the final representative set: `rep_set.fna`

#Step 6) Making the OTU tables and trees An OTU table is built using the `final_otu_map_mc2.txt` file to produce: `otu_table_mc2.biom`

#As long as the `--suppress_taxonomy_assignment` flag is NOT passed, then taxonomy will be assigned to each of the representative sequences in the final `rep_set` produced in Step 5, producing: `rep_set_tax_assignments.log` `rep_set_tax_assignments.txt` This taxonomic metadata is then added to the `otu_table_mc2.biom` to produce: `otu_table_mc_w_tax.biom`

#As long as the `--suppress_align_and_tree` is NOT passed, then the `rep_set.fna` file will be used to align the sequences and build the phylogenetic tree, which includes the de novo OTUs. Any sequences that fail to align are omitted from the OTU table and tree to produce: `otu_table_mc_no_pynast_failures.biom` `rep_set.tre`

#If both `--suppress_taxonomy_assignment` and `--suppress_align_and_tree` are NOT passed, the script will produce: `otu_table_mc_w_tax_no_pynast_failures.biom`

#It is important to remember that with a large workflow script like this that the user can jump into intermediate steps. For example, imagine that for some reason the script was interrupted on Step 2, and the user did not want to go through the process of re-picking OTUs as was done in Step 1. They can simply rerun the script and pass in the: `--step_1_otu_map_fp` `--step1_failures_fasta_fp` parameters, and the script will continue with Steps 2 - 4. [http://qiime.org/scripts/pick\\_open\\_reference\\_otus.html](http://qiime.org/scripts/pick_open_reference_otus.html)

##Output files loaded into R studio

---

---

#R script Data analysis

#Taxonomic classification

```
setwd("E:/Qiime files/")  
library(biomformat)  
library(microbiome)  
library(readr)  
library(phyloseq)  
library(dplyr)
```

#these steps are for importing data and renaming the merged files which are used in subsequent steps

```
biom_file <- paste("sorted_otu_w_tax.biom", sep = "")  
map_file <- paste("mapping_finalnew2.txt", sep = "")
```

```
# Now import the .biom-formatted otu_table-tax_table file.
```

```
biom_otu_tax <- import_biom(biom_file)
```

```
# Add sample data to the dataset using merge
```

```
map <- import_qiime_sample_data(map_file)
```

```
# merge the two into a single phyloseq object containing the otu info and sample metadata
```

```
merged <- merge_phyloseq(biom_otu_tax, map)
```

```
#renames taxonomy shorter names
```

```
colnames(tax_table(merged))=c("Domain", "Phylum", "Class", "Order", "Family", "Genus", "OTU")
```

```
#lists names of rows and columns
```

```
rownames(sample_data(merged))
```

```
colnames(sample_data(merged))
```

```
treefile = "rep_set.tre"
```

```
tree.obj = import_qiime(treefilename = treefile)
```

```
otu.table = merge_phyloseq(merged, tree.obj)
```

```
#summary information on dataset used
```

```
#used phyloseq object which was the file called "merged"
```

```
summarize_phyloseq(merged)
```

```
# do other things as in phyloseq_analysis.R script
```

```
mergedtaxonomy <- tax_table(merged)
```

```
mergedranks <- rank_names(merged)
```

```
meta <- meta(merged)
```

```
#####
```

```
##To SUBSET data
```

```
# e.g. by Type == "SAMPLE" or Time_point == "Baseline" etc.
```

```
sub_base <- merged %>% subset_samples(Time_point == "Baseline") %>% prune_taxa(taxa_sums(.) > 0, .)
```

```
sub_base_case <- merged %>% subset_samples(Time_point == "Baseline" & TB_treatment == "CASE") %>% prune_taxa(taxa_sums(.) > 0, .)
```

```
sub_base_control <- merged %>% subset_samples(Time_point == "Baseline" & TB_treatment
=="CONTROL") %>% prune_taxa(taxa_sums(.) > 0, .)
```

```
#####month2 subset
```

```
sub_month2 <-merged %>% subset_samples(Time_point == "month2") %>% prune_taxa(taxa_sums(.) >
0, .)
```

```
sub_month2_case <- merged %>% subset_samples(Time_point == "month2" & TB_treatment == "CASE")
%>% prune_taxa(taxa_sums(.) > 0, .)
```

```
sub_month2_control <- merged %>% subset_samples(Time_point == "month2" & TB_treatment
=="CONTROL") %>% prune_taxa(taxa_sums(.) > 0, .)
```

```
#####month6 subset
```

```
#split the cases into clinically diagnosed and bacteriologically confirmed
```

```
sub_month6 <-merged %>% subset_samples(Time_point == "month6") %>% prune_taxa(taxa_sums(.) >
0, .)
```

```
sub_month6_case <- merged %>% subset_samples(Time_point == "month6" & TB_treatment == "CASE")
%>% prune_taxa(taxa_sums(.) > 0, .)
```

```
sub_month6_control <- merged %>% subset_samples(Time_point == "month6" & TB_treatment
=="CONTROL") %>% prune_taxa(taxa_sums(.) > 0, .)
```

```
#####SUBSET BY CONTROL
```

```
subcon_base <- merged %>% subset_samples(Time_point == "contamination CONTROL") %>%
prune_taxa(taxa_sums(.) > 0, .)
```

```
subMock_base <- merged %>% subset_samples(Time_point == "Mock CONTROL2") %>%
prune_taxa(taxa_sums(.) > 0, .)
```

```
subMock1_base <- merged %>% subset_samples(Time_point == "Mock CONTROL1") %>%
prune_taxa(taxa_sums(.) > 0, .)
```

```
sub_type <- merged %>% subset_samples(SAMPLE_type == "NP") %>% prune_taxa(taxa_sums(.)
> 0, .)
```

```
sub_typeIS <- merged %>% subset_samples(SAMPLE_type == "IS") %>% prune_taxa(taxa_sums(.)
> 0, .)
```

```
#####
```

```
#Use the subset data to generate the graphs or the complete data depending on what you are looking at
```

#e.g. for generating a graph at the phylum level for the baseline cases; the sub\_case file generated above was used in the command below and the taxonomic level was selected as phylum (this can be done for every other taxonomic level, just replace phylum with genus, order, class etc..)

# the command included a filter(Abundance > 0.02) which filters out low abundance taxa i.e. everything above 2% was included?

#edited script to family/Genus etc. changed colour palette

#set colours and theme for Phyla so that they will always be the same

```
theme_set(theme_bw())
```

#phyla colours used are dodgerblue3, mediumseagreen, mediumpurple4, mediumturquoise, mediumslateblue, mediumvioletred, mediumspringgreen

```
getPalette = colorRampPalette(c("mediumslateblue", "mediumturquoise", "mediumvioletred",
"dodgerblue3", "mediumseagreen","burlywood", "ivory", "beige", "aquamarine", "aliceblue",
"darkslategray", "darkorange", "firebrick", "deepskyblue", "floralwhite","brown", "green",
"mediumspringgreen","violet","blue","coral2","cornsilk1",black,"forestgreen","darkorchid","hon
eydew","hotpink1","tan1","gray82","blueviolet","magenta","navy","mediumpurple4","pink","golde
nrod","red","seagreen","turquoise4","slateblue","antiquewhite", "cyan","yellow"))
```

```
phylumList = unique(tax_table(merged),"Phylum")
```

```
phylumPalette = getPalette(length(phylumList))
```

```
names(phylumPalette) = phylumList
```

```
all_sample_abundance<-
```

```
merged%>%tax_glom(taxrank="Phylum")%>%transform_sample_counts(function(x) {x/sum(x)})
%>% psmelt() %>% filter(Abundance > 0.02) %>% arrange(Phylum)
```

```
Phy_abundance_bca <-sub_base_case %>% tax_glom(taxrank= "Phylum") %>%
transform_sample_counts(function(x) {x/sum(x)}) %>% psmelt() %>% filter(Abundance > 0.02) %>%
arrange(Phylum)
```

```
Phy_abundance_bco <-sub_base_control %>% tax_glom(taxrank= "Phylum") %>%
transform_sample_counts(function(x) {x/sum(x)}) %>% psmelt() %>% filter(Abundance > 0.02) %>%
arrange(Phylum)
```

```
Phy_abundance_2ca <-sub_month2_case %>% tax_glom(taxrank= "Phylum") %>%
transform_sample_counts(function(x) {x/sum(x)}) %>% psmelt() %>% filter(Abundance > 0.02) %>%
arrange(Phylum)
```

```

Phy_abundance_2co <-sub_month2_control %>% tax_glom(taxrank= "Phylum") %>%
transform_sample_counts(function(x) {x/sum(x)}) %>% psmelt() %>% filter(Abundance > 0.02) %>%
arrange(Phylum)

Phy_abundance_6ca <-sub_month6_case %>% tax_glom(taxrank= "Phylum") %>%
transform_sample_counts(function(x) {x/sum(x)}) %>% psmelt() %>% filter(Abundance > 0.02) %>%
arrange(Phylum)

Phy_abundance_6co <-sub_month6_control %>% tax_glom(taxrank= "Phylum") %>%
transform_sample_counts(function(x) {x/sum(x)}) %>% psmelt() %>% filter(Abundance > 0.02) %>%
arrange(Phylum)

#controls at phylum level

mock_control1 <-subMock1_base %>% tax_glom(taxrank= "Phylum") %>%
transform_sample_counts(function(x) {x/sum(x)}) %>% psmelt() %>% filter(Abundance > 0.02) %>%
arrange(Phylum)

mock_control <-subMock_base %>% tax_glom(taxrank= "Phylum") %>%
transform_sample_counts(function(x) {x/sum(x)}) %>% psmelt() %>% filter(Abundance > 0.02) %>%
arrange(Phylum)

sequencing_controls <-subcon_base %>% tax_glom(taxrank= "Phylum") %>%
transform_sample_counts(function(x) {x/sum(x)}) %>% psmelt() %>% filter(Abundance > 0.02) %>%
arrange(Phylum)

```

#### Plots at phylum level: to generate multiple plots at once

# load the following

```

library(ggpubr)

library(gridExtra)

getPalette=colorRampPalette(c("mediumslateblue","mediumturquoise","mediumvioletred","dodg
erblue3",mediumseagreen","burlywood","ivory","beige","aquamarine",
"aliceblue","darkslategray","darkorange","firebrick","deepskyblue",
"floralwhite","brown","green",
"mediumspringgreen","violet","blue","coral2","cornsilk1","black","forestgreen","darkorchid","ho
neydew","hotpink1","tan1","gray82","blueviolet","magenta","navy","mediumpurple4","pink","gol
denrod","red","seagreen","turquoise4","slateblue","antiquewhite","cyan","yellow"))

phylumList = unique(tax_table(merged)[,"Phylum"])

phylumPalette = getPalette(length(phylumList))

names(phylumPalette) = phylumList

```

#assign a variable to each graph

```

a <-ggplot(Phy_abundance_bca, aes(x = X.SampleID, y = Abundance, fill = Phylum))+
  geom_bar(stat = "identity")+ scale_fill_manual(values = phylumPalette) +
  theme(axis.title.x = element_blank(), axis.text.x = element_text(size = 8, angle = 270)) +
  guides(fill = guide_legend(reverse = TRUE, keywidth = 1, keyheight = 1))+
  ylab("Relative Abundance (Phyla > 2%)" ) + ggtitle("Phylum Composition of baseline cases")
b <-ggplot(Phy_abundance_bco, aes(x = X.SampleID, y = Abundance, fill = Phylum))+
  geom_bar(stat = "identity")+ scale_fill_manual(values = phylumPalette) +
  theme(axis.title.x = element_blank(), axis.text.x = element_text(size = 8, angle = 270)) +
  guides(fill = guide_legend(reverse = TRUE, keywidth = 1, keyheight = 1))+
  ylab("Relative Abundance (Phyla > 2%)" ) + ggtitle("Phylum Composition of baseline controls")
c <-ggplot(Phy_abundance_2ca, aes(x = X.SampleID, y = Abundance, fill = Phylum))+
  geom_bar(stat = "identity")+ scale_fill_manual(values = phylumPalette) +
  theme(axis.title.x = element_blank(), axis.text.x = element_text(size = 8, angle = 270)) +
  guides(fill = guide_legend(reverse = TRUE, keywidth = 1, keyheight = 1))+
  ylab("Relative Abundance (Phyla > 2%)" ) + ggtitle("Phylum Composition of Month 2 cases")
d <-ggplot(Phy_abundance_2co, aes(x = X.SampleID, y = Abundance, fill = Phylum))+
  geom_bar(stat = "identity")+ scale_fill_manual(values = phylumPalette) +
  theme(axis.title.x = element_blank(), axis.text.x = element_text(size = 8, angle = 270)) +
  guides(fill = guide_legend(reverse = TRUE, keywidth = 1, keyheight = 1))+
  ylab("Relative Abundance (Phyla > 2%)" ) + ggtitle("Phylum Composition of Month 2 controls")
e <-ggplot(Phy_abundance_6ca, aes(x = X.SampleID, y = Abundance, fill = Phylum))+
  geom_bar(stat = "identity")+ scale_fill_manual(values = phylumPalette) +
  theme(axis.title.x = element_blank(), axis.text.x = element_text(size = 8, angle = 270)) +
  guides(fill = guide_legend(reverse = TRUE, keywidth = 1, keyheight = 1))+
  ylab("Relative Abundance (Phyla > 2%)" ) + ggtitle("Phylum Composition of Month 6 cases")
f <-ggplot(Phy_abundance_6co, aes(x = X.SampleID, y = Abundance, fill = Phylum))+

```

```
geom_bar(stat = "identity")+ scale_fill_manual(values = phylumPalette) + theme(axis.title.x =
element_blank(), axis.text.x = element_text(size = 8, angle = 270)) + guides(fill =
guide_legend(reverse = TRUE, keywidth = 1, keyheight = 1))+ ylab("Relative Abundance (Phyla >
2%))" + ggtitle("Phylum Composition of Month 6 controls")
```

```
#use the variables as shown below
```

```
#arrange by column and row
```

```
grid.arrange(a,b,c,d,e,f, ncol = 2, nrow = 3)
```

```
#####
```

```
#graphs for controls
```

```
sequencing_controls<-
```

```
subcon_base%>%tax_glom(taxrank="Phylum")%>%transform_sample_counts(function(x)
{x/sum(x)}) %>% psmelt() %>% filter(Abundance > 0.02) %>% arrange(Phylum)
```

```
S <-ggplot(sequencing_controls, aes(x = X.SampleID, y = Abundance, fill = Phylum))+geom_bar(stat
= "identity")+ scale_fill_manual(values = phylumPalette) + theme(axis.title.x = element_blank(),
axis.text.x = element_text(size = 8, angle = 270)) + guides(fill = guide_legend(reverse = TRUE,
keywidth = 1, keyheight = 1))+ ylab("Relative Abundance (Phyla > 2%))" + ggtitle("Phylum
Composition of sequencing controls")
```

```
M <-ggplot(mock_control, aes(x = X.SampleID, y = Abundance, fill = Phylum))+geom_bar(stat =
"identity")+ scale_fill_manual(values = phylumPalette) + theme(axis.title.x = element_blank(),
axis.text.x = element_text(size = 8, angle = 270)) + guides(fill = guide_legend(reverse = TRUE,
keywidth = 1, keyheight = 1))+ylab("Relative Abundance (Phyla > 2%))" + ggtitle("Phylum
Composition of mock control 2")
```

```
m <-ggplot(mock_control1, aes(x = X.SampleID, y = Abundance, fill = Phylum))+geom_bar(stat =
"identity")+ scale_fill_manual(values = phylumPalette) + theme(axis.title.x = element_blank(),
axis.text.x = element_text(size = 8, angle = 270)) + guides(fill = guide_legend(reverse = TRUE,
keywidth = 1, keyheight = 1))+ylab("Relative Abundance (Phyla > 2%))" + ggtitle("Phylum
Composition of mock control 1")
```

```
grid.arrange(S,M,m, ncol = 2, nrow = 3)
```

```
#the tables below can be used for alpha and beta diveristy
```

```
# Absolute abundances
```

```
#additional commands
```

```
phyla.otu.absolute <- abundances(phyla)
```

```

genus.otu.absolute <- abundances(genus)

# Relative abundances

phyla.otu.relative <- abundances(phyla, "compositional")

genus.otu.relative <- abundances(genus, "compositional")

```

```
#Rarefaction (did not use this in the study)
```

```
#rarified <- rarefy_even_depth(merged)
```

```
#####
```

```
#don't run all commands it takes a while to complete, if you only need the genus table only run that line in the script
```

```
#relative abundance at the different taxonomic level
```

```
#DIVERSITY MEASURES
```

```
# add tree file to phyloseq object
```

```
#treefile = "rep_set.tre"
```

```
#tree.obj = import_qiime(treefilename = treefile)
```

```
#otu.table = merge_phyloseq(merged, tree.obj)
```

```
# removing otus that are 0 and transforming to relative abundance
```

```

otu.table = subset_taxa(otu.table, rowSums(otu_table(otu.table)) != 0)

normalizeSample = function(x) {
    x/sum(x)
}

otu.relative.table = transformSampleCounts(merged, normalizeSample)

Phylum.rel.table = tax_glom(otu.relative.table, taxrank = "Phylum")

Class.rel.table = tax_glom(otu.relative.table, taxrank = "Class")

Order.rel.table = tax_glom(otu.relative.table, taxrank = "Order")

Family.rel.table = tax_glom(otu.relative.table, taxrank = "Family")

Genus.rel.table = tax_glom(otu.relative.table, taxrank = "Genus")

OTU.rel.table = tax_glom(otu.relative.table, taxrank = "OTU")

```

```
#Alpha diversity
```

#alpha diversity colour palette

```
library(vegan)

Shannon_diversity = diversity(otu_table(Family.rel.table), index = "shannon", MARGIN = 2, base = exp(1))

write.table(Shannon_diversity, file="shannon.txt", sep="\t")
```

#edited shannon.txt to include Time\_point and TB\_treatment metadata and saved it as shannon\_edit.txt

#removed the controls from the file and only looked at samples this was done for all the tests (Shannon and Simpson)

#used to read saved text file e.g. Shannon file

```
shannon_edit <- read.delim("shannon_edit.txt")
```

#diversity by timepoint (can't remember if ggboxplot shows "mean +- std error" or "median and IQR" (think it's the latter, just check))

```
library(ggpubr)
```

```
library(ggplot2)
```

#subset by time point

```
BLsub <- subset(shannon_edit, Time_point=="Baseline")
```

```
M6sub <- subset(shannon_edit, Time_point=="month6")
```

```
case_sub <- subset(shannon_edit, TB_treatment=="CASE")
```

###exclude data substitue equal sign with "!"

```
BL6sub <- subset(shannon_edit, Time_point!="month2")
```

```
BL6sub <- subset(case_sub, Time_point!="month2")
```

```
M6con <- subset(shannon_edit, Time_point=="month6")
```

#HEALTHY VS CASE SUBSET

```
M6con2 <-subset(M6sub, TB_treatment=="CONTROL")
```

```
BLTB <- subset(BLsub, TB_treatment=="CASE")
```

```
spplot <- ggboxplot(shannon_edit, x = "Time_point", y = "Shannon_Diversity", ylab = "Shannon", xlab = "Time point", title = "Shannon test", add = "jitter", outlier.shape = NA, width = 0.5)
```

```
BLsplot <- ggboxplot(BLsub, x = "TB_treatment", y = "Shannon_Diversity", ylab = "Shannon", xlab
= "TB treatment", title = "Shannon test Baseline", fill= "TB_treatment", palette =c("#00AFBB",
"#E7B800", "#FC4E07"), add = "jitter", outlier.shape = NA, width = 0.5)
```

**BLsplot**

#is there a difference between the cases and control group at baseline

```
BLplot1 <- BLsplot + facet_grid(.~Time_point)
```

#####STATS BETWEEN TWO GROUPS#####

```
BLstatsplot <- BLsplot + stat_compare_means(method= "wilcox.test")
```

**BLstatsplot**

```
splot <- ggboxplot(shannon_edit, x = "Time_point", y = "Shannon_Diversity", ylab = "Shannon", xlab
= "Time point", title = "Shannon test", color = "Time_point", palette =c("#00AFBB", "#E7B800",
"#FC4E07"), add = "jitter", outlier.shape = NA, width = 0.5)
```

```
M6splot <- ggboxplot(M6sub, x = "TB_treatment", y = "Shannon_Diversity", ylab = "Shannon", xlab
= "TB treatment", title = "Shannon test Month6", palette =c("#00AFBB", "#E7B800", "#FC4E07"), add
= "jitter", outlier.shape = NA, width = 0.5)
```

```
BL6splot <- ggboxplot(BL6sub, x = "Time_point", y = "Shannon_Diversity", ylab = "Shannon", xlab
= "Time point", title = "Shannon test BL vs M6", fill= "TB_treatment", palette =c("#00AFBB",
"#E7B800", "#FC4E07"), add = "jitter", outlier.shape = NA, width = 0.5)
```

```
BL6statsplot <- BL6splot + stat_compare_means(method= "wilcox.test")
```

# outlier.shape = NA makes outliers invisible (boxplots show outliers as points, so adding jitter points may make some points plot twice)

# jitter makes points not clump on each other, but spread out

#diversity by timepoint, Split view into case vs control:

```
splot2 <- splot + facet_grid(.~TB_treatment)
```

**splot2**

#is there a difference between time points in the cases and control groups?

```
statsplot2 <- splot2 + stat_compare_means(method="kruskal.test")
```

```
casplot <- ggboxplot(case_sub, x = "Time_point", y = "Shannon_Diversity",
```

```
      ylab = "Shannon", xlab = "Time point", title = "Shannon test (Case group)", color =
"Time_point", palette =c("#00AFBB", "#E7B800", "#FC4E07"), add = "jitter", outlier.shape = NA,
width = 0.5)
```

```
#####stats measure
```

```
#stat compare inputs the p value on the graph
```

```
#the data not normally distributed used kruskal wallis test (non-parametric test)
```

```
statsplot1 <- splot + stat_compare_means(method="kruskal.test")
statsplot2 <- splot2 + stat_compare_means(method="kruskal.test")
#diversity between cases and controls
splot3 <- ggboxplot(shannon_edit, x = "TB_treatment", y = "Shannon_Diversity", ylab = "Shannon",
xlab = "Treatment arm", title = "Shannon test", color = "TB_treatment", palette =c("#00AFBB",
"#E7B800", "#FC4E07"),add = "jitter", outlier.shape = NA, width = 0.5)
splot3
#diversity between case vs control, Split view into timepoints:
splot4 <- splot3 + facet_grid(~Time_point)
splot4
```

```
#####Simpsons#####
```

```
#USED THE FAMILY RELATIVE ABUNDANCE TO GENERATE PLOTS
```

```
library(vegan)
Simpson_diversity = diversity(otu_table(Family.rel.table), index = "simpson", MARGIN = 2, base =
exp(1))
write.table(Simpson_diversity, file="simpsonF1.txt", sep="\t")
#edited shannon.txt to include Time_point and TB_treatment metadata and saved it as
shannon_edit.txt
simpson_edit <- read.delim("simpson_edit.txt")
splot <- ggboxplot(simpson_edit, x = "Time_point", y = "Simpson_index",ylab = "Simpson", xlab =
"Time point", title = "Simpson test", color = "Time_point", palette =c("#00AFBB", "#E7B800",
"#FC4E07"),add = "jitter", outlier.shape = NA, width = 0.5)
splot <- ggboxplot(simpsonF_edit, x = "Time_point", y = "Simpson_index", ylab = "Simpson", xlab
= "Time point", title = "Simpson test", color = "Time_point", palette =c("#00AFBB", "#E7B800",
"#FC4E07"),add = "jitter", outlier.shape = NA, width = 0.5)
```

```
splot <- ggboxplot(simp_ISM1, x = "Time_point", y = "Simpson_index", ylab = "Simpson", xlab =
"Time point", title = "Simpson test", color = "Time_point", palette =c("#00AFBB", "#E7B800",
"#FC4E07"),add = "jitter", outlier.shape = NA, width = 0.5)
```

```
splot
```

#diversity by timepoint, Split view into case vs control:

```
splot2 <- splot + facet_grid(~TB_treatment)
```

```
splot2
```

#is there a difference between time points in the cases and control groups?

```
statsplot2 <- splot2 + stat_compare_means(method="kruskal.test")
```

```
statsplot2
```

#diversity between cases and controls

```
splot3 <- ggboxplot(simpson_edit, x = "TB_treatment", y = "Simpson_index", ylab = "Simpson", xlab
= "Treatment arm", title = "Simpson test", color = "TB_treatment", palette =c("#00AFBB",
"#E7B800", "#FC4E07"),add = "jitter", outlier.shape = NA, width = 0.5)
```

```
splot3 <- ggboxplot(simp_ISM1, x = "TB_treatment", y = "Simpson_index", ylab = "Simpson", xlab
= "Treatment arm", title = "Simpson test", color = "TB_treatment", palette =c("#00AFBB",
"#E7B800", "#FC4E07"), add = "jitter", outlier.shape = NA, width = 0.5)
```

```
splot3
```

#diversity between case vs control, Split view into timepoints:

```
splot4 <- splot3 + facet_grid(~Time_point)
```

```
splot4
```

```
statsplot4 <- splot4 + stat_compare_means(method="kruskal.test")
```

```
statsplot4
```

#is there a difference between the cases and control group at baseline

```
BLsplot <- ggboxplot(BLsub, x = "TB_treatment", y = "Simpson_index", ylab = "Simpson", xlab =
"TB treatment", title = "Simpson test Baseline", fill= "TB_treatment",palette =c("#00AFBB",
"#E7B800", "#FC4E07"), add = "jitter", outlier.shape = NA, width = 0.5)
```

```
BLsplot
```

```
BLplot1 <- BLsplot + facet_grid(~Time_point)
```

```
BLstatsplot <- BLsplot + stat_compare_means(method= "wilcox.test")
```

**BLstatsplot**

```
###Baseline cases vs m6 cases
```

```
BL6splot <- ggboxplot(BL6sub, x = "Time_point", y = "Simpson_index", ylab = "Simpson", xlab =
"Time point", title = "Simpson test BL vs M6", fill= "TB_treatment", palette =c("#00AFBB",
"#E7B800", "#FC4E07"), add = "jitter", outlier.shape = NA, width = 0.5)
```

```
BL6splot <- ggboxplot(simpblTBcon6, x = "Time_point", y = "Simpson_index", ylab = "Simpson",
xlab = "Time point", title = "Simpson test BL vs M6", fill= "TB_treatment", palette =c("#00AFBB",
"#E7B800", "#FC4E07"), add = "jitter", outlier.shape = NA, width = 0.5)
```

```
BL6statsplot <- BL6splot + stat_compare_means(method= "wilcox.test")
```

```
#####Beta diversity
```

```
subsample <- merged %>% subset_samples(Sample_seqcon == "sample") %>%
prune_taxa(taxa_sums(.) > 0, .)
```

```
sub_base <- merged %>% subset_samples(Time_point == "Baseline") %>%
prune_taxa(taxa_sums(.) > 0, .)
```

```
cas_conTB <- subsample %>% subset_samples(TB_diagnosis != "control" & TB_treatment ==
"CASE") %>% prune_taxa(taxa_sums(.) > 0, .)
```

```
any(taxa_sums(subsample) ==0)
```

```
merged1 <- prune_taxa(taxa_sums(sub_base)>1, sub_base)
```

```
merged2 <- subset_taxa(merged1, Phylum != "p_")
```

```
merged2
```

```
merged2a <- subset_taxa(merged2, Class!="Chloroplast")
```

```
merged3 <- subset_taxa(merged2a, Order!="Mitochondria")
```

```
merged3
```

```
otu.table = merge_phyloseq(merged3, tree.obj)
```

```
otu.table = subset_taxa(otu.table, rowSums(otu_table(otu.table)) != 0)
```

```
normalizeSample = function(x) {
```

```
x/sum(x)
```

```
}
```

```
otu.relative.table = transformSampleCounts(otu.table, normalizeSample)
```

```
Family.rel.table = tax_glom(otu.relative.table, taxrank = "Family")
```

```
#####Graphs
```

```
beta.ps1 <- plot_ordination(Family.rel.table, bx.ord_pcoa_bray, color="TB_treatment", label =
"TB_treatment") + geom_point(aes(shape = Time_point), size= 4) +
  theme(plot.title = element_text(hjust = 0, size = 12))
beta.ps1
beta.ps1 <- beta.ps1 + theme_bw(base_size = 14) + theme(panel.grid.major = element_blank(),
panel.grid.minor = element_blank())
beta.ps2 <- beta.ps1 + geom_line() + scale_color_brewer(palette = "Dark2")
beta.ps2
beta.ps3 <- plot_ordination(Family.rel.table, bx.ord_pcoa_bray, color="TB_treatment", label =
"TB_treatment") + geom_point(size= 4) +
  theme(plot.title = element_text(hjust = 0, size = 12))
beta.ps3 <- beta.ps3 + theme_bw(base_size = 14) + theme(panel.grid.major = element_blank(),
panel.grid.minor = element_blank())
beta.ps3 + scale_color_brewer(palette = "Dark2") + stat_ellipse()
metadf.bx <- data.frame(sample_data(ps4fam.rel))
bray_ps.bxn <- phyloseq::distance(physeq = ps4fam.rel, method = "bray")
##Stats test
adonis.test <- adonis(bray_ps.bxn ~ TB_treatment, data = metadf.bx)
adonis.test
adonis.test2 <- adonis(bray_ps.bxn ~ Time_point, data = metadf.bx)
adonis.test2
```

```
#####
```

```
#Heatmap
```

```
##load the following libraries
```

```
library("ggplot2")
theme_set(theme_bw())
library("RColorBrewer")
```

```

library("gplots")

library('ape')

library("plyr")

library("d3heatmap")

library("vegan")

library("Heatplus")

library("igraph")

library('Hmisc')

library("reshape2")

theme_set(theme_bw())

sample_data(Fam.Rel.table1B)$TB_treatment

SampleVector = sample_data(Fam.Rel.table1B)$TB_treatment

sample_data(Fam.Rel.table1B)$TB_treatment

```

#duplicate to create a color vector and replace value w/ color

#Colorvector can only replace numbers!

```

Colorvector <- SampleVector

Colorvector <- replace(Colorvector, which (Colorvector == "3"), "chocolate")

Colorvector <- replace(Colorvector, which (Colorvector == "4"), "chartreuse")

Colorvector <- replace(Colorvector, which (Colorvector == "5"), "cadetblue")

FamilyData <- otu_table(Fam.Rel.table1B)

FamilyData.Bray.dist <- vegdist(FamilyData, method = "bray")

Family.Bray.clus <- hclust(FamilyData.Bray.dist, "aver")

Bray.dist = distance(FamilyData, method="bray")

cluster.Bray = hclust(Bray.dist, "aver")

```

#Here we are able to change the names for genus or family that are labelled as "g\_\_/" f\_\_"--> Come back to this

```

tax_table(Fam.Rel.table1B)

```

```
Fam.Rel.table1B.New.Names=prune_taxa(tail(names(sort(taxa_sums(Fam.Rel.table1B))),
ntaxa(Fam.Rel.table1B)), Fam.Rel.table1B)
```

```
tax_table(Fam.Rel.table1B.New.Names)<-cbind(tax_table(Fam.Rel.table1B.New.Names),
Strain=taxa_names(Fam.Rel.table1B.New.Names))
```

```
# Define the ranks you want to include
```

```
myranks = c("Phylum", "Family")
```

```
mylabels=apply(tax_table(Fam.Rel.table1B.New.Names)[,myranks],1,paste,sep="", collapse="_")
```

```
tax_table(Fam.Rel.table1B.New.Names)<-
```

```
cbind(tax_table(Fam.Rel.table1B.New.Names),catglab=mylabels)
```

```
tax_table(Fam.Rel.table1B.New.Names)
```

```
#Now Plot Heat map with dendograms
```

```
mypalette <- colorRampPalette(c('#ffffff', '#4169E1', '#0000CD'))
```

```
pdf("Bray Curtis heatmap Family.pdf", height = 10, width = 15)
```

```
heatmap.2(FamilyData, margins = c(10,20),
```

```
  density.info = "none",
```

```
  trace = "none",
```

```
  keysize = 0.75,
```

```
  key.title = "Relative abundance",
```

```
  offsetRow = 1, offsetCol = 1,
```

```
  dendrogram = "both",
```

```
  Rowv = as.dendrogram(Family.Bray.clus),
```

```
  Colv = as.dendrogram(cluster.Bray),
```

```
  labRow=tax_table(Fam.Rel.table1B.New.Names)[,"catglab"],
```

```
  cexRow = .9,
```

```
  labCol = sample_data(Fam.Rel.table1B)$heatmap,
```

```
  cexCol = .9,
```

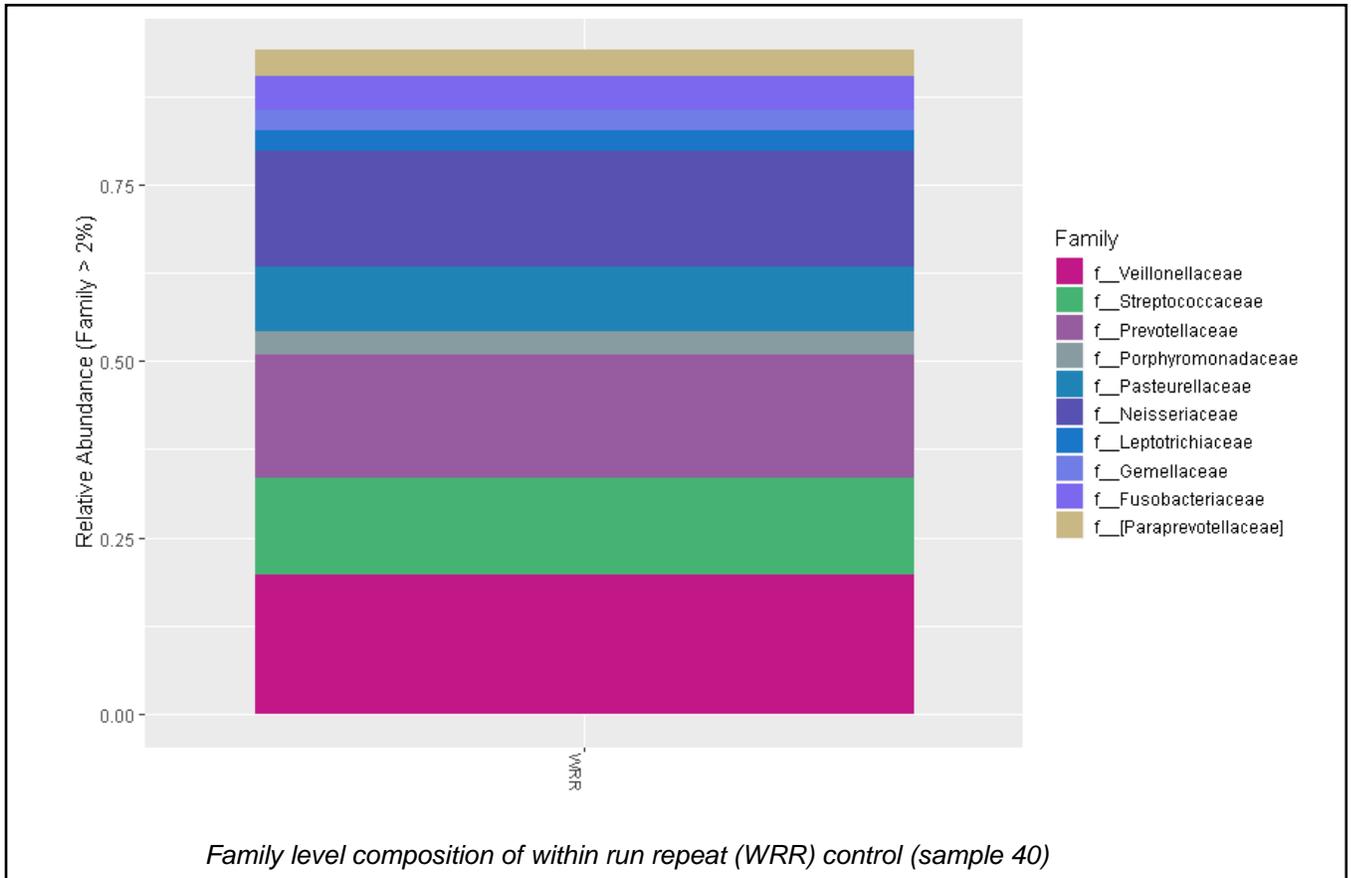
```
  col = mypalette(17),
```

```
  symm=F,symkey=F,symbreaks=T, scale="none",
```

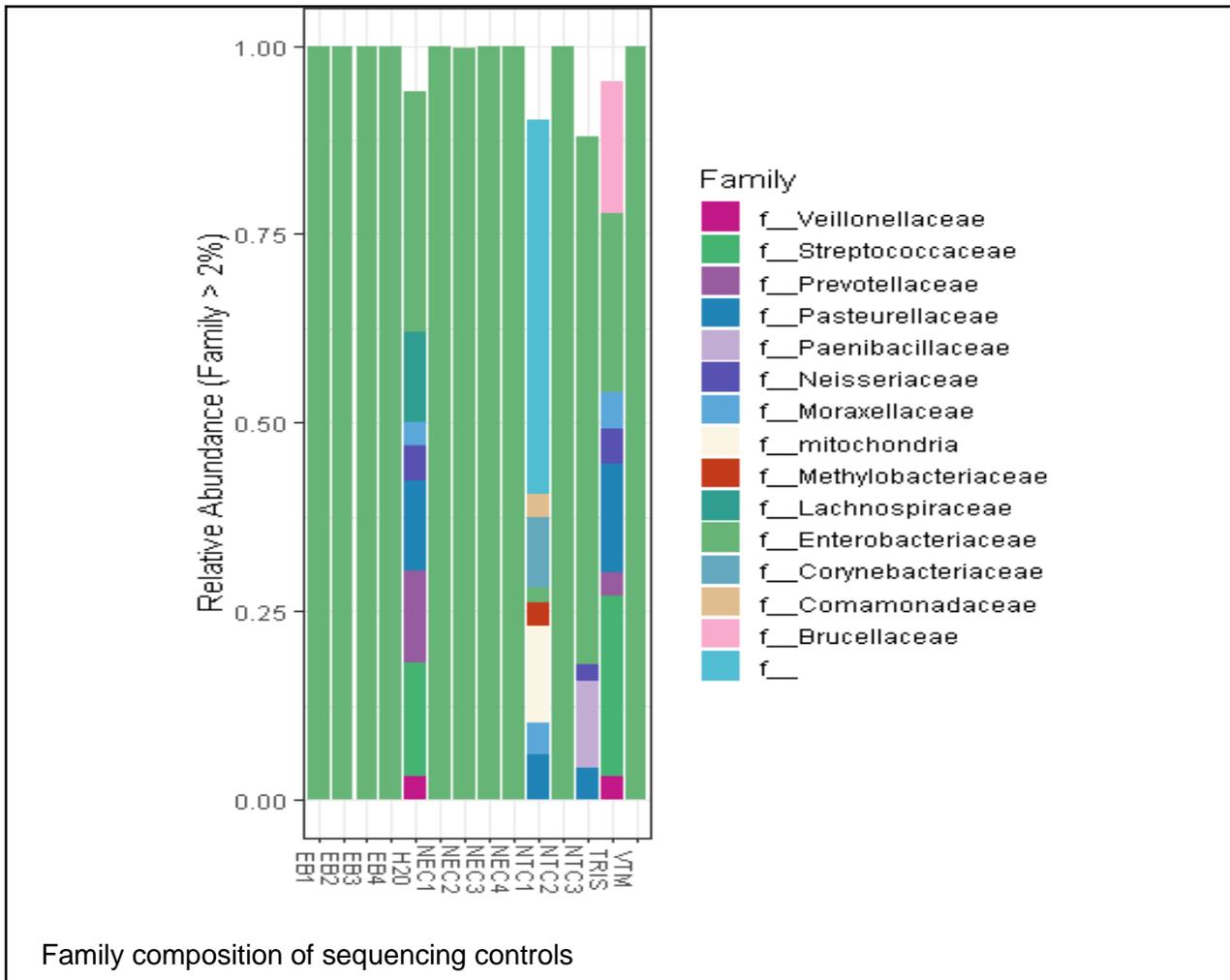
```
  breaks =c(seq(0,.1,length=10),seq(.11,0.3,length=4),seq(0.31,.7,length=4)),
```

```
main = "Partipants")  
dev.off()
```

### Addendum 8



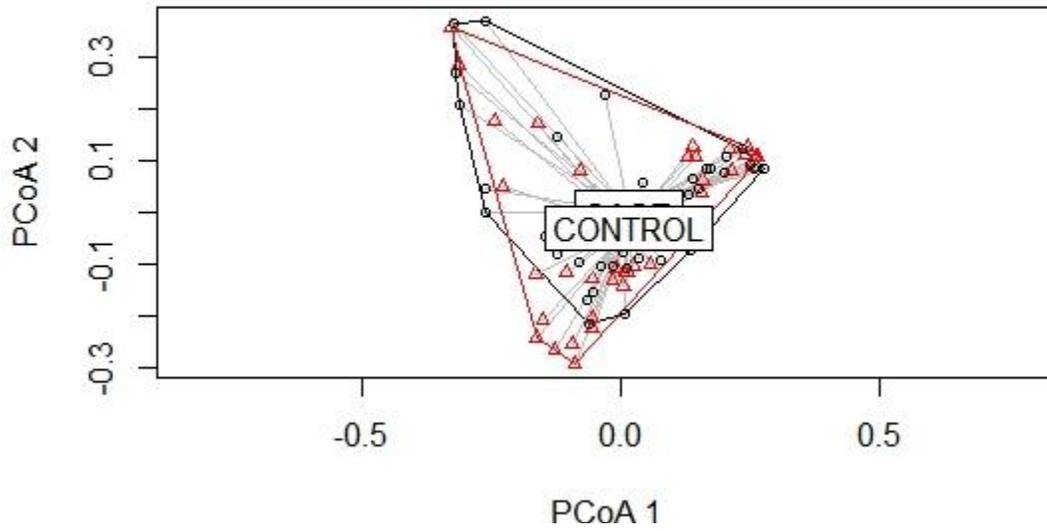
### Addendum 9



## Addendum 10

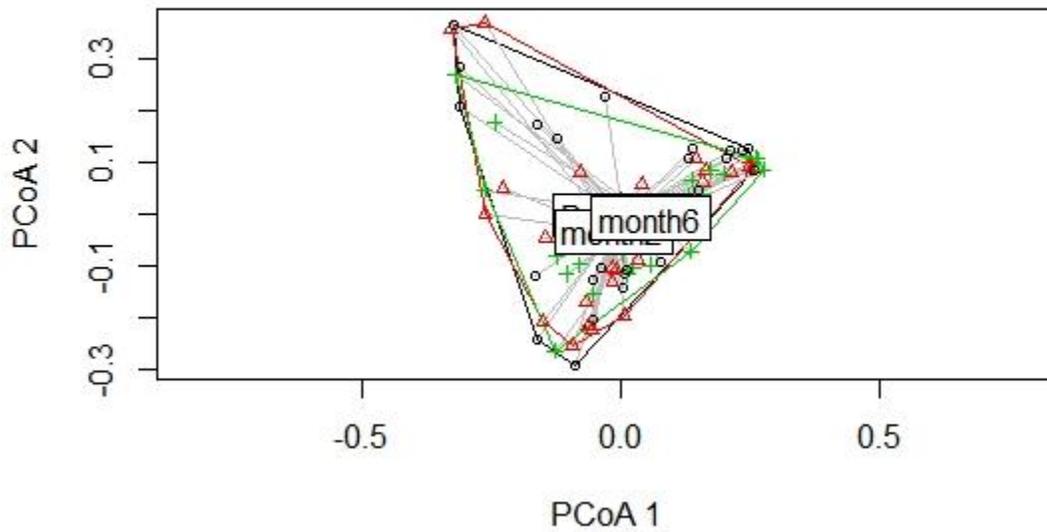
Betadisper function for weighted Unifrac distances

TB case and ill control comparison



Method= Weighted UniFrac

Time point comparison between all samples



Method= Weighted UniFrac

Call:

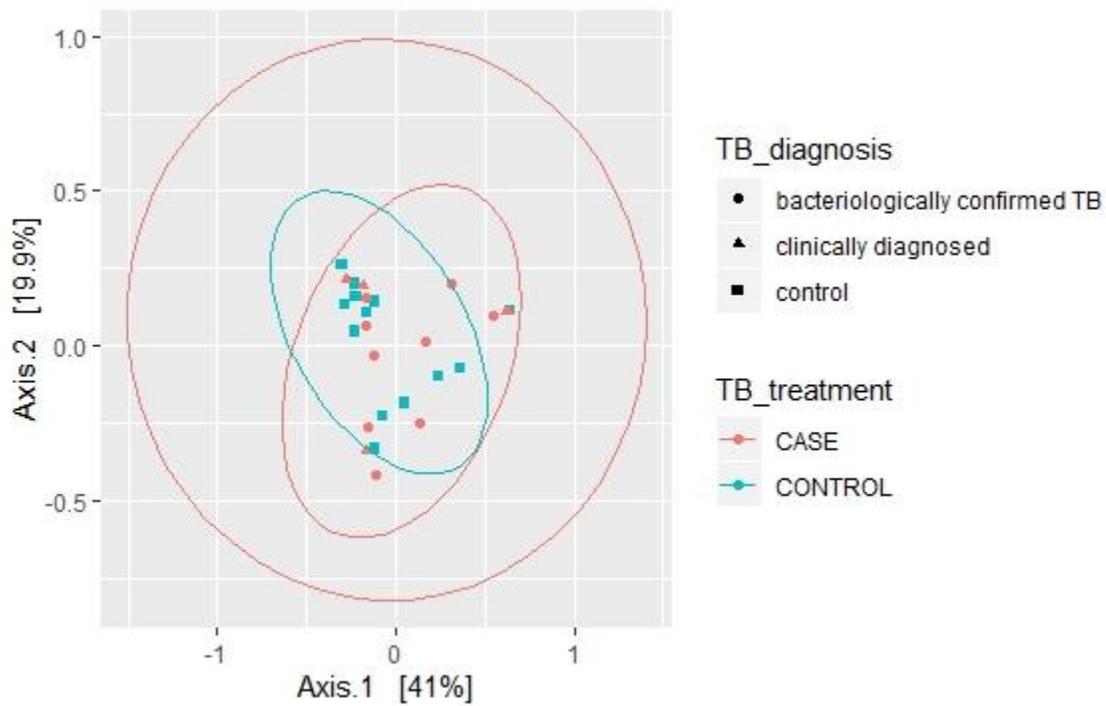
```
adonis(formula = wUniF.dist ~ sample_data(Family.rel.table)$TB_treatment)
```

Permutation: free

Number of permutations: 999

Terms added sequentially (first to last)

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
sample_data(Family.rel.table)\$TB_treatment	1	0.1161	0.11612	0.56406	<b>0.02296</b>	<b>0.72</b>
Residuals	24	4.9409	0.20587	0.97704		
Total	25	5.0570		1.00000		



*Weighted unfrac: Baseline comparison between TB cases and ill controls*

Call:

```
adonis(formula = wUniF.dist ~ sample_data(Family.rel.table)$TB_treatment)
```

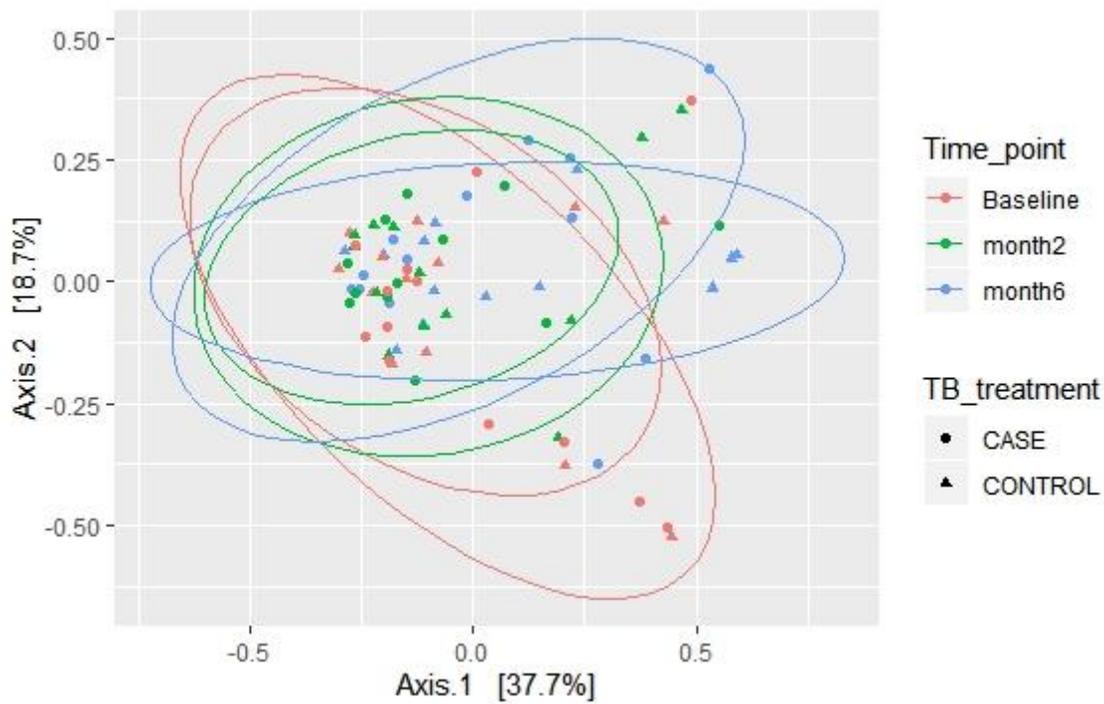
Permutation: free

Number of permutations: 999

Terms added sequentially (first to last)

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
sample_data(Family.rel.table)\$TB_treatment	1	0.1288	0.12885	0.68547	<b>0.00906</b>	<b>0.625</b>
Residuals	75	14.0979	0.18797	0.99094		
Total	76	14.2267		1.00000		

TB case and Ill control at all time points



Call:

```
adonis(formula = wUniF.dist ~ sample_data(Family.rel.table)$Time_point)
```

Permutation: free

Number of permutations: 999

Terms added sequentially (first to last)

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
sample_data(Family.rel.table)\$Time_point	2	0.5644	0.28222	1.5286	<b>0.03967</b>	<b>0.121</b>
Residuals	74	13.6623	0.18463	0.96033		
Total	76	14.2267		1.00000		

TB case and ill-control group over time

