Profiling of Bacterial Communities in Chronic Obstructive Pulmonary Disease

Thesis submitted for the degree of
Doctor of Philosophy
at the University of Leicester

by
Kairabi Sadhu Haldar
Department of Infection, Immunity and Inflammation
University of Leicester

October 2014
I would like to dedicate this work to my late Dadai (my grandfather) who had a dream that his granddaughter may one day study for and complete a PhD. Sadly I will not have the opportunity to share this with him in person.
Abstract

Profiling of bacterial communities in Chronic Obstructive Pulmonary disease

Kairabi Haldar

Introduction: It is hypothesized that bacteria are important in the pathogenesis of COPD and exacerbations. Most bacteriological research in COPD has utilised culture based methods. Novel molecular approaches enable detailed evaluation of the airway microbiome that may better inform the role of bacteria in COPD. This project aimed to characterise the microbial community in COPD at stable state and during acute exacerbations through assessment of serial sputa at stable (S), exacerbation (E), follow up (F) and recovery (R) visits.

Methods: Sputum from 145 clinical trial COPD patients was collected at multiple stable visits and at each exacerbation (E, F and R visit) over 12 months. Real-time quantitative PCR (qPCR) was performed on sputum DNA using universal 16S gene primers and specific gene targets to quantify total bacterial load and the specific pathogens *H. influenzae*, *S. pneumoniae*, *M. catarrhalis* and *S. aureus*. In a subgroup of 30 exacerbating patients, 454 high-throughput pyrosequencing of 16S rDNA amplicons was performed at each of the 4 visits.

Results: There was no significant difference in total bacterial load or any specific pathogen between longitudinal stable and exacerbation samples. 454 pyrosequencing identified Proteobacteria and Firmicutes to be the dominant groups contributing >80% of the sequence reads at phylum level. *Haemophilus, Moraxella* and *Streptococcus* were the dominant groups at genus level. No significant within-subject change in the microbial community was observed across visits. Cluster analysis, based on the ratio of Proteobacteria to Firmicutes (P:F) characterised three subgroups. The high P:F subgroup was characterised by a significant increase in P:F from S to E visits, associated with raised blood CRP and sputum IL-1β levels, suggesting a role for bacteria in exacerbation pathogenesis for this subgroup.

Conclusions: Molecular profiling identifies heterogeneity in the airway microbiome of COPD patients, with a role for bacteria suggested in a subgroup.
Acknowledgement

I would like to start by expressing my gratitude and enormous thanks to my supervisor Professor Mike Barer who gave me the opportunity to work and undertake my post graduate studies in his lab at a time when I was new to this country. His constant guidance and support has allowed me to get to where I am today.

I would like to thank Professor Chris Brightling for providing the opportunity to be a part of his clinical studies and Dr Mona Bafadhel for helping and providing the clinical data and samples which formed my primary study. I would also like to thank Hemu Patel for coordinating clinical sample storage and collection.

Special thanks to Kelvin (Dr Kelvin Lau) who introduced me to the world of scripting and bioinformatics. Much of the analytical work I have been able to do has only been possible with his help. Thanks also to my husband, Pranab, who chipped in with a few statistical pearls of wisdom at different times.

I would also like to thank Gurdeep, Hediah and Amiee for their technical support processing the study samples and Adam for helping with building the sequence library.

I would like to acknowledge "The Gang" in the lab who kept me sane and entertained between PCR reactions. At different times, the gang has included Eddy, Farah, Gosia, Irina, Andrew and Kelvin.

I am forever thankful for the love, support and blessings of my parents and the support and patience of my husband.

And finally, love, hugs and lots of kisses to my daughter Ayanna. I hope I may now be able to answer her question 'Have you finished your PhD?'
Declaration of Joint Effort

All sputum samples were collected at Glenfield hospital as part of the clinical trial study. An aliquot of the homogenised sputum was provided for the bacteriological study from Leicester Royal Infirmary (LRI) hospital after the standard culturing assays. Standard operating procedures (SOP) were set in place by me for sputum collection and storage. I was helped with performing the optimised qPCR assays for the clinical samples based on the SOPs set by me. Clinical data was collected by Dr Mona Bafadhel and was kindly provided for analysis with microbial work. I was assisted for selecting 454 primers and building 16S amplicon library by Dr Adam Berg. 454 sequencing were performed at the Liverpool Genomic center from the normalised 16S amplicon library provided by me. Dr Kelvin Lau provided the preliminary scripts for work with 'R programming language' some of which were later improvised by me for my work. All statistical and biostatistical analysis for qPCR and 454 sequencing work was performed by me.
Peer-reviewed Publications and Presentation from this study


## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>16S primers</td>
<td>16S rRNA gene targeting primers</td>
</tr>
<tr>
<td>1-D</td>
<td>Simpson diversity index</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of variance</td>
</tr>
<tr>
<td>ATP</td>
<td>Adenosine triphosphate</td>
</tr>
<tr>
<td>BAL</td>
<td>Bronchial alveolar lavage</td>
</tr>
<tr>
<td>BHI</td>
<td>Brain heart infusion</td>
</tr>
<tr>
<td>bp</td>
<td>Base pair</td>
</tr>
<tr>
<td>C.F.U</td>
<td>Colony forming unit</td>
</tr>
<tr>
<td>CI</td>
<td>Confidence interval</td>
</tr>
<tr>
<td>COPD</td>
<td>Chronic obstructive pulmonary disease</td>
</tr>
<tr>
<td>CRP</td>
<td>C reactive protein</td>
</tr>
<tr>
<td>Ct</td>
<td>Cycle threshold</td>
</tr>
<tr>
<td>DGGE</td>
<td>Denaturing gel gradient electrophoresis</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>dNTP</td>
<td>Deoxyribonucleotide triphosphate</td>
</tr>
<tr>
<td>dsDNA</td>
<td>Double stranded DNA</td>
</tr>
<tr>
<td>DTT</td>
<td>Dithiothreitol</td>
</tr>
<tr>
<td>E</td>
<td>Exacerbation COPD visit</td>
</tr>
<tr>
<td>EDTA</td>
<td>Ethylenediamine-tetra-acetic acid</td>
</tr>
<tr>
<td>E_H</td>
<td>Shannon equitability index</td>
</tr>
<tr>
<td>em PCR</td>
<td>Emulsion PCR</td>
</tr>
<tr>
<td>F</td>
<td>Follow-up COPD visit</td>
</tr>
<tr>
<td>FEV1</td>
<td>Forced expiratory volume in one second</td>
</tr>
<tr>
<td>FVC</td>
<td>Forced vital capacity</td>
</tr>
<tr>
<td>GIT</td>
<td>Gastrointestinal tract</td>
</tr>
<tr>
<td>GOLD</td>
<td>Global initiative for Obstructive Lung Disease</td>
</tr>
<tr>
<td>H</td>
<td>Shannon-Weiner diversity index</td>
</tr>
<tr>
<td>HF</td>
<td>High Firmicutes cluster</td>
</tr>
<tr>
<td>HP</td>
<td>High Proteobacteria cluster</td>
</tr>
<tr>
<td>HPA</td>
<td>Health Protection Agency</td>
</tr>
<tr>
<td>LB</td>
<td>Luria Bertani</td>
</tr>
<tr>
<td>LPS</td>
<td>Lipopolysaccharide</td>
</tr>
<tr>
<td>LRI</td>
<td>Leicester Royal infirmary</td>
</tr>
<tr>
<td>LRT</td>
<td>Lower respiratory tract</td>
</tr>
<tr>
<td>LSU</td>
<td>Large subunit</td>
</tr>
<tr>
<td>Mtb</td>
<td>Mycobacterium tuberculosis</td>
</tr>
<tr>
<td>NAD</td>
<td>Nicotinamide-adenine-dinucleotide</td>
</tr>
<tr>
<td>NCBI</td>
<td>National center for Biotechnology Information</td>
</tr>
<tr>
<td>NES</td>
<td>Non exacerbated individuals stable samples for 454 study</td>
</tr>
<tr>
<td>NGS</td>
<td>Next generation sequencing</td>
</tr>
<tr>
<td>NICE</td>
<td>National Institute for Health and Clinical Excellence</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>NTHI</td>
<td>Non typeable <em>Haemophilus influenzae</em></td>
</tr>
<tr>
<td>NYC</td>
<td>Not yet been cultured</td>
</tr>
<tr>
<td>OD</td>
<td>Optical density</td>
</tr>
<tr>
<td>OTU</td>
<td>Operational taxonomic unit</td>
</tr>
<tr>
<td>P:F</td>
<td>Proteobacteria to Firmicutes ratio</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component analysis</td>
</tr>
<tr>
<td>PCoA</td>
<td>Principal Coordinate analysis</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PF</td>
<td>Proteobacteria and Firmicutes cluster</td>
</tr>
<tr>
<td>PPB</td>
<td>Potential pathogenic bacteria</td>
</tr>
<tr>
<td>PSB</td>
<td>Protected brush specimen</td>
</tr>
<tr>
<td>QIIME</td>
<td>Quantitative insights into microbial ecology</td>
</tr>
<tr>
<td>qPCR</td>
<td>Quantitative PCR</td>
</tr>
<tr>
<td>R</td>
<td>Recovery COPD visit</td>
</tr>
<tr>
<td>RA</td>
<td>Relative abundance</td>
</tr>
<tr>
<td>rDNA</td>
<td>Ribosomal DNA</td>
</tr>
<tr>
<td>RM-ANOVA</td>
<td>Repeated measures Anova</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>rRNA</td>
<td>Ribosomal RNA</td>
</tr>
<tr>
<td>RT</td>
<td>Respiratory tract</td>
</tr>
<tr>
<td>S</td>
<td>Stable COPD visit</td>
</tr>
<tr>
<td>SBL</td>
<td>Specific bacterial load</td>
</tr>
<tr>
<td>SD</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>SEFR</td>
<td>Stable, exacerbation, follow-up and recovery visits of 454 study</td>
</tr>
<tr>
<td>SEM</td>
<td>Standard error mean</td>
</tr>
<tr>
<td>S_est</td>
<td>Number of species estimated</td>
</tr>
<tr>
<td>S_obs</td>
<td>Number of species observed</td>
</tr>
<tr>
<td>SSE</td>
<td>Sequential stable samples in exacerbated individuals of 454 study</td>
</tr>
<tr>
<td>SSU</td>
<td>Small subunit</td>
</tr>
<tr>
<td>T-RFLP</td>
<td>Terminal restriction fragment length polymorphism</td>
</tr>
<tr>
<td>TAE</td>
<td>Tris-acetate-EDTA</td>
</tr>
<tr>
<td>TBL</td>
<td>Total bacterial load</td>
</tr>
<tr>
<td>URT</td>
<td>Upper respiratory tract</td>
</tr>
<tr>
<td>VGS</td>
<td>Viridans group streptococci</td>
</tr>
<tr>
<td>WGS</td>
<td>Whole genome sequencing</td>
</tr>
<tr>
<td>WHO</td>
<td>World Health Organization</td>
</tr>
<tr>
<td>γproteo</td>
<td>Gammaproteobacteria</td>
</tr>
</tbody>
</table>
1 INTRODUCTION ........................................................................................................... 23

1.1 General introduction .............................................................................................. 23

1.2 Chronic Obstructive Pulmonary Disease (COPD) .............................................. 25
   1.2.1 What is COPD? .................................................................................................. 25
   1.2.2 COPD definition and stages .......................................................................... 25
   1.2.3 COPD symptoms ............................................................................................. 26
   1.2.4 COPD burden ................................................................................................ 26
   1.2.5 COPD cause and risk factors ......................................................................... 27
   1.2.6 COPD exacerbation ....................................................................................... 28
   1.2.7 COPD treatment ............................................................................................. 31

1.3 COPD and Bacteriology ....................................................................................... 31
   1.3.1 The respiratory microbiome in healthy individuals ....................................... 31
   1.3.2 Bacterial role in COPD .................................................................................. 32
   1.3.3 Controversy on the use of antibiotics in the treatment of COPD exacerbation .................................................................................................................. 35
   1.3.4 General problems with bacterial studies in COPD ....................................... 36
1.4 16S rDNA as marker gene for enumeration and characterization of bacterial community

1.4.1 Molecular biology and classification
1.4.2 Ribosomes
1.4.3 16S rRNA gene: pros and cons
1.4.4 Other molecular markers
1.4.5 16S gene based databases

1.5 Enumeration of bacterial community

1.5.1 Difficulties of Quantitative Analysis in Normal PCR
1.5.2 Real-Time PCR: chemistries, application and other methods
1.5.3 Mathematics of absolute quantification
1.5.4 Real-Time PCR fluorescence chemistry
1.5.5 Real-Time PCR applications

1.6 Bacterial community characterization

1.6.1 Microbial complexity of normal human microbiome
1.6.1.1 Host-Microbe interactions
1.6.1.2 Interactions between bacteria (microbe and microbe)
1.6.2 Molecular approaches for 16S rDNA bacterial profiling
1.6.3 Next generation sequencing (NGS) technologies
1.6.3.1 454 sequencing

1.7 Aims and Objectives

1.7.1 Specific objectives using qPCR
1.7.2 454 Sequencing

2 MATERIALS AND METHODS

2.1 Bacterial Strains, Media and Buffers

2.1.1 Bacterial Strains
2.1.2 Media
2.1.3 Reagents and Buffers

2.2 Preparation, Cultivation and Storage of Bacteria

2.2.1 Glycerol Stocks of Bacteria
2.2.2 Enumeration of Colony Forming Units (C.F.U)
2.2.3 Optical Density (OD) Measurements........................................................................... 69
2.2.4 Cultivation of E. coli, P. aeruginosa ........................................................................... 70
2.2.5 Cultivation of S. pneumoniae ..................................................................................... 70
2.2.6 Cultivation of M. catarrhalis .................................................................................... 70
2.2.7 Cultivation of H. influenzae ...................................................................................... 70
2.2.8 Cultivation of B. fragilis .......................................................................................... 70

2.3 Patient Samples ........................................................................................................... 71
2.3.1 COPD subjects and sample collection ..................................................................... 71
2.3.2 Sample preparation, Bacteriological plating and storage .......................................... 72

2.4 Molecular Methods ..................................................................................................... 73
2.4.1 Real time PCR and primers .................................................................................... 73
2.4.2 High-throughput parallel sequencing ...................................................................... 73
2.4.3 Gel analysis of PCR amplicons .............................................................................. 74
2.4.4 Cloning and sequence analysis of Streptococcus species ........................................ 74

3 CHANGES IN THE TOTAL AND SPECIFIC BACTERIAL LOAD AT STABLE AND EXACERBATION STATE OF COPD ................. 76

3.1 Introduction: .................................................................................................................. 76

3.2 Materials and Methods: .............................................................................................. 77
3.2.1 Bacterial strains and Study sample ........................................................................ 77
3.2.2 DNA Extraction....................................................................................................... 78
  3.2.2.1 DNA extraction from sputum samples ............................................................... 78
  3.2.2.2 DNA extraction from Gram positive cultures .................................................. 78
  3.2.2.3 DNA extraction from Gram Negative cultures ............................................... 78
  3.2.2.4 Other DNA release and purification methods used ........................................... 79
3.2.3 DNA Concentration measurement .......................................................................... 80
3.2.4 Primers and oligonucleotide sequence used for the real time assays .................... 80
3.2.5 Real Time PCR ....................................................................................................... 82
  3.2.5.1 DNA standards for quantification assays ......................................................... 83
  3.2.5.2 Quantification Assays ...................................................................................... 84
  3.2.5.3 qPCR data analysis ......................................................................................... 86
  3.2.5.4 Calculations of genome/ml from real time assay ............................................. 86
3.2.6 Statistical analysis .................................................................................................... 87
3.3 Results ............................................................................................................................................ 87

3.3.1 Developing and optimizing qPCR assays for quantifying TBL and for SBL in COPD sputum ................................................................................................................................. 87

3.3.1.1 DNA Extraction Methods ..................................................................................................... 87

3.3.1.2 Development and Optimisation of TBL and SBL qPCR .................................................. 90

3.3.1.2.1 Optimisation of 16S rDNA based TBL qPCR ................................................................. 90

3.3.1.2.2 Correction of 16S gene copy variation and the no template control (NTC) reading for enumerating TBL .......................................................................................................... 92

3.3.1.2.3 Comparison between SYBR Green and TaqMan assays for 16S DNA quantification ................................................................................................................................. 94

3.3.1.2.4 Optimisation and assay chemistry of SBL qPCRs .......................................................... 96

3.3.1.2.5 Correction of high baseline signals in SYBR Green assays ........................................ 97

3.3.2 Comparison between culture and qPCR for changes in TBL and SBL in stable and exacerbation visit ................................................................................................................................ 100

3.3.2.1 Detection rate of individual bacteria tested by qPCR and its concordance to culture result .............................................................................................................................. 100

3.3.2.2 Comparing qPCR counts of SBL between culture positive and culture negative samples ................................................................................................................................. 102

3.3.2.3 Change in the TBL of COPD samples between stable and exacerbation states ......... 104

3.3.2.4 Comparison of detection rate and quantitative readings of individual SBL between stable and exacerbation state ............................................................................................. 105

3.4 Discussion ........................................................................................................................................... 107

3.4.1 Developing and optimizing qPCR assays for quantifying TBL and for SBL in COPD sputum ........................................................................................................................................ 107

3.4.1.1 DNA Extraction Methods .................................................................................................. 107

3.4.1.2 Development and Optimisation of TBL and SBL qPCR .................................................... 109

3.4.2 Comparison between culture and qPCR for changes in TBL and SBL in stable and exacerbation visit .................................................................................................................... 110

3.4.2.1 Detection rate of individual bacteria tested by qPCR and its concordance to culture result .............................................................................................................................. 110

3.4.2.2 Change in TBL and SBL between stable and exacerbation samples ............................ 112

3.5 Conclusion .......................................................................................................................................... 114

4 16S RDNA BASED COPD BACTERIAL COMMUNITY ANALYSIS ................................................................................................................................. 115
4.1 Introduction ..................................................................................................................................... 115

4.2 Materials and Methods ...................................................................................................................... 121
  4.2.1 454 sequencing samples ............................................................................................................ 121
  4.2.2 Production of 16S rDNA bacterial amplicon library ................................................................. 122
    4.2.2.1 Primer design and sequence ............................................................................................... 122
    4.2.2.2 PCR conditions, Gel run and purification ......................................................................... 123
    4.2.2.3 DNA quantification, standardization and pooling ............................................................. 124
  4.2.3 Sequence data analysis using QIIME (Quantitative insights into microbial ecology) .............. 125
    4.2.3.1 Sequence processing and curation ..................................................................................... 125
    4.2.3.2 Bacterial community classification and diversity analysis ................................................. 126
  4.2.4 Statistical analysis and graphical output using other programs ................................................. 127

4.3 Results .............................................................................................................................................. 129
  4.3.1 Sequencing plate design, patient characteristics and pre-processing of reads 129
    4.3.1.1 Study design and sample sets for 454 sequencing: ......................................................... 129
    4.3.1.2 Raw sequence read depth and pre-processing of raw sequence files ................................ 130
  4.3.2 Taxonomic Characterisation of the bacterial community ........................................................... 133
    4.3.2.1 Phylum level classification ............................................................................................... 133
    4.3.2.2 Genus level classification ............................................................................................... 138
    4.3.2.3 Concordance between the 454 sequencing, qPCR and culture results for the 3 dominant genera .................................................................................................................. 142
    4.3.2.4 Changes in the bacterial community at different visit stages in SEFR set   ................................. 143
    4.3.2.5 Stability of microbial pattern in longitudinal stable samples in SSE set ............................ 148
  4.3.3 Sequence similarity based classification and diversity analysis of COPD microbiota .................. 149
    4.3.3.1 Alpha diversity measurement ......................................................................................... 150
    4.3.3.2 Beta diversity ................................................................................................................... 154
  4.3.4 Metadata based bacterial community analysis ............................................................................. 160
    4.3.4.1 Bacterial Community comparison between smokers and ex-smokers ........................... 160
    4.3.4.2 Bacterial community changes based on COPD severity .................................................. 163
    4.3.4.3 Treatment ....................................................................................................................... 166
4.3.5 *Streptococcus* OTU discrepancy - an observed artefact of the denoising algorithm ................................................................. 169

4.4 Discussion: ........................................................................................................................................................................... 174
4.4.1 Sequencing plate design and pre-processing of reads ............................................................ 174
4.4.2 Taxonomic Characterisation of the bacterial community ..................................................... 175
4.4.3 Sequence similarity based classification and diversity analysis of COPD microbiota .............................................................................................................................................................................................................. 177
4.4.4 Metadata based bacterial community analysis ..................................................................... 179
4.4.5 *Streptococcus* OTU discrepancy - an observed artefact of the denoising algorithm .............................................................................................................................................................................................................. 182

4.5 Conclusions ........................................................................................................................................................................... 183

5 COMPARISON OF SEFR SAMPLES BASED ON THE RELATIVE ABUNDANCE OF THE 2 DOMINANT PHYLA .............................................................................................................................................................................................................. 184

5.1 Introduction: ........................................................................................................................................................................... 184

5.2 Materials and Methods .......................................................................................................................................................... 186
5.2.1 Study Sample ........................................................................................................................................................................ 186
5.2.2 qPCR based enumeration of Proteobacteria and Firmicutes ................................................. 186
5.2.2.1 DNA standards for quantification assays ........................................................................... 186
5.2.2.2 Primers ........................................................................................................................................................................... 186
5.2.2.3 Quantification Assays ....................................................................................................................... 187
5.2.3 Statistical analyses for sequencing and qPCR data ................................................................ 188

5.3 Results ............................................................................................................................................................................... 188
5.3.1 Stratification of samples based on two dominant phyla ......................................................... 188
5.3.1.1 454 sequencing derived Proteobacteria and Firmicutes P:F ratio cluster model .............................................................................................................................................................................................................. 188
5.3.1.2 Comparison of the P:F ratio between clusters in individual visit groups ...................... 191
5.3.1.3 Within-Cluster changes in the microbiome across visits ................................................. 192
5.3.1.4 Baseline clinical characteristics of the three P:F clusters ................................................. 194
5.3.2 Validating P:F microbiome cluster by qPCR ........................................................................ 195
5.3.2.1 Developing 16S rDNA based Proteobacteria and Firmicutes qPCR ... 195
5.3.2.2 Correlation and validation of qPCR derived P:F ratio ...................................................... 199
5.3.2.3 Clustering of all 30 subjects based on qPCR P:F ratio ...................................................... 201
5.3.2.4 Changes in the qPCR based P:F ratio across visits in 30 subjects .... 203
5.3.2.5 Clinical characteristics of qPCR derived P:F clusters............................. 203
5.3.3 Grouping of COPD subjects based on PCoA pattern across visit in subjects 207

5.4 Discussion ........................................................................................................... 210
5.4.1 Stratification of samples based on two dominant phyla ............................. 210
5.4.2 Validating P:F microbiome cluster by qPCR ........................................... 212
5.4.3 Grouping of COPD subjects based on PCoA pattern across visit in subjects 214

5.5 Conclusions ......................................................................................................... 214

6 CONCLUDING REMARKS AND FUTURE WORK ............................................. 215

6.1 Limitations of the study ....................................................................................... 218
6.1.1 Sample and clinical study design ................................................................. 218
6.1.1.1 Sample type ............................................................................................ 218
6.1.1.2 Sample collection and storage ................................................................. 219
6.1.2 Factors contributing to bias and variability in molecular studies:.............. 220
6.1.2.1 DNA extraction ...................................................................................... 220
6.1.2.2 qPCR reaction variability: .................................................................... 221
6.1.2.3 16S rDNA based microbial community PCR reaction .......................... 222
6.1.2.4 16S rDNA primers and 16rDNA hypervariable region for bacterial community study: ................................................................. 223
6.1.2.5 Analysis pipelines ................................................................................ 224

6.2 Conclusions ......................................................................................................... 224

6.3 On-going and Future work .................................................................................. 225
6.3.1 COPD microbiome and P:F ratio based analysis with larger sample cohort: 225
6.3.2 Deep sequencing for greater sampling depth to understand the role of rarer species in COPD microbiome: ................................................................. 225
6.3.3 PMA based microbial community analysis: .............................................. 226
6.3.4 Whole genome sequencing (WGS) and Metatranscriptomics: .................. 226

REFERENCES ........................................................................................................... 227
APPENDICES ........................................................................................................254

Additional tables and figures .............................................................................254

Definitions ...........................................................................................................258

Scripts for Qiime .................................................................................................260
List of Figures

Figure 1.1: Vicious Cycle Hypothesis ................................................................. 33
Figure 1.2: 16S rDNA primary structure .......................................................... 42
Figure 1.3: Real-time PCR amplification curves in linear and logarithmic scale ...... 45
Figure 1.4: Amplification signals and linear regression graph from DNA standards .... 48
Figure 1.5: SYBR Green and TaqMan fluorescence chemistry .......................... 50
Figure 1.6: Flowgram of 454 sequencing .......................................................... 63
Figure 3.1: Correlation between TaqMan and SYBR Green 16S qPCR ............... 94
Figure 3.2: Bland Altman plot of agreement between SYBR Green and TaqMan 16S assay ................................................................. 95
Figure 3.3: Comparing the oligo sequence used for SYBR Green and TaqMan 16S qPCR ................................................................. 96
Figure 3.4: SYBR green qPCR producing high baseline signals ....................... 98
Figure 3.5: Difference in quantification readings for high baseline samples with different normalization methods used ............................................. 99
Figure 3.6: Detection rate of pathogens and comparison of the detection pattern between culture and qPCR ................................................................. 101
Figure 3.7: Comparison of qPCR counts between culture positive and culture negative samples for individual pathogen ........................................... 103
Figure 3.8: TBL in COPD sputum at stable and exacerbation visit ...................... 105
Figure 3.9: Comparison of the qPCR counts of pathogens between longitudinal stable and exacerbation samples ........................................... 107
Figure 4.1: 454 fusion primer design and sequence ........................................ 123
Figure 4.2: Flowgram of microbial community 16SrDNA sequence analysis ......... 127
Figure 4.3: Schematic representation of the sample set used for 454 sequencing ...... 133
Figure 4.4: Bacterial community structure at phylum level in each visit type of the SEFR set ................................................................. 134
Figure 4.5: Prevalence of bacterial phyla in samples of the SEFR set ............... 135
Figure 4.6: Dominance of Proteobacteria and Firmicutes in individual samples ..... 135
Figure 4.7: Bacterial community structure at phylum level in SSE and NES sets .... 136
Figure 4.8: Prevalence of bacterial groups in SSE and NES sets ....................... 137
Figure 4.9: Bacterial composition and their relative abundance in individual samples in SSE and NES set samples ............................................... 137
Figure 4.10: Average relative abundance, Prevalence and Total abundance in SEFR set ................................................................................................................................................................................139
Figure 4.11: Bacterial composition at genus level in individual samples in SEFR set 140
Figure 4.12: Average relative abundance, Prevalence and Total abundance in SSE and NES sets................................................................................................................................................................................................................................................................................................................................................................................141
Figure 4.13: Bacterial composition at genus level in individual samples in SSE and NES sets ................................................................................................................................................................................................................................................................................................................................................................................142
Figure 4.14: Samples form three major cluster based on Proteobacteria and Firmicutes dominance rather than SEFR visit types ................................................................................................................................................................................................................................................................................................................................................................................145
Figure 4.15: Separation of SEFR samples influenced by dominance of three genera rather than visit types ................................................................................................................................................................................................................................................................................................................................................................................146
Figure 4.16: Low diversity and bacterial richness in Haemophilus dominant samples 147
Figure 4.17: Hierarchical clustering of the COPD microbiota in longitudinal stable samples ................................................................................................................................................................................................................................................................................................................................................................................148
Figure 4.18: Comparison of bacterial community between exacerbated and non-exacerbated stable samples ................................................................................................................................................................................................................................................................................................................................................................................149
Figure 4.19: Rarefaction plot present a wide variation in exacerbation samples among the 4 visit type ................................................................................................................................................................................................................................................................................................................................................................................151
Figure 4.20: No difference in mean observed species richness across the 4 visit types ................................................................................................................................................................................................................................................................................................................................................................................151
Figure 4.21: Rarefaction curves of stable samples in SSE and NES sets ................................................................................................................................................................................................................................................................................................................................................................................153
Figure 4.22: NES set showed a lower bacterial richness trend compared to SSE set stable samples ................................................................................................................................................................................................................................................................................................................................................................................153
Figure 4.23: No significant difference in richness and diversity measures at multiple stable points in SSE and NES stable samples ................................................................................................................................................................................................................................................................................................................................................................................154
Figure 4.24: Negative association between Haemophilus and other dominant bacterial groups ................................................................................................................................................................................................................................................................................................................................................................................................................................................155
Figure 4.25: PCoA plot on unweighted UniFrac distance in SEFR visit types ....... 156
Figure 4.26: Boxplots of phylogenetic distance measure within visit type in SEFR ... 157
Figure 4.27: Box plot of distance measure between visits in each individual .......... 157
Figure 4.28: PCoA plot of SSE set based on phylogenetic distance measure .......... 159
Figure 4.29: Comparison of beta diversity pattern of NES samples and SSE samples 159
Figure 4.30: PCoA plot comparing diversity pattern of samples between two 454 runs ................................................................................................................................................................................................................................................................................................................................................................................160
Figure 4.31: Changes in relative abundance of phyla across SEFR visit type between smokers and ex-smokers ................................................................. 161
Figure 4.32: Comparison of between sample diversity measures based on smoking status .................................................................................................................. 163
Figure 4.33: Changes in relative abundance of phyla across SEFR visit types between COPD severity stages .................................................................................................................. 164
Figure 4.34: Comparison of between sample diversity measures based on COPD severity .................................................................................................................. 166
Figure 4.35: Changes in relative abundance of phyla across SEFR visit types between treatment groups .................................................................................................................. 167
Figure 4.36: Comparison of between sample diversity measures based on treatment received at exacerbation .................................................................................................................. 169
Figure 4.37: Beta diversity PCoA plots of SEFR group using Bray-Curtis and Morisita-Horn .................................................................................................................. 170
Figure 4.38: Alignment of the dominant reads from the two OTU groups .................. 173
Figure 4.39: PCoA plots using non-phylogenetic distance measure on Ampliconnoise corrected samples .................................................................................................................. 174
Figure 5.1: Classification of samples by ratio of relative abundance of Proteobacteria to Firmicutes .................................................................................................................. 190
Figure 5.2: The three P:F clusters and their membership confirmed by bootstrapping test .................................................................................................................. 190
Figure 5.3: Changes in P:F ratio between clusters and their concordance with OTU (0.03) level microbial composition .................................................................................................................. 192
Figure 5.4: Change in the P:F ratio across 4 visits within each cluster .................... 193
Figure 5.5: Within-subject changes in the microbiome at OTU level in each P:F microbiome cluster .................................................................................................................. 194
Figure 5.6: Performance of Firmicutes and Yproteo primer pairs in qPCR of mixed cultures .................................................................................................................. 198
Figure 5.7: Correlation between 454 and qPCR for P:F ratio .................................... 200
Figure 5.8: Comparison of cluster models based on P:F ratio obtained by qPCR and microbiomics .................................................................................................................. 200
Figure 5.9: Grouping of qPCR derived P:F ratio subjects labelled according to 454 P:F clusters .................................................................................................................. 201
Figure 5.10: Clustering of 30 subjects based on qPCR P:F ratio .................................. 202
Figure 5.11: Change in the qPCR derived P:F ratio across 4 visits within each cluster ................................................................................................................................. 203
Figure 5.12: Grouping of subjects based on change in microbiome at OTU level................................................................. 208
Figure A1.1: Variability between fresh, frozen and automated DNA extraction methods ................................................................................................................................................................................................. 254
Figure A1.2: QPCR interassay variability ................................................................................................................................. 254

List of Tables
Table 1.1: GOLD stages of COPD severity ................................................................................................................................. 26
Table 1.2: Classification of COPD exacerbation based on Anthonisen criteria ................................................................................................. 29
Table 1.3: Comparison of SYBR Green and TaqMan qPCR chemistry ................................................................................................. 51
Table 1.4: Other molecular techniques for microbial quantification ................................................................................................................................. 53
Table 1.5: Comparison of sequencing technologies ................................................................................................................................. 60
Table 3.1: Primer and probe sequence for TBL and SBL qPCR ................................................................................................................................. 81
Table 3.2: Bacterial DNA standards used for TBL and SBL qPCR ................................................................................................................................. 84
Table 3.3: Efficiency of Qiagen and microLYSIS DNA extraction method ................................................................................................. 89
Table 3.4: Representative sputum samples comparing 16S amplicon copies using different extraction protocol ................................................................................................................................................................................................. 89
Table 3.5: Optimization of primer pair concentration for 16S SYBR Green assay ................................................................................................. 91
Table 3.6: Optimisation of primer pair and probe concentration for 16S TaqMan assay ................................................................................................................................................................................................. 92
Table 3.7: 16S rRNA gene copy number per genome in bacteria associated with COPD ................................................................................................................................................................................................. 93
Table 3.8: Specificity of individual specific bacterial load qPCR ................................................................................................................................. 97
Table 3.9: Numerical summary of pathogen positive samples detected by both methods ................................................................................................................................................................................................. 101
Table 3.10: DNA inhibition for samples identified to be qPCR negative but culture positive ................................................................................................................................................................................................. 102
Table 3.11: ROC curve table based on quantitative readings for individual pathogen between culture positive and culture negative samples ................................................................................................................................................................................................. 104
Table 3.12: Detection rate of pathogens in paired stable and exacerbation samples by qPCR and culture ................................................................................................................................................................................................. 106
List of Equations

Equation 1.1: Mathematical principle of PCR.......................................................... 46
Equation 1.2: Quantification of initial target molecule by Ct method....................... 46
Equation 1.3: Standard curve equation for absolute quantification.......................... 47
Equation 3.1: DNA concentration measurement..................................................... 80
Equation 3.2: Determining genome equivalent for pure culture standards .............. 83
1 Introduction

1.1 General introduction

Bacteria are single celled organisms that have an influence on every aspect of ecology and life (Pace, 1997). These microscopic organisms, first observed in the 17th century by Antonie Van Leeuwenhoek, are able to colonise almost everywhere from soil, water, atmosphere and living matter to extreme and apparently inimical growth conditions (Makarova et al., 2001; Nicholson et al., 2013).

Bacteria play a fundamental role in maintaining the equilibrium of the ecosystem through cycling essential nutrients like nitrogen, carbon, oxygen and sulphur as part of fixation and decomposition processes. Bacterial assemblages present in living systems preserve health by assisting in the supply of essential nutrients and vitamins, the regulation of metabolism and protection against invasion by pathogens.

An aggregate of the microbial constituents present in an environment constitutes its microbiome or the microbiota. The human body harbours, 10 fold more microbial cells than human cells, that is about 10-100 trillion microbes (Foxman et al., 2008; Ursell et al., 2012b). An array of diverse bacterial groups exists on skin and in oral and vaginal cavities, which assist in maintaining pH balance, production of antimicrobials and competitive colonisation provide a healthy host environment (Grice and Segre, 2011; Wade, 2013). The human gut bacterial community plays an important role in host nutritional requirements by synthesizing vitamins, fermenting complex carbohydrates, assisting in the absorption of essential minerals. It is well recognised that the gut biota in health is essential for development and maturation of both the host gastrointestinal and immune systems (Cummings and Macfarlane, 1997; Round and Mazmanian, 2009; Witkin et al., 2007). In this context, studies have identified associations between constitution of the gut microbiome and host body mass index, ageing and development of allergy (Björkstén et al., 2001; Ley et al., 2006; Shanahan, 2013).

As pathogens, bacteria are a cause of considerable morbidity and mortality. There is increasing recognition of the diverse influence of bacteria on disease processes. On the
one hand, pathogenic bacteria are a direct cause of infectious disease. The causative relationship between a microbe and a disease was first established based on the four criteria of the Henle-Koch postulates:

1. The microorganism must be present in abundance in all cases of the disease.
2. Isolation and growth of the disease causing microorganism in the pure cultures should be achieved.
3. The cultured microorganism should cause disease when introduced into a healthy organism.
4. The microorganism must be re-isolated from the inoculated, diseased experimental host and identified as being identical to the original specific causative agent.

These postulates are based on the idea that an individual organism is the causative agent of a specific disease. This simplistic view has been increasingly challenged with conflicting observations, notably the presence of disease causing microbes in asymptomatic carriers and failure to grow many infectious microbes *in vitro*. With the development of molecular analysis techniques it has become apparent that, more than one microbial constituent and the interaction between these constituents in a microbial community are associated with many disorders (Sekirov and Finlay, 2006; Shanahan, 2013; Wade, 2013). In this context, there is increasing awareness of the importance of bacterial communities in the pathogenesis of both communicable and non-communicable diseases through disruption of the assemblages that maintain the balance of ecosystems for health (Foxman et al., 2008; Young et al., 2008). A typical example of this is colitis caused by *Clostridium difficile* colonisation following antibiotic therapy (Bartlett, 1994).

The evaluation of microbial communities is complicated by their complexity and diversity (Ursell et al., 2012a). Recent advances in techniques for profiling microbial communities may help to better understand their role in health and disease, informing new strategies for treatment. My project is concerned with the bacterial community
associated with, Chronic obstructive Pulmonary disease (COPD), a human respiratory condition.

1.2 Chronic Obstructive Pulmonary Disease (COPD)

1.2.1 What is COPD?
COPD is an inflammatory disease of the respiratory tract characterized by chronic inflammation of the airways leading to fixed airflow obstruction and restricted gas exchange. The term COPD includes two basic respiratory pathologies: chronic bronchitis and emphysema (GOLD, 2006) that are described below. The condition is not reversible and may progressively worsen with time (NICE, 2010).

Chronic bronchitis: is the narrowing of the bronchi and bronchioles due to thickening of the airway walls and enlargement of mucus glands from chronic inflammation (Hogg, 2008). COPD caused by chronic bronchitis is characterized by excessive mucus production and expectoration, with mucus plugging contributing to obstruction of the bronchial tree, leading to symptoms of wheeze and shortness of breath.

Emphysema: is characterized by destruction of the lung parenchyma and alveolar enlargement leading to loss of alveolar attachments, loss of functional gas exchange surface and loss of elastic tissue recoil (Cosio Piqueras and Cosio, 2001). Together these changes lead to airflow obstruction and breathlessness.

1.2.2 COPD definition and stages
Clinically, COPD and its severity are defined by the presence and magnitude of airflow obstruction. This is assessed with spirometry. Airflow obstruction is measured, post-bronchodilator, as the ratio of forced expiratory volume in one second (FEV1) to forced vital capacity (FVC), with a value of < 0.7 indicating presence of airflow flow obstruction and therefore COPD (GOLD, 2006). The severity of COPD is clinically categorised into four stages by the Global initiative for Obstructive Lung Disease (GOLD), based on the predicted percentage (%) of normal FEV1, in subjects presenting with < 0.7 FEV1/FVC post-bronchodilator ratio (GOLD,2006) (Table 1.1).
Table 1.1: GOLD stages of COPD severity

<table>
<thead>
<tr>
<th>COPD groups</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 1</td>
<td>Mild</td>
<td>FEV1 ≥ 80% predicted</td>
</tr>
<tr>
<td>Stage 2</td>
<td>Moderate</td>
<td>50% ≤ FEV1 &lt; 80% predicted</td>
</tr>
<tr>
<td>Stage 3</td>
<td>Severe</td>
<td>30% ≤ FEV1 &lt; 50% predicted</td>
</tr>
<tr>
<td>Stage 4</td>
<td>Very severe</td>
<td>FEV1 &lt; 30% predicted or FEV1 &lt; 50% predicted accompanied with chronic respiratory failure</td>
</tr>
</tbody>
</table>

All the stages had FEV1/FVC ratio <0.7, post-bronchodilator

Although a useful guide, severity assessment based on FEV1 alone is considered a weak marker for predicting morbidity and mortality as there is a poor association between extent of airflow obstruction and common COPD symptoms like breathlessness, exercise capacity and health status, (Celli et al., 2004). Several other measures such as the BODE (body mass index (B), airflow obstruction (O), dyspnoea (D), exercise capacity (E)) and ADO (age (A), dyspnoea (D) and airflow obstruction (O) indices have been developed and are superior for accurately defining disease severity (Puhan et al., 2009).

1.2.3 COPD symptoms
COPD patients experience shortness of breath and suffer from tightness in their chest, which leads to wheezing and laboured breathing. Damage to the lung parenchyma and recoiling mechanisms causes poor oxygenation of blood and accumulation of carbon dioxide (GOLD, 2006). COPD is characterized by persistent cough, sputum production due to excess mucus secretion by damaged airways and impaired ciliary mucus clearance. Expectoration of green or yellow coloured purulent sputum is also observed in some cases and has been associated with bacterial infective COPD (Miravitlles et al., 2010). All these symptoms lead to COPD patients having an impaired quality of life, weight loss, anxiety, depression and fatigue.

1.2.4 COPD burden
COPD is a major cause of morbidity and mortality both in well- and poorly- resourced countries. COPD is one of the leading causes of death contributing 5.8% of all deaths and, globally, ranks fourth (WHO, 2008). In the UK it ranks third among respiratory
diseases and is the fifth most frequent cause of death (National Clinical Guideline Centre, 2010). The World Health Organization (WHO) predicts that COPD will be the third biggest killer globally by 2030.

The economic burden of healthcare for COPD is substantial. In England, COPD is one of the most costly inpatient conditions treated and is the 2nd commonest cause of emergency admission to hospital (National Clinical Guideline Centre, 2010). The economic impact includes both healthcare costs and costs associated with loss of earnings and productivity due to morbidity and premature mortality in patients of working age (National Clinical Guideline Centre, 2010).

The prevalence of COPD is grossly underestimated due to sufferers being unaware, stigma associated with smoking related diseases and misdiagnosis of the condition when healthcare advice is sought. It is estimated that in the UK about two thirds of COPD cases are undiagnosed (British Lung Foundation, 2007), representing a reservoir of disease with implications for future healthcare. Many patients will eventually present at a late stage of disease and require hospital admission. COPD is also associated with several comorbidities including cardiovascular disease, cancer and mental health problems (Decramer and Janssens, 2013). The cost and complexity of care escalates with the number of these co-morbid conditions.

1.2.5 COPD cause and risk factors
Several environmental and genetic factors play a key role in development of COPD. Topmost among these are smoking and working in a polluting environment.

**Tobacco smoking and inhalation**: Approximately 85% of COPD cases are attributed to smoking making it the leading risk factor for COPD (Department of Health, 2012). Tar and harmful chemicals released from cigarettes and tobacco smoke causes damage to all parts of respiratory system from the major to smaller peripheral airways and gas exchange areas. Damage can be caused by both active and passive smoking. Smokers compared to non-smokers show significant shortening and reduced beating of their airway cilia, leading to impairment of mucocilliary defence and increased risk of infection (Leopold et al., 2009). Long term smoking causes hyper-mucus secretion, infiltration of inflammatory cells and many of the structural changes described for COPD. Smoking is a global problem and mortality from COPD is projected to increase
by more than 30% in the next decade without interventions to reduce the number of smokers (WHO, 2012).

**Coal, Air pollutants and other harmful toxic fumes:** COPD is also commonly seen in coal miners due to long term exposure to coal. Long term and persistent exposure to indoor pollutants such as biomass fuel used for cooking and heating is a significant risk factor in developing countries. Outdoor pollutants like toxic chemical fumes and fumes associated with vehicular pollution are also potential risk factors for COPD (WHO, 2012).

**Other factors:**

**Genetic factors:** Alpha-1 antitrypsin deficiency is a rare autosomal recessive genetic condition that is more common in Europeans and associated with COPD and pulmonary emphysema (GOLD, 2006).

**Childhood Infection:** Frequent chest infection in childhood (Sethi, 2000a) and impaired lung growth and development at birth are also implicated in COPD (Barker et al., 1991).

**Gender:** Although previously COPD was more commonly seen in men, increased use of tobacco among women has led to the COPD prevalence rate to now be almost equal between men and women (National Clinical Guideline Centre, 2010). It has also been suggested that women are genetically more susceptible to smoking related risks compared to men (Silverman et al., 2000).

**Asthma:** This is another chronic inflammatory disease of the airways and shares some characteristics with COPD. Severe asthma is also considered a risk factor for developing COPD (Silva et al., 2004).

**1.2.6 COPD exacerbation**

COPD that is in a controlled state with the prescribed medication is termed stable while flaring up of symptoms that is acute and beyond normal day-to-day variations of an individual's stable state is known as exacerbation (National Clinical Guideline Centre, 2010). Symptoms experienced during exacerbation are, in general, worsening breathlessness, cough, and increased sputum production and change in sputum colour.
Exacerbations are a major cause of morbidity and mortality in COPD and can accelerate disease progression.

Diagnosis of exacerbation severity and treatment decisions are based on fulfilling at least one of the three criteria proposed by Anthonisen (Anthonisen et al., 1987): increased dyspnoea, increased sputum production and increased sputum purulence compared to stable state (Table 1.2). Presence of only one of the criteria is considered a mild exacerbation that can be treated with corticosteroids alone while more than one symptom is treated with either antibiotics alone or a combination of antibiotics and steroids (Anthonisen et al., 1987). In some studies, exacerbation severity is categorized according to whether the subject self-medicated (mild), accessed healthcare support as an outpatient (moderate) or required hospital admission (severe) (Hurst et al., 2010).

<table>
<thead>
<tr>
<th>Severity of exacerbation</th>
<th>Type of exacerbation</th>
<th>Characteristics</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severe</td>
<td>Type 1</td>
<td>Increased dyspnea, sputum production and purulence</td>
<td>Combination of Steroid and antibiotics</td>
</tr>
<tr>
<td>Moderate</td>
<td>Type 2</td>
<td>Any 2 of the above 3 symptoms</td>
<td>Antibiotics only</td>
</tr>
<tr>
<td>Mild</td>
<td>Type 3</td>
<td>Any 1 of the above 3 symptoms along with 1 or more of following minor symptoms: Cough, Wheezing, Fever without an obvious source, Respiratory and/or Heart rate increase &gt;20% over baseline.</td>
<td>Steroid only</td>
</tr>
</tbody>
</table>

Table 1.2: Classification of COPD exacerbation based on Anthonisen criteria

Table is based on the data from paper (Anthonisen et al., 1987)

Exacerbations are considered common in moderate and severe (GOLD stage 2 to 4) cases of COPD, with increase in both mean exacerbation frequency and proportion of COPD subjects exacerbating, with increasing severity of COPD stage (Hurst et al., 2010). Hospitalization due to exacerbation was reported to be nearly fivefold higher in COPD with stage 4 severity compared to stage 2 (Hurst et al., 2010). However, there
exists a wide variation in exacerbation frequency with many patients never or very infrequently having exacerbations to a few experiencing them almost every month.

The socioeconomic burden of COPD exacerbations is considerable. While mild to moderate exacerbations may be managed with a change in medication costing approximately £100 per patient on average, severe exacerbations that require hospital admissions are very expensive costing thousands of pounds per patient. A quarter of COPD hospital admission cases die within a year of admission while 15% die within three months of hospital admission (Department of Health, 2012).

It is likely that there are multiple triggers for COPD exacerbations. However, infections caused by bacteria and viruses are thought to be associated with the majority (~80%) of exacerbation events (Murphy et al., 2000; Sethi, 2000a, b). Around 50% of COPD exacerbations are considered to be bacteria related with isolation of *H. influenzae*, *S. pneumoniae*, *M. catarrhalis*, *P. aeruginosa* and *S. aureus* most frequent. Viruses such as rhinovirus, coronavirus, and parainfluenza contribute 30%, while atypical bacteria (*Mycoplasma pneumoniae*, *Chlamydia pneumoniae*) are reported to be isolated in a minority of COPD exacerbations by some authors (Lieberman et al., 2002; Sethi, 2000b). In contrast, Erkan et al reported the presence of atypical bacteria (*C. pneumoniae* and *M. pneumoniae*) in an unpredictably high proportion of 26% of exacerbations in their study (Erkan et al., 2008).

In up to 30% of exacerbation cases the cause is unidentified (National Clinical Guideline Centre, 2010). Exacerbations may also be result of a congestive heart failure, systemic infections, pulmonary embolism, pneumonia, air pollution, cold air, allergies, and/or smoking.

COPD exacerbations are multifactorial and heterogeneous in nature and sometimes more than one factor may be concurrently contributing to the episode (Gao et al., 2013; Sethi et al., 2009). The challenge and uncertainty of establishing the cause of exacerbation is partly due to presence of both infective and non-infective factors at stable state. In one study, pathogenic respiratory bacteria have been isolated from 25% of stable subjects, (Monsó et al., 1999). In this setting the significance of isolating the same bacteria at exacerbation is uncertain. There is also considerable debate over the definition of severity gradation and mechanism of exacerbation (Celli and Barnes, 2007). Unlike the defined stable COPD stages based on FEV1 predicted values, there is
Introduction

no defined objective measurement for exacerbation severity apart from the number of new symptoms identified. Severity of exacerbation is dependent on the underlying disease severity at stable state as well as the acute changes during exacerbation (Siddiqi and Sethi, 2008). Thus, for subjects with severe COPD, a relatively small change from baseline may lead to exacerbations with significant clinical symptoms while milder disease may tolerate the effect of triggers better and cope with a much larger change in symptoms from their baseline.

1.2.7 COPD treatment

**Smoking cessation**: For current smokers with COPD, stopping smoking can help with decreasing the rate of lung function decline and further damage to the airways. Nicotine replacement therapy and counselling are provided and can be effective.

**Immunization**: COPD patients are susceptible to infections, due to impaired airway defence and mucociliary clearance. Influenza and pneumococcal vaccines are recommended for all COPD patients. These vaccines can reduce the risk of infective exacerbations (Varkey et al., 2009).

**Inhaled Bronchodilators**: These drugs are designed to improve airflow obstruction and help with symptoms of wheeze and breathlessness. All COPD patients are given short-acting bronchodilators. Depending on the severity of COPD, patients may also be given a combination of one or more long-acting bronchodilators as well (NICE, 2010).

**Steroids and Antibiotics**: As mentioned in section 1.3.6, these drugs are prescribed during COPD exacerbations according to Anthonisen criteria. In addition, exacerbations may also need oxygen therapy.

Surgery is also considered in some patients with very severe COPD.

1.3 COPD and Bacteriology

1.3.1 The respiratory microbiome in healthy individuals

The normal microbiome of the healthy upper respiratory tract (URT), comprising the nose, mouth, nasopharynx and oropharynx, harbours an extremely varied bacterial community (Austrian, 1968). The oral cavity is densely populated with a complex
microbial community that is composed mainly of bacterial groups belonging to Firmicutes, Bacteroidetes, Proteobacteria and Actinobacteria phyla (Aas et al., 2005). At genus level *Streptococcus*, *Gemella*, *Granulicatella*, *Veillonella*, *Prevotella*, *Porphyromonas*, *Neisseria* and *Rothia* are among the most abundant and widespread constituents (Aas et al., 2005; Bik et al., 2010). Firmicutes and Actinobacteria represent the most dominant phyla in the nose with *Staphylococcus*, *Corynebacterium* and *Propionibacterium* being the most prevalent groups while the nasopharynx is predominantly colonised with non-haemolytic and alpha-haemolytic *Streptococcus* and *Neisseria* species (Austrian, 1968; Beasley et al., 2012; Lemon et al., 2010). Pathogens such as *Streptococcus pneumoniae*, *Streptococcus pyogenes*, *Haemophilus influenzae* and *Neisseria meningitidis* are also present in the pharynx (Austrian, 1968; Lemon et al., 2010).

In contrast, the lower respiratory tract has been considered to be sterile in healthy non-smokers, due to mechanical clearance of bacteria by ciliated epithelium to prevent colonisation. Lower respiratory tract sampling with bronchoscopy has cultured potential pathogens in <10% of healthy population at very low bacterial load (Beasley et al., 2012). However application of molecular techniques has highlighted the short-comings of standard culture, identifying dominance of anaerobes from the *Prevotella* and *Veilonella* genera in healthy lungs (Charlson et al., 2011; Free, 2005; Hilty et al., 2010).

### 1.3.2 Bacterial role in COPD

In the pathogenesis of COPD, bacteria are considered a secondary factor, with chronic exposure to cigarette smoke and other irritants leading to damage of primary airway defences and allowing chronic colonisation with potentially pathogenic bacteria (PPB). The role of bacteria in the development and progression of COPD is not well understood. The ‘Vicious Cycle Hypothesis’ of Sethi (Sethi, 2000a) proposes that chronic bacterial infection of the airway is a significant factor in the progressive decline of lung function. Expression of toxic bacterial products causes epithelial damage and impaired ciliary clearance and makes the airway susceptible to further colonisation by pathogens. Additionally, low grade infection propagates chronic airway inflammation that acts synergistically to cause airway damage and lung function decline (Figure 1.1).
In vitro and animal model based bacterial studies have shown bacterial strains belonging to non typeable *H. influenzae (NTHI), S. pneumoniae, M. catarrhalis and P. aeruginosa* are the major and commonly isolated pathogens in COPD. In addition, *Haemophilus parainfluenzae, Staphylococcus aureus* and members of the family *Enterobacteriaceae* have also been reported in stable COPD (Siddiqi and Sethi, 2008) (Banerjee et al., 2004). Products released by these pathogens cause mucus hypersecretion, reduction in ciliary beat frequency and airway epithelial injury (Adler et al., 1986; Read et al., 1991; Wilson et al., 1985). Bacterial products also stimulate airway inflammation with host neutrophil influx and release of neutrophil elastase that disrupts the elastase-antielastase balance causing emphysema and airway narrowing (Sethi, 2000a).

There is a body of evidence to support a role for bacteria in both stable and exacerbation state of COPD. Potentially pathogenic bacteria (PPB) have been shown to be present in approximately 25% to 35% of stable COPD (Monsó et al., 1999; Sethi et al., 2006). In one study, quantitative culture of samples obtained at bronchoscopy from the lower respiratory tract demonstrated colonisation with PPB in stable COPD patients, while
cultures from control subjects were negative (Cabello et al., 1997), suggesting an association of COPD with bacterial airway colonisation.

PPB in stable COPD are suggested to be colonisers although increase in host inflammatory biomarkers expression level associated with COPD pathogenesis has been shown in relation to pathogenic bacteria load. For example, detection of PPB in stable sputum and bronchoalveolar lavage (BAL) has been shown to be associated with higher percentage of neutrophils and tumour necrosis factor alpha (TNF-α) (Banerjee et al., 2004; Soler et al., 1999). A significant correlation between markers of neutrophilic airway inflammation and bacterial load, in sputum of stable COPD patients has also been reported, especially in samples with \( > 10^6 \) CFU/ml aerobic counts, (Hill et al., 2000). This group also showed an association between inflammation and the bacterial species identified with inflammation being greatest for Pseudomonas colonisation, followed by H. influenzae and M. catarrhalis.

The role of bacteria in acute exacerbations of COPD is poorly understood (Murphy et al., 2000). Increase in exacerbation frequency and accelerated lung function decline has been linked with bacterial colonisation at stable state (Patel et al., 2002; Wilkinson et al., 2003). Cross sectional studies of samples collected from sputum or bronchoscopy at the time of COPD exacerbation have reported evidence of bacterial infection with PPB such as H. influenzae, S. pneumoniae, M. catarrhalis, S. aureus and P. aeruginosa in about 50% to 70% (Ball, 1995; Sethi, 2000b; Soler et al., 1998). In a few other studies, both the rate of isolation of PPB and bacterial load were significantly greater in exacerbation samples than stable samples (Monsó et al., 1995; Rosell et al., 2005). However, there is some scepticism over these findings due to the cross-sectional study designs, resulting in the possibility of exacerbating individuals belonging to a population with higher bacterial carriage levels at stable state (Hirschmann, 2000).

In contrast, early longitudinal studies reported no difference in the rate of positive culture detection of stable and exacerbation samples and concluded that bacterial infection did not play a significant role in acute exacerbations (Gump et al., 1976; McHardy et al., 1980). Other longitudinal culture studies have shown association of exacerbation with increased prevalence and higher bacterial load of H. influenzae (Wilkinson et al., 2006) and S. pneumoniae numbers (Gump et al., 1976) compared to stable state.
Sethi et al have investigated further the role of bacteria in the pathogenesis of COPD exacerbations by using molecular typing methods to examine the change in strain of bacterial species across visits (Sethi et al., 2002; Sethi et al., 2007). The authors identified a new strain of a PPB, most often *H. influenzae* and *M. catarrhalis*, in the sputum of COPD patients at 33% of visits associated with isolation of a new strain compared with an exacerbation diagnosed at 15% of visits when no new strain was identified. The authors concluded that these results supported the view that acquisition of new strains are important in the pathogenesis of exacerbations (Sethi et al., 2002). However, identification of any pathogen, irrespective of whether it was a new strain was also more often seen for exacerbations and no comparison was made in the study between whether a new strain had or had not been detected among the subgroup associated with isolation of a PPB. In a follow-on study, Sethi et al. observed no change or a decrease in bacterial load, at exacerbation from stable state, with pre-existing strains of *H. influenzae*, *M. catarrhalis* and *S. pneumoniae* load, while new strains of *H. influenzae* and *M. catarrhalis* were detected more frequently (Sethi et al., 2007). However, the differences reported were small (< 1 log fold) and further studies are needed. Though not conclusive, studies using quantitative culture and molecular methods do support a role for acute bacterial infection during exacerbations (Garcha et al., 2012; Sethi et al., 2007; Wilkinson et al., 2006).

### 1.3.3 Controversy on the use of antibiotics in the treatment of COPD exacerbation

The effectiveness of antibiotic therapy for the treatment of COPD exacerbations is likely to be related to the contribution of bacterial infection in their pathogenesis. This is an important question as COPD exacerbations are heterogeneous and a blanket approach to treatment with antibiotics may not be appropriate.

Indeed, the use of antibiotics to treat exacerbations has been controversial due to conflicting and inconclusive results of studies investigating the role of bacteria at stable and exacerbation states (Sethi, 2004). The most successful study showing benefits from antibiotic treatment is the Anthonisen study. This led to the exacerbation type and therapy strategy being defined according to the number of symptoms diagnosed (Anthonisen et al., 1987). Nouira and co-workers demonstrated significant benefit from antibiotics in a COPD exacerbation study of patients requiring mechanical ventilation support, suggesting effectiveness of antibiotic treatment for severe cases (Nouira et al.,
However there is no study that shows how well the exacerbation treatment criteria correlate with the cause of the exacerbation. Meta-analysis of nine placebo-controlled antibiotic studies has reported conflicting results with either small (in 7 studies) or no benefits (in 2 studies) (Saint et al., 1995). Sethi et al. have highlighted a number of limitations with these antibiotic studies that may have contributed to inconclusive results (Sethi, 2004). There is also concern that bias may be introduced with inclusion of non-infectious aetiologies and mild exacerbations (Sachs et al., 1995).

There is growing concern regarding the increasing prevalence of multiple antibiotic resistant bacteria, and efforts need to be made to stop unnecessary antibiotic medication. While some support the use of antibiotics to treat COPD exacerbations (Murphy et al., 2000) others are not certain if the benefits outweigh the pitfalls of increasing antibiotic resistance (Hirschmann, 2000). Therefore it will be important to identify biomarkers or a bacterial signature that defines bacterially caused exacerbation that could benefit from antibiotic therapy.

**1.3.4 General problems with bacterial studies in COPD**

The studies described illustrate a number of problems that are encountered in the investigation of COPD exacerbations that make it difficult to characterize the role of bacteria. Firstly, with chronic bacterial colonisation at stable state, sensitive techniques are required to detect changes in the bacterial load and composition during exacerbations. As discussed below, this cannot be readily achieved with conventional culture techniques. Secondly, within a COPD population, there is wide variability and considerable overlap in the bacterial load and organisms detected at stable and exacerbation states. Cross sectional studies are therefore of limited value and ideally longitudinal studies are needed to detect changes within individuals between stable and exacerbation states. This is difficult to achieve practically and only a limited number of longitudinal studies have been performed to date, all of these in specialised centres. Further, all of these studies have been performed using culture methods, which have their own limitations. Finally, COPD exacerbations are considered to be heterogeneous as PPB are not isolated in all COPD exacerbation (Wilson, 1998).

Apart from bacterial culture studies and antibiotic trials, serological studies against bacterial antigens, sputum characteristics and measurement of airway inflammatory markers have also been utilised to determine the role of bacteria in exacerbations (Sethi,
Serology has been performed not only for atypical bacteria but also for testing *H. influenzae*. One important limitation of this technique is that different strains of bacteria such as *H. influenzae* have considerable variation in their surface antigenic structure, yet the majority of the serological studies have been flawed by the use of laboratory strains of *H. influenzae* as antigen instead of the strains recovered from exacerbations (Sethi, 2004). In addition to this, there are a lack of longitudinal studies and conflicting results with comparison of paired sera before and after onset of exacerbation (Sethi, 2004).

Sputum purulence and colour are claimed to be effective indicators of distinguishing bacterial and non-bacterial exacerbations of COPD (Miravitlles et al., 2010). Although sputum purulence and a number of pro-inflammatory markers have been associated with bacterial related COPD they have not proven satisfactory as specific indicators of bacterial related exacerbation. Sethi and co-workers demonstrated that *H. influenzae* and *M. catarrhalis* associated exacerbations were associated with significant increases of inflammatory cytokines compared with non-pathogenic exacerbations (Sethi et al., 2000). In contrast, although Aaron and co-workers showed an increase in the inflammatory cytokines at exacerbation from stable and post-exacerbation state, they did not show any difference in inflammatory markers between exacerbations that were considered to be either associated with bacterial infection or not (Aaron et al., 2001).

### 1.3.5 Limitation of culturing techniques used in characterizing COPD samples

The "Great Plate Count Anomaly", was a term coined by Staley and Konopka, referring to the observation that bacterial populations estimated by culture counts or most-probable-number (MPN) methods grossly underestimated the true bacterial count (Amann et al., 1995). The explanation for this discrepancy was the presence of either species which have not yet been cultured (NYC) due to lack of knowledge of suitable methods or known species that are viable but have entered a non-culturable state (VNC) (Amann et al., 1995).

Conventional culture methods have been used extensively for characterizing the bacterial community in COPD. However, these techniques have a number of limitations and on their own are inadequate to properly explore microbial communities and their role in COPD pathology. Culture techniques are labour intensive and time consuming.
with poor sensitivity. COPD samples are usually cultured aerobically, thus excluding anaerobes, at 37°C for 48hrs on three to four different plates to identify the major bacterial pathogens associated with respiratory disease. The most commonly used bacterial media plates are:

**Blood agar**: is used for isolating *S. pneumoniae*, *S. aureus* and *M. catarrhalis*, together with study of haemolytic characteristics for organism identification.

**Chocolate agar**: is specifically used for growth of *H. influenzae* as this organism requires haemin and nicotinamide-adenine-dinucleotide (NAD), released during heating of the media.

**Cysteine Lactose Electrolyte Deficient agar (CLED)**: is used for enumeration and growth of Gram negative bacteria

**MacConkey agar**: is used commonly for culturing coliforms.

Although the use of a number of different media allows the identification of a broad range of bacteria, bias clearly exists in the bacterial populations that may be detected and there is a high probability of missing potentially important novel populations that require different and often difficult growth conditions. Examples include bacteria that require extremes of temperature, a longer growth time or anaerobic conditions. Atypical bacteria such as *C. pneumoniae* and *M. pneumoniae* are difficult to isolate by culture but are commonly associated with respiratory tract infection and therefore likely to be relevant in the study of COPD (Lieberman et al., 2002; Sethi, 2000a). Recently it was reported that intermittent non-antibiotic treated samples, preceded and followed by positive cultures with the same *H. influenzae* strain, were *H. influenzae* negative by culture but tested positive using molecular methods (Murphy et al., 2004). Intracellular or biofilm associated bacterial populations have been reported as reasons for a failure to detect using culture methods (Legnani, 2009; Rayner et al., 1998). Biofilm formation by *P. aeruginosa* and *H. influenzae* strains has been demonstrated in Cystic Fibrosis (CF) and COPD patients (Martinez-Solano et al., 2008; Murphy and Kirkham, 2002; Singh et al., 2000). Evidence that non-typeable *H. influenzae* (NTHI) can penetrate and colonise in the sub-epithelial layers of the respiratory tract and adenoidal macrophages has also been demonstrated (Bandi et al., 2001; Forsgren et al., 1996). Furthermore, the presence of *H. influenzae* in a metabolically active but non-proliferative state within host cells,
likely a consequence of inadequate antibiotic treatment, would evade detection by routine culturing techniques (Morey et al., 2011; Roberts et al., 1984).

The poor sensitivity of culture methods is supported by studies that have compared this approach with molecular methods (discussed below). It is estimated that culture methods identify less than 1% of some bacterial populations (Klappenbach et al., 2001). The development of molecular quantification and sequencing techniques has resulted in new approaches to overcome the limitations of culture and use of some of these methods might provide more insights into role of bacteria in COPD.

1.4 16S rDNA as marker gene for enumeration and characterization of bacterial community

1.4.1 Molecular biology and classification
The discovery of helical DNA structure (Watson and Crick, 1953) and following this, extensive study by Kornberg characterising the mechanisms of DNA replication (Lehman, 2008) led to the birth of the molecular era of biology and development of two ground-breaking molecular biology techniques that still form the basis of modern molecular DNA technologies: *in vitro* sequencing of DNA (Sanger et al., 1977) and gene amplification with polymerase chain reaction (PCR) (Saiki et al., 1988).

The "molecular evolutionary clock" hypothesis (Zuckerandl and Pauling, 1965) and neutral theory of molecular evolution (Kimura, 1968) in eukaryotes laid the foundation for comparing structures with homologous informational macromolecules to establish phylogenetic relationships. Based on these principles Woese and Fox performed comparative analysis of a gene sequence coding for the ribosomal RNA (rRNA) (Woese and Fox, 1977). This work pioneered the building of a universal tree of life defining the three domains of life (Woese et al., 1990) and making this a key marker gene for bacterial identification enumeration and diversity studies (Pace, 1997; Tringe and Hugenholtz, 2008). In 1984 Bergey's manual of Systematic bacteriology was amended to include this rRNA molecule frame work for prokaryote classification (Hugenholtz, 2002).
1.4.2 Ribosomes

Three different rRNA molecules (5S, 16S and 23S) along with approximately 50 additional proteins form the complex macromolecular structure of the ribosome. Ribosomes are the protein-synthesis machinery and structurally are comprised of two subunits. The small subunit (SSU) binds to messenger RNA (mRNA) reading the genetic codon and the larger subunit (LSU) binds to the appropriate amino acid carrying transfer RNA (tRNA) extending the linear polypeptide chain during protein translation. Prokaryotes have 70S (S stands for Svedberg unit for sedimentation co-efficient) ribosomes, which consists of a 30S SSU that is itself made up of a 16S RNA subunit (consisting of ~1540 nucleotides) bound to 21 proteins. The 50S LSU is composed of a 5S RNA subunit (~120 nucleotides), a 23S RNA subunit (~2900 nucleotides) and 31 proteins (Pace, 1973).

1.4.3 16S rRNA gene: pros and cons

To characterize and record both quantitative and qualitative changes, of a bacterial community, a target molecule is required that is conserved across the bacterial kingdom but also has enough variability to allow bacterial profiling. In this context, the 16S rDNA, coding the SSU rRNA, is well suited and the most widely used gene for bacterial identification, quantification, assessing microbial diversity and defining phylogenetic relationship in ecological studies (Acinas et al., 2004). The past 3 decades has seen an exponential increase (465%) in the number of approved bacterial species from 1,791 to 8,168 based on 16S gene identification techniques (Janda and Abbott, 2007).

The 16S gene ticks all the essential criteria required for a comparative phylogenetic analysis between prokaryotes. Firstly, as a component of the cellular translational protein synthesis machinery it is universally distributed among all bacteria (Woese and Fox, 1977). Secondly, being a core informational gene coding for cellular function essential for sustenance of life, it is believed that most 16S regions would be resistant to horizontal gene transfer (HGT) events as they could compromise structural and functional integrity of the gene which might prove deleterious for the bacteria (Daubin et al., 2003; Jain et al., 1999; Woese, 1987). Thirdly, the slow evolutionary rate and functional constraint on the 16S rRNA domains mean that this gene sequence is characterised by stretches of highly conserved regions interspersed by variable regions, providing a wide phylogenetic range for comparative analysis (Sogin et al., 1971;
Woese, 1987; Woese et al., 1975). Finally, 16S rDNA is ~ 1540 bp long making it an ideal gene size for study using PCR, cloning techniques and sequence analysis.

Bacterial 16S rRNA gene consists of nine hypervariable (V1-V9) regions (Figure 1.2) that exhibit wide-ranging sequence variation among bacterial groups (Van de Peer et al., 1996) and can be utilised for targeted design of primers and hybridisation probes, enabling characterisation of samples at different levels of taxonomic specificity (Chakravorty et al., 2007; Greisen et al., 1994; Jonasson et al., 2002). These hypervariable regions are flanked by conserved stretches of nucleotide sequences and primers designed based on these sequences are useful for enumeration and profiling of mixed bacterial communities (Greisen et al., 1994; Lane et al., 1985; Munson et al., 2004; Nossa et al., 2010; Youssef et al., 2009).

Although the primers that target the conserved regions of 16S rDNA are known as universal or broad range primers, none of them have 100% bacterial coverage and there is considerable variability in their range of coverage (Wang and Qian, 2009). One of the major limitations of the 16S gene is the presence of multiple copies within a genome. The copy number varies from 1 to 15 among bacterial groups to species level (Crosby and Criddle, 2003; Farrelly et al., 1995). Acinas and colleagues have demonstrated that variation in the copy number of this gene exists even at strain level of many bacterial species (Acinas et al., 2004). Utilisation of the 16S is therefore problematic for bacterial enumeration studies, particularly of mixed bacterial samples where the genome content may be either overestimated or underestimated. Similarly, in bacterial community studies, this can generate a skewed bacterial composition favouring the high copy number bacterial groups. Presence of intra genomic heterogeneity in the 16S rDNA has also added to complexity with this marker gene based bacterial study leading to biases in bacterial community diversity estimates and on phylogenetic resolution at species and genus level (Acinas et al., 2004; Fox et al., 1992; Janda and Abbott, 2007; Větrovský and Baldrian, 2013).
**16S rDNA primary structure:** Picture adapted from (Nossa et al., 2010). The nine hypervariable regions (1-9) spanned nucleotides 69-99, 137-242, 433-497, 576-682, 822-879, 986-1043, 1117-1173, 1243-1294 and 1435-1465 for V1 through V9 respectively [numbering based on the E. coli system of nomenclature (Brosius et al., 1978)], flanked by the blue coloured conserved regions of the 16S rDNA gene. Arrows at the top show examples of the forward primers (F) regions and bottom arrows representing the reverse primers (R) regions for targeting 16S rDNA conserved regions.

**1.4.4 Other molecular markers**

To overcome the limitations of multiple copies and intragenomic heterogeneity of 16S rRNA, other marker genes have also been explored. One promising marker gene in this regard is the rpoB gene (Dahllöf et al., 2000; Mollet et al., 1997). The rpoB gene encodes for the β-subunit of the RNA polymerase (RNAP), an essential enzyme for DNA transcription, synthesising RNA molecules in bacteria (Borukhov and Nudler, 2003; Dahllöf et al., 2000). This gene has been used in phylogenetic studies of various bacterial groups such as *Mycobacterium* species (Ben Salah et al., 2008), *Staphylococcus* species (Rowland et al., 1993) and *Enterobacteriaceae* (Mollet et al., 1997). The rpoB gene was shown to produce phylogenetic resolution comparable with 16S rRNA gene at all taxonomic levels, with better resolution at subspecies level compared to 16S rDNA (Case et al., 2007). Another phylogenetic marker is the *recA* gene, from the recombinase A gene family, essential for DNA repair and maintenance (Eisen, 1995; Lloyd and Sharp, 1993). However, it is very difficult to design universal primers for phylogenetic studies with protein encoding genes due to saturation of all third codon positions over a long evolutionary timescale (Case et al., 2007).
Some studies have also explored the 5S rRNA gene but this gene is limited for informing phylogenetic relationships (Olsen et al., 1986; Stahl et al., 1985). The 23S rRNA gene and 16S-23S rRNA internal transcribed spacer sequences have also been used for discrimination between bacterial species and phylogenetic studies (Barry et al., 1991; Ludwig and Schleifer, 1994; Rantakokko-Jalava et al., 2000). A major limitation of these phylogenetic markers in culture-independent applications is the lack of extensive gene-specific reference databases against which sequences obtained from mixed communities can be compared.

1.4.5 16S gene based databases

The culture independent PCR amplification and sequencing of ssu-rRNA from an unprecedented variety of communities is enriching the 16S rRNA sequence collection as well as contributing significantly to the identification of novel bacteria (Cole et al., 2014; DeSantis et al., 2006). With next generation sequencing, billions of available partial ssu-rRNA sequences as well as completely sequenced genes have been added to the GenBank (Benson et al., 2009), the largest databank of nucleotide sequences, as "unclassified" or "environmental" sequences.

Apart from this there are many other curated databases and programs dedicated to developing tools for ecological studies that are based on rRNA sequence only. Among these a few frequently used are:

i. Ribosomal Database Project (RDP; http://rdp.cme.msu.edu/) with 2,809,406 aligned and annotated bacterial and archaeal small subunit (SSU) rRNA gene sequences (Cole et al., 2014)

ii. SILVA ssu-rRNA database (http://www.arb-silva.de) (Pruesse et al., 2007; Quast et al., 2013) which has a collection of around 3,194,778 small subunit and 288,717 large subunit rRNA gene sequences.

iii. The Greengenes databank (http://greengenes.lbl.gov) provides a chimera check along with aligned and a taxonomically classified list of 16S rRNA gene sequences (DeSantis et al., 2006).

iv. The rrndb database provides information on the number of rRNA operons contained in the genomes of prokaryotic microorganisms and is accessible at http://rrndb.cme.msu.edu(Klappenbach et al., 2001).
The cumulative outcome of this work has been expansion of previously known 11 bacterial phyla (Woese, 1987) to currently 52 phyla. Most of these divisions are dominated by uncultured organisms (Handelsman, 2004). Thus in addition to ssu- rRNA use as a phylogenetic marker, it also became a key “ecological marker” for studies of microbes.

1.5 Enumeration of bacterial community

1.5.1 Difficulties of Quantitative Analysis in Normal PCR

During a bacterial infection there is active cellular proliferation of the disease causing pathogen within the host. Although PCR is a highly sensitive method, it is limited for understanding longitudinal changes in bacterial load between samples at different time-points. Standard PCR fails to accurately represent the amount of starting template from amplified end product due to plateau phase at which a similar amount of amplified product will be obtained irrespective of the starting material (Elizabeth van, 2008). Furthermore, traditional PCR requires performing DNA electrophoresis to visualise the end product of the reaction on agarose gel, which has a poor resolution for quantifying the amount of DNA and is time consuming (Holland et al., 1991).

1.5.2 Real-Time PCR: chemistries, application and other methods

Real-Time PCR has introduced the opportunity for enumeration and characterization of specific or mixed bacterial populations by DNA directed quantitative PCR known as quantitative PCR (qPCR). RNA directed quantitative PCR is a related technique exploring levels of gene expression and is known as Reverse-transcriptase qPCR (RT-qPCR) (Gibson et al., 1996).

Quantitative PCR (qPCR) exploits the principle that there is a direct relationship between the amount of DNA present during the log phase of the amplification cycle and the starting amount of DNA of the target gene (Heid et al., 1996). This is achieved by detection of a fluorescent signal that accumulates in direct proportion to the yield of amplified PCR product in every cycle using either dsDNA binding fluorescent dyes or fluorescent oligonucleotide probe chemistry (Chehab and Kan, 1989; Higuchi et al., 1992). The ability to monitor the progression of amplification reaction in real time eliminates the need for running time consuming and labour intensive downstream post-
PCR analysis allowing fast high-throughput assays (Heid et al., 1996; Higuchi et al., 1992). A typical quantitative positive amplification reaction plot is shown in (Figure 1.3). The linear fluorescence scale presents a sigmoidal shape positive amplification curve that consists of three phases:

**Linear-ground phase** - in this phase, there is no detectable amplification occurring over the background fluorescence and statistical noise. This phase is used to evaluate the baseline fluorescent "noise" (Karlen et al., 2007; Wong and Medrano, 2005).

**Log-linear phase** - this phase represents cyclical amplification when detectable fluorescence starts accumulating at an exponential rate with every PCR cycle. Figure 1.3B represents this phase on a log scale that imposes linearity on the increase in PCR product. Fluorescence data accrued during this phase is utilized for quantification analysis (Karlen et al., 2007; Rutledge and Cote, 2003; Wong and Medrano, 2005).

**Plateau phase** - in this phase the increase in DNA concentration decelerates to zero (plateau) because of various limiting factors coming into play, including depletion of one or more PCR reactants and accumulation of PCR product (Wong and Medrano, 2005).

**Figure 1.3**

Real-time PCR amplification curves in linear and logarithmic scale: (A) represents the sigmoidal curve with the three phases of PCR reaction in a positive amplification plot on a linear scale. (B) represents the amplification curves of serial dilutions of a sample on a log fluorescence scale and a
threshold is determined in the log-linear phase for quantification of samples. Figure adapted from (Karlen et al., 2007).

1.5.3 Mathematics of absolute quantification

Equation 1.1 (1A) describes the basic numerical principle of the PCR reaction where \( N_c \) is the number of amplicon molecules, \( N_0 \) is the number of initial target molecule, \( E \) is the fractional amplification efficiency and \( C \) is the number of thermocycles. Reordering of this equation provides the mathematical relationship Equation 1.1(1C) upon which qPCR is based (Rutledge and Cote, 2003).

\[
\text{Equation 1.1: Mathematical principle of PCR}
\]

\[
N_c = N_0 \cdot (E+1)^C \quad \text{equation 1A}
\]
\[
E = \frac{N_{c+1}}{N_c} \quad \text{equation 1B}
\]
\[
N_0 = \frac{N_c}{(E+1)^C} \quad \text{equation 1C}
\]

The increase in number of amplicon molecules produced after each thermocycle is proportional to the amplification efficiency and dependant on the number of starting molecules present. Therefore to deduce the original amount of molecules in an unknown sample, rearrangement of equation 1A produces equation 1C.

Using DNA fluorescence techniques, Higuchi et al. simplified and standardised determination of the \( N_c \) of all samples in a reaction by comparing all individual amplification reactions at the same point known as the Cycle-threshold (Ct) (Higuchi et al., 1992). The number of cycles needed for the amplification-associated fluorescence to reach this threshold level of detection (Ct) is inversely correlated with the amount of nucleic acid that was in the original sample. The \( C_t \) is chosen in the log-linear phase as the efficiency reaches a maximal value, which is nearly constant, during this phase (Karlen et al., 2007). By the threshold method, \( N_c \) becomes a constant and Equation 1.1C becomes modified to Equation 1.2 (below), where, \( N_t \) is the number of amplicons at that threshold.

\[
\text{Equation 1.2: Quantification of initial target molecule by Ct method}
\]

\[
N_0 = \frac{N_t}{(E+1)^{C_t}}
\]
After the determination of the Ct value, there are two different quantification strategies: i. absolute quantification or ii. relative quantification/expression. In absolute qPCR, serial dilutions of a known amount of target gene or cells are used as PCR standards. PCR standards can be prepared from pure culture genomic DNA or purified conventional PCR product or target gene inserted into plasmids for DNA quantification studies (Wong and Medrano, 2005). The standard curve is obtained by amplification of the PCR standards and plotting the resultant Ct values against log of the standard DNA dilutions (Figure 1.4) (Rutledge and Cote, 2003). Based on Equation 1.2 the logarithmic standard curve equation becomes Equation 1.3A and rearranging this equation to Equation 1.3C takes the form of general equation for a straight line \((y = mx+c)\), where constants \(m\) (slope) = - \(\log (E+1)\) and \(c\) (y-intercept) = \(\log(N_i)\) (Rutledge and Cote, 2003). Therefore the amplification efficiency of absolute quantification is the mean efficiency derived by \(E = 10^{-Slope} - 1\). Quantification of microbial load in unknown samples is achieved by extrapolating from this standard curve.

**Equation 1.3: Standard curve equation for absolute quantification**

\[
\begin{align*}
\log(N_0) &= \log (N_i) - \log [(E+1)^d] \quad \text{equation 3A} \\
\log (N_0) &= \log(N_i) - \log(E + 1).Ct \quad \text{equation 3B} \\
\log (N_0) &= - \log(E+1).Ct + \log (N_i) \quad \text{equation 3C}
\end{align*}
\]

Relative quantification studies the expression ratio of a target gene compared to a reference gene; it demonstrates the changes of messenger RNA (mRNA) levels of a gene across multiple samples, relative to levels of another gene transcript (Livak and Schmittgen, 2001; Pfaffl, 2006).
Figure 1.4

Amplification signals and linear regression graph from DNA standards: (A) represents the amplification in the logarithmic fluorescence scale signals obtained from DNA standards ranging from $10^7$ to $10^2$ genome/ml. The red horizontal line represents the fluorescence threshold selected for determining the Ct value for each of the concentration with $10^7$ (red curve) detected with the lowest Ct value and $10^2$ (blue curve) with highest Ct. (B) shows the plot of Ct values against the log concentration of the standards.

1.5.4 Real-Time PCR fluorescence chemistry

To monitor DNA amplification at each cycle of qPCR, various types of specific and non-specific fluorescence chemistry are available from DNA binding dyes (SYBR Green, SYTO9); hydrolysis probes (TaqMan), hybridisation probes (FRET probes) and hairpin probes (molecular beacons, scorpion primers)(Mackay, 2004; Wong and Medrano, 2005). Hybridisation and hairpin probes are very specific and usually used in allele discrimination studies (Hnatyszyn et al., 2001; Piatek et al., 1998; Solinas et al., 2001). Two of the most commonly used qPCR assay chemistries for the detection and enumeration of microbial load are SYBR Green qPCR and TaqMan qPCR (Cao and Shockey, 2012; Maeda et al., 2003; Tajadini et al., 2014).
**SYBR Green qPCR chemistry**

SYBR Green is an asymmetrical cyanine dye and on binding to the minor groove of the DNA double helix emits fluorescence at a ~1,000 fold higher level than in the unbound state (Dragan et al., 2012; Zipper et al., 2004). The resulting DNA-dye-complex absorbs blue light ($\lambda_{\text{max}} = 497 \text{ nm}$) and emits green light ($\lambda_{\text{max}} = 520 \text{ nm}$). During the extension step of the PCR cycle, multiple SYBR green molecules bind per molecule of dsDNA and quantification is achieved by acquisition of fluorescence emitted during formation of dsDNA amplicons. As the double-stranded PCR product accumulates during cycling, the fluorescence intensity increases proportionally to the dsDNA concentration (Wittwer et al., 1997).

Figure 1.5A shows the SYBR green qPCR assay chemistry. Binding of SYBR green to dsDNA is non-specific as it does not distinguish between amplicon from the gene of interest and dsDNA formed from primer-dimers or non-specific amplification (Ririe et al., 1997). Therefore post amplification melt curve analysis is performed that differentiates the DNA product of interest from non-specific co-products based on the melt temperature profiles (Ririe et al., 1997). The advantage of this sequence-independent detection of DNA with SYBR Green is that it allows quantification of any target gene with the use of gene-specific primers. This makes SYBR green assays rapid, and flexible and ideal for preliminary studies, without the need for extra investment designing individual probes (Simpson et al., 2000).

**TaqMan qPCR chemistry**

These assays use dual labelled probes, which are short oligos with a fluorescent reporting dye attached to the 5’-end and a quencher molecule attached to the 3’-end (Heid et al., 1996). Little fluorescence is emitted from the native probe due to close proximity of the reporter dye and quencher. When the probe anneals to the target sequence, the bound and quenched probe are degraded by the DNA polymerase’s 5’ nuclease ability during the annealing and extension step of the PCR. Probe degradation allows for separation of the reporter from the quencher dye, resulting in increased fluorescence emission: Figure 1.5B (Gibson et al., 1996; Heid et al., 1996).
Probe based assays are more specific than SYBR green assays, but as the probe gets denatured during the amplification cycle, melt curve analysis cannot be performed. A comparison of the assay chemistries is shown in Table 1.3

**Figure 1.5**

**SYBR Green and TaqMan fluorescence chemistry:** (A) shows the unbound SYBR green in presence of single stranded DNA. During polymerization and extension of ds DNA multiple SYBR green molecules binds to ds DNA and starts fluorescing. (B) shows a dual labelled probe with a fluorophore at the 5’ end of the oligo sequence whose fluorescence is quenched by the presence of a quencher at the 3’ end. During polymerization Taq polymerase 5’ exonuclease activity cleaves the fluorophore and it fluoresces separated from its quencher. Adapted from [http://www.thermoscientificbio.com/applications/pcr-and-qpcr/](http://www.thermoscientificbio.com/applications/pcr-and-qpcr/); basic principles of qPCR.
### Table 1.3: Comparison of SYBR Green and TaqMan qPCR chemistry

<table>
<thead>
<tr>
<th>SYBR Green qPCR chemistry</th>
<th>TaqMan qPCR chemistry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Each amplicon molecule is represented by multiple SYBR green</td>
<td>Each amplicon molecule is represented by a single fluorescent molecule when</td>
</tr>
<tr>
<td>molecules and biases introduced in presence of variable length</td>
<td>the probe anneals to the template DNA</td>
</tr>
<tr>
<td>amplicons</td>
<td></td>
</tr>
<tr>
<td>Non-specific binding to any dsDNA</td>
<td>Fluorescence acquired only on binding of probe to target specific region</td>
</tr>
<tr>
<td>Cheaper and multiple assays can be performed by changing the</td>
<td>Relatively expensive and time consuming as each target region would require</td>
</tr>
<tr>
<td>target region and target primers</td>
<td>designing of a new probe sequence</td>
</tr>
<tr>
<td>Reversible nature of this assay allows performing melt curve</td>
<td>Irreversible nature and therefore melt curve analysis cannot be performed</td>
</tr>
<tr>
<td>analysis</td>
<td></td>
</tr>
<tr>
<td>Cannot perform multiplex assay</td>
<td>Can be used to design multiplex assay</td>
</tr>
</tbody>
</table>

#### 1.5.5 Real-Time PCR applications

Real-Time PCR has been applied for the detection and quantification of various different pathogens, including viruses, bacteria, fungi and protozoa, in a range of clinical and environmental samples (Mackay, 2004). Due to its high sensitivity, it has been a valuable method for exploring changes in microbial load as an indicator of active infection, disease progression and efficacy of antimicrobial treatment (Garcha et al., 2012; Gentili et al., 2012; Zhang et al., 2004) that is helping to better understand the pathogenesis and association of microbes with various chronic conditions (Capone et al., 2000; Kennedy et al., 1997; Smith-Vaughan et al., 2006). The speed and sensitivity of real-time PCR has also revolutionised the diagnosis of infections caused by slow growing and fastidious organisms (Mackay, 2004; Maurin, 2012; Miller et al., 2002). This rapid diagnosis turnaround time is particularly advantageous in the setting of severe or fulminant infection such as meningitis, sepsis and food poisoning (Fortin et al., 2001; Ke et al., 2000). In clinical practice, real-time PCR has become a favoured technique for rapidly identifying antibiotic resistant strains of various organisms such as *S. aureus*, *Mycobacterium tuberculosis* and *Helicobacter pylori*, enabling the use of appropriate treatment more quickly (Chisholm et al., 2001; Garcia de Viedma et al., 2002; Tan et al., 2001). qPCR has also been developed for characterization of bacterial...
communities, pathotyping and genotyping (Asai et al., 2002; Cheah et al., 2010; Klaschik et al., 2002; Woo et al., 1997)

Apart from real-time PCR, other molecular techniques such as fluorescence in situ hybridisation (FISH), nucleic acid sequence based amplification (NASBA), quantitative southern/dot blot hybridisation, microarray and flow cytometry have also been utilised for quantification. Table 1.4 below lists the methodology and limitations of these techniques compared with qPCR.
### Table 1.4: Other molecular techniques for microbial quantification

<table>
<thead>
<tr>
<th>Other quantification techniques</th>
<th>Methodology</th>
<th>Advantage/Application</th>
<th>Limitation relative to qPCR</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluorescent in-situ hybridisation (FISH)</td>
<td>Whole cell hybridisation technique where fluorescent labelled probes are hybridised to RNA/DNA of cells. Quantification by microscopy based cell counting or automated fluorescence intensity based quantification.</td>
<td>Considered as GOLD standard for quantification</td>
<td>Requires intact cells and fresh sample. Time consuming, very low throughput, prone to human error</td>
<td>(Coskuner et al., 2005; Manz et al., 1994)</td>
</tr>
<tr>
<td>Nucleic acid sequence based amplification (NASBA)</td>
<td>An isothermal, transcription based amplification assay. Amplification takes place at a temperature lower than required for DNA denaturation and preventing their amplification</td>
<td>Eliminates contamination arising from DNA amplification. Commonly used in diagnostic tests</td>
<td>Cannot use for DNA based assays. Greater variability between samples than qPCR assay might occur due to different PCR kinetics of three enzymes used for NASBA.</td>
<td>(Compton, 1991; Patterson et al., 2005)</td>
</tr>
<tr>
<td>Dot blot hybridisation</td>
<td>Extracted rRNA bound to nylon membrane is hybridised to radiolabelled probes. Amount of probe hybridised is quantified using densitometry relative to reference standards.</td>
<td>Used commonly prior to fluorescent labelled probe and real-timePCR technique</td>
<td>Use of radioactive substances. Time consuming</td>
<td>(Raskin et al., 1994; Stahl et al., 1988)</td>
</tr>
<tr>
<td>Microarray</td>
<td>Hybridisation of fluorophore labelled target genes to chip bound thousands of oligo probes. Quantification based on fluorescence intensity</td>
<td>High-throughput.</td>
<td>Time consuming, labour intensive, lot of optimisation, expensive</td>
<td>(Small et al., 2001; Treimo et al., 2006)</td>
</tr>
<tr>
<td>Flow cytometry</td>
<td>Flow cytometry is a sensitive method and can be used to quantify either fluorescent labelled viable cells (Flow-Fish) or quantification of fluorescent labelled amplified nucleic acid.</td>
<td>sensitive technique</td>
<td>Time consuming, labour intensive, expensive</td>
<td>(Dreier et al., 2009; Wedemeyer et al., 2000)</td>
</tr>
</tbody>
</table>
1.6 Bacterial community characterization

Although important, quantitative and PCR techniques only provide information on the presence of known target specific bacterial populations and fluctuations in their load. However, profiling of directly amplified 16S rRNA gene of various biotic and abiotic samples has taken forward our understanding of the complex and diverse array of existent microbes. A good example of this is the extensive study of the human gut microbiome with these techniques. E. coli was previously thought to be the classic gut microbe. However it is now recognised that the entire Gamma-proteobacteria class, that includes E. coli, usually forms < 1% of the gut biome (Eckburg et al., 2005; Hamady and Knight, 2009). In contrast, anaerobic bacterial species from Firmicutes and Bacteroidetes phyla dominate the gut microbiome, with the majority being novel and yet to be cultivated (Eckburg et al., 2005; Hold et al., 2002).

1.6.1 Microbial complexity of normal human microbiome

In health, resident microbial cells outnumber human cells by a factor of 10 and together comprise the healthy human microbiome (Foxman et al., 2008). In addition to the gastrointestinal tract (GIT), bacteria are abundant in skin, the respiratory tract and vagina (Dethlefsen et al., 2007). These consortia of microorganisms and their composite interactions are essential for human health.

Although more than 50 phyla are known, only four (Proteobacteria, Firmicutes, Actinobacteria and Bacteroidetes) dominate the human microbiota at various sites. At phylum level the bacterial composition of these body sites tends to be consistent, however there is wide diversity at species level across sites and between individuals (Costello et al., 2009; Dethlefsen et al., 2007). Factors such as genetic background, diet, ethnicity, race, geographical location and age are considered some of the reasons for the observed interpersonal differences (Ursell et al., 2012a; Yatsunenko et al., 2012; Zoetendal et al., 2011). Both genetics and diet are thought to influence the ratio between Bacteroidetes and Firmicutes in the gut microbiome and this has been associated with differentiating obese and lean groups (Ley et al., 2006; Turnbaugh et al., 2009). Differences in the bacterial composition of the surface of hands has been reported between U.S and Tanzanian subjects with dominance of Propionibacteriaceae, Staphylococcaceae and Streptococcaceae families in the former population compared to abundance of soil-associated Rhodobacteraceae and Nocardioidaceae in the latter
(Hospodsky et al., 2014). A study on vaginal microbiota of asymptomatic women from four different ethnic groups showed significantly differences in microbial composition between ethnic groups, with abundance of *Lactobacillus* genera in Asians and Whites and dominance of *Prevotella, Sneathia, Megasphaera*, or *Streptococcus* in Hispanic and Black populations (Ravel et al., 2011).

The core microbiome is difficult to define because of both inter and intraindividual variability of the microbiome at and across different body sites (Costello et al., 2009; Ursell et al., 2012a; Ursell et al., 2012b). Costello and co-workers showed that variability in bacterial composition of skin from different sites was greater than inter or intrapersonal variation within sites (Costello et al., 2009). Erb-Downward and colleagues demonstrated significant variation in bacterial communities identified from tissue samples collected from distinct lung sites within subjects (Erb-Downward et al., 2011). They also showed significant inter-individual differences in healthy respiratory microbiome. In contrast, Charlson et al. reported a homogenous microbiota from URT to lungs with dominance of *Streptococcaceae, Prevotellaceae* and *Veillonellaceae* across individuals (Charlson et al., 2011). Variation in microbial composition based on sample type has also been demonstrated for the gut and airway microbiomes (Cabrera-Rubio et al., 2012; Eckburg et al., 2005; Hilty et al., 2010).

To help overcome the problem of heterogeneity, the "National Institute of Health (NIH) Common Fund" launched the Human Microbiome Project (HMP) as a community resource program [http://commonfund.nih.gov/hmp/](http://commonfund.nih.gov/hmp/) with the objective of creating an overview of the healthy human microbiome at five major body sites (airways, skin, oral cavity, GIT, and vagina) (Gevers et al., 2012).

### 1.6.1.1 Host-Microbe interactions

Healthy humans are inherently equipped with a sophisticated innate and adaptive immune system for effectively clearing and preventing accumulation of foreign particles. The tight epithelial lining at body sites exposed to the external environment forms the first line of defence by containing microbes at these sites. Mucus secretion prevents the binding of microbes to epithelial cell layers while the synchronised beating of cilia in the airways leads to expulsion of foreign antigens (Abusriwil and Stockley, 2007; Wilson et al., 1996). Several antibacterial factors such as bacteriostatic proteins (lysozyme, lactoferrin), bacteriocidal factors (β defensins), local immunity (secretory IgA) and resident phagocytes (macrophages) ensure effective clearance of antigens.
Introduction

(Abusriwil and Stockley, 2007). But despite these mechanisms commensal bacteria are able to colonise and proliferate on the skin and mucosal surfaces of the GIT, RT and vagina. Establishment of this human microbiome has been possible by a mutual relationship being established with co-evolution of the host and commensals. From an immunological perspective, host mucosal surfaces have developed specialised regulatory, anti-inflammatory immunity to tolerate and allow colonisation of commensal bacteria (Tlaskalova-Hogenova et al., 2004). These consist of specialised pattern recognition receptors (toll-like receptors TLR) that have evolved to differentiate between commensal and pathogens (Beisswenger et al., 2009; Kelly and Conway, 2005). It has been demonstrated that commensal bacteria play an active role in modulating host innate immune response, priming the host adaptive immune response against pathogens and stimulating epithelial repair (Hooper, 2004). Commensals release proteins with immunosuppressive properties such as the Arg-1 gingipain proteinase of Porphyromonas gingivalis that cleaves and inactivates pro-inflammatory cytokines IL-1 and IL-6 (Tlaskalova-Hogenova et al., 2004). Another example of this mutualistic behaviour is induction of release of host Angiogenin 4 (Ang4) protein by Bacteroides thetaiotaomicron. This protein selectively eliminates Gram positive bacteria in the gut, (Ismail and Hooper, 2005). Metabolites produced by gut anaerobic and lactic acid producing commensals have been shown to suppress pro-inflammatory responses induced by pathogenic bacteria (Menard et al., 2004; Vinolo et al., 2011).

While commensals assist in maintaining mucosal immunity and homeostasis, pathogenic bacteria have, in contrast, developed several mechanisms to evade host defense mechanisms and cause host tissue injury (Wilson et al., 1996). Respiratory pathogens (H. influenzae, S. pneumonias and P. aeruginosa) have several virulence factors that cause epithelial damage, disruption of ciliary action, bacterial adherence and increased mucus secretion (Abusriwil and Stockley, 2007). Various bacterial products such as lipopolysaccharide (LPS), proteases, capsular polysaccharides, toxins and enzymes released by these pathogens directly cause tissue injury and inflammation. A pro-inflammatory host immune response is triggered by disrupting the protease-antiprotease balance, stimulating generation of pro-inflammatory mediators and neutrophil chemo attractants, exacerbating host inflammation caused by indirect tissue damage (Abusriwil and Stockley, 2007; Beisswenger et al., 2009).
1.6.1.2 Interactions between bacteria (microbe and microbe)

Bacteria can co-exist through positive interactions between them. Mutualistic or synergistic relationships between gut commensals exist for production of metabolites and vitamins that in turn benefit the host. Synergistic relationships also exist between pathogens for establishing colonisation leading to chronic, polymicrobial infection. Examples include dental plaque and otitis media and Cystic Fibrosis. The colonic degradation of butyrate producing polysaccharides is achieved through a cooperative microbial process. *Bifidobacterium* species produces lactate as a by-product, from fermentation of the polysaccharides, which is then converted to butyrate by lactate fermenters, such as *Eubacterium hallii* and utilized by the host (Dethlefsen et al., 2007; Flint et al., 2007). In addition, some of the resulting sugars are consumed by strains that cannot ferment the polymers but provide growth factors to strains that do (Dethlefsen et al., 2007). Synergistic induction of amplified epithelial pro-inflammatory response by *S. pneumoniae* and *H. influenzae* have been reported using both *in vitro* and *in vivo* models (Ratner et al., 2005). These bacteria are common pathogens in various nasopharyngeal infections, otitis media and COPD and may contribute to accelerated inflammation and tissue damage of heavily colonised mucosal barriers. *In vitro* studies have demonstrated synergistic associations between *Fusobacterium nucleatum* and *P. gingivalis*, commonly detected pathogen in periodontal and cardiovascular diseases. While *P. gingivalis* stimulates biofilm formation by *F. nucleatum*, presence of the latter enhances invasion by *P. gingivalis* of gingival epithelial and aortic endothelial cells (Saito et al., 2008a; Saito et al., 2008b).

Associations between commensals and pathogenic bacteria are frequently negative, with competition for resources. These negative associations provide protection to the host from infection with competing pathogenic bacteria. Another mechanism for an antagonistic association is the release of metabolic by-products, reactive oxygen species and antimicrobial substances like bacteriocin that have an inhibitory effect on other bacteria (Fons, 2000). *Clostridium difficile* is a hospital acquired infection of the gut that is caused primarily by antibiotic therapy. Recently a fecal isolate of a *Bacillus thuringiensis* strain was shown to produce bactreicin, thuricin CD that was effective in clearing infection with *C. difficile* (Rea et al., 2010). In the URT, the normal flora that includes α-haemolytic and non-haemolytic *Streptococcus*, *Prevotella* and *Peptostreptococcus* species have also been shown to produce inhibitory products that
may have potential for therapeutic utilisation to treat various URT infections (Brook, 2005). Lemon and co-workers analysed the nostril microbiota of healthy individuals and showed a negative correlation between prevalence of Staphylococcaceae family and the Corynebacteriaceae and Propionibacteriaceae families, suggesting possible antagonism between these groups. The Staphylococcaceae include important community-acquired and nosocomial pathogens such as S. aureus and coagulase-negative Staphylococcus species. These observed inverse correlations point to the possibility that nostril bacterial communities in which Actinobacteria are most prevalent might protect against carriage of Staphylococcus species including the pathogen S. aureus (Lemon et al., 2010).

1.6.2 Molecular approaches for 16S rDNA bacterial profiling

As discussed in section 1.4.1, development of PCR, cloning and sequencing techniques has allowed detection and characterization of novel and uncultivated bacteria from the microbiome of environmental and human samples (Eckburg et al., 2005; Giovannoni et al., 1990). Sanger sequencing based on the di-deoxy chain termination method became the most popular method for sequencing and characterization of bacterial communities using 16S rDNA clone libraries (Tewari et al., 2011; Zimmermann et al., 1988). Automation of Sanger sequencing, utilizing fluorescent dyes and capillary gel electrophoresis made this process faster and more accurate (Liu et al., 2012). However building clone libraries to separate individual 16S rDNA genes from the amplified mixture proved a major bottleneck, making the process of bacterial characterization extremely laborious and time-consuming with low throughput.

A number of other molecular approaches have been utilized for bacterial community analysis. These include DNA fingerprinting techniques like Temperature/Denaturing Gradient Gel Electrophoresis (T/DGGE), Single-strand conformation polymorphism (SSCP), Terminal Restriction Fragment Length Polymorphism (T-RFLP) Amplified rDNA restriction analysis (ARDRA); length heterogeneity PCR (LH-PCR) and hybridization techniques such as 16S rRNA gene microarray (PhyloChip) (Rastogi G. and Sani R. K., 2011). Among these, DGGE and T-RFLP are the most extensively used (Siqueira et al., 2010). These provide a qualitative picture of bacterial diversity and dynamics of a community based on the generated sequence or length polymorphism profiles. With T/DGGE, the amplified marker gene is electrophoresed on a polyacrylamide gel containing a linear gradient of temperature or DNA denaturant such as a mixture of urea and formamide. This allows separation of the amplicon mixture
forming a gel pattern that is based on the variation in sequence composition of the mixture (Muyzer et al., 1993). In T-RFLP the amplified marker gene tagged with a fluorescently labelled molecule at the 5' end during amplification is digested with restriction enzyme to produce terminal-restriction fragments (T-RF). An automated sequencer generates a profile of peaks with each peak corresponding to a specific T-RF size, representing a single ribotype. The size of the peak provides relative abundance estimates for different bacterial groups (Marsh, 1999).

These methods provide a rapid approach for crude profiling to determine similarities and difference between bacterial communities. However they have low resolution for providing taxonomic identities and estimating relative abundance of community constituents. Other common limitations are generation of the same profile with different DNA fragments and/or multiple profiles of a bacterium arising from sequence heterogeneity among multiple rRNA operons. These technical confounders are a cause of bias in diversity measurements (Rastogi G. and Sani R. K., 2011). DeSantis et al. designed a universal high-density 16S microarray containing ~300,000 probes of the 16S rRNA gene, targeting ~8,935 operational taxonomic units (OTUs) (DeSantis et al., 2007). Although microarrays have a higher throughput than cloning-sequencing and finger printing methods, they are technically and analytically challenging and cannot identify novel bacteria (DeSantis et al., 2007).

1.6.3 Next generation sequencing (NGS) technologies

In less than a decade after their advent, next/second generation sequencing technologies have replaced the conventionally used dye terminator capillary gel electrophoresis based sequencing methods for complex microbial community studies (Liu et al., 2012). The enormous parallel high-throughput sequencing capabilities of NGS platforms allow higher depth of coverage and detection of low abundance taxonomic groups/alleles at a relatively low cost compared to Sanger sequencing, making them the preferred method for 16S based community analysis, variant analysis and whole genome sequencing (WGS) (Liu et al., 2012).

Some of the leading platforms for high-throughput sequencing utilised in metagenomic studies include 454 (Roche), Solexa (Illumina), SOLiD (Applied Biosystems) and Ion Torrent PGM (Life Technologies). One of the major limitations of high throughput sequencing is read length. Reads from NGS technologies are considerably shorter than
Introduction

those from Sanger sequencing. Among the NGS platforms, 454 sequencing with its shorter time per run (10 hours) and longer read lengths (~500 bases) compared to others has been widely used for bacterial community analysis (Liu et al., 2012). However, more recently MiSeq (Illumina platform) has delivered improvements in the read-length, producing lengths of up to 400bp at a lower cost and 10 fold higher coverage depth, compared with 454. This has led to its increasing application in microbiomic studies (Caporaso et al., 2012; Loman et al., 2012). A comparison of common NGS technologies and Sanger sequencing for microbiome analysis is shown in Table 1.5.

<table>
<thead>
<tr>
<th>Method (sequencing platform)</th>
<th>454 (454 GS FLX)</th>
<th>Solexa/Illumina (HiSeq, MiSeq)</th>
<th>SOLID (SOLiDv4)</th>
<th>Ion torrent (Ion personal genome sequencer (PGM))</th>
<th>Sanger sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clone library</td>
<td>Emulsion PCR(EmPCR) Clonal amplification of individual DNA template by oil emulsion PCR creating a clone library of one read/bead</td>
<td>Solid phase bridge amplification on sequencing slides to generate spatially separated clonal clusters</td>
<td>EmPCR</td>
<td>EmPCR</td>
<td>Bacterial clone library</td>
</tr>
<tr>
<td>Sequencing methodology</td>
<td>Pyrosequencing: Intensity of light emitted, due to enzymatic reaction on released inorganic pyrophosphate when nucleotide incorporated into DNA, is measured</td>
<td>Sequencing by synthesis: Simultaneous flow of all four terminator dNTPS, each labelled with a unique fluorescent dye. Sequence imaging by detection of fluorescence on nucleotide incorporation and chain elongation by cyclic irreversible chain termination.</td>
<td>Sequencing by ligation: Ligase enzyme catalyzed extension of the DNA fragment and fluorescence detection of the incorporated nucleotide dye and cleavage for next cycle of ligation reaction</td>
<td>Semiconductor technology: Change in voltage is measured which is proportional to the change in pH caused due to release of proton during incorporation of nucleotide in the DNA molecule</td>
<td>Chain termination: Electrophoretic separation of DNA fragments produced due to termination of extension by deoxy nucleotide incorporation</td>
</tr>
<tr>
<td>Read length</td>
<td>700 bp</td>
<td>50 to 300 bp</td>
<td>50±35 or 50±50 bp</td>
<td>200bp</td>
<td>up to 900 bp</td>
</tr>
<tr>
<td>Accuracy</td>
<td>99.9%</td>
<td>99%</td>
<td>99.9%</td>
<td>99%</td>
<td>99.9%</td>
</tr>
<tr>
<td>Reads per run</td>
<td>1 million</td>
<td>up to 3 billion</td>
<td>1.2 to 1.4 billion</td>
<td>up to 5 million</td>
<td>96/384</td>
</tr>
<tr>
<td>Time per run</td>
<td>24 hours</td>
<td>1 to 10 days, depending upon sequencer and specified read length</td>
<td>1 to 2 weeks</td>
<td>2 hours</td>
<td>20 minutes to 3 hours</td>
</tr>
<tr>
<td>Cost per 1 million bases (in US$)</td>
<td>$10</td>
<td>$0.05 to $0.15</td>
<td>$0.13</td>
<td>$1</td>
<td>$2400</td>
</tr>
<tr>
<td>Disadvantages</td>
<td>Runs are expensive. Homopolymer errors.</td>
<td>Equipment can be very expensive. Requires high concentrations of DNA.</td>
<td>Slower than other methods. Have issue sequencing palindromic sequence.</td>
<td>Homopolymer errors.</td>
<td>More expensive and impractical for larger sequencing projects.</td>
</tr>
</tbody>
</table>

Data presented in the table is based on following publications (Liu et al., 2012; Metzker, 2010; Quail et al., 2012).
1.6.3.1 454 sequencing

454 sequencing technology was the first to be developed among the NGS technologies by 454 Sciences that later became part of Roche (Liu et al., 2012; Margulies et al., 2005). Within three years of the first paper using this method, around 250 peer-reviewed papers were published with studies involving WGS, metagenomics, transcriptome profiling, ultra-deep sequencing for rare variant analysis and the study of ancient DNA (Roche, 2008; Rothberg and Leamon, 2008). Currently there are more than 1000 publications utilising the 454 Genome Sequencer (GS) FLX system for studies related to metagenomics and microbial diversity (http://my454.com/publications/publications.asp?postback=true website).

The principle of 454 sequencing is based on pyrosequencing, which is a non-electrophoretic, bioluminescence method that measures the release of inorganic pyrophosphate (PPi) by proportionally converting it to visible light. This process is achieved by a series of enzymatic reactions. The DNA fragment to be sequenced is incubated with a four enzyme mixture of DNA polymerase, ATP sulfurylase, luciferase and apyrase, an unmatched deoxyribonucleoside triphosphate (dNTP) degrading enzyme. On binding of a dNTP to the extending DNA fragment, PPi is released that is converted to adenosine triphosphate (ATP) by sulfurylase. Following this, a luciferase catalysed reaction converts the ATP to light which is measured by a light-sensitive device, like the charge-coupled device (CCD) camera (Ronaghi et al., 1998).

NGS technologies involve a two-step method for sequencing. The first step is the parallel bulk clonal amplification of individual templates which has overcome the time consuming production of bacterial clone libraries needed with Sanger sequencing (Holt and Jones, 2008). Clonal amplification with 454 sequencing is an oil emulsion based PCR reaction mixture (emPCR) (Figure 1.6A). EmPCR is performed either on the sheared DNA fragment library, with adapter sequences ligated at both 5' and 3' ends, or on amplicon libraries, with the adaptor sequence incorporated during the amplification cycle. Building the 16srDNA amplicon library for emPCR and primer design is described in (4.2.2.1). The adaptor sequences are 454 proprietary oligonucleotide sequences that allow binding of the DNA fragment to the beads used for emPCR and sequencing. Limiting conditions are used to facilitate binding of a single DNA template molecule per bead. Each bead is compartmentalized in a private droplet of aqueous PCR reaction mixture within an oil emulsion. This allows amplification of individual
molecules without contamination or competing sequences, producing millions of clonally amplified sequencing templates on each bead (Margulies et al., 2005; Metzker, 2010). In the second phase, parallel sequencing of these individual beads is performed by the 454 sequencing machine. This consists of four parts: a fluidic assembly, a flow cell along with the well-containing fibre-optic slide, a CCD camera-based imaging assembly, and a computer. The fibre optic slide, containing ~1.6 million wells, is loaded with beads and mounted in a flow cell chamber through which the sequencing reagents flow (Figure 1.6B) (Margulies et al., 2005). The fluidic subsystem delivers the sequencing buffer and the 4 nucleotides sequentially in a fixed order (GS FLX titanium series-200 cycles) across the well during the sequencing reaction. During the flow of nucleotides, complementary nucleotides to the template strand in each well are incorporated in the new DNA strand that is being synthesized and the camera captures the light generated by the chemiluminescence reaction. This information is translated into DNA bases on the computer. The entire flowgram for 454 sequencing is shown in Figure 1.6.

454 sequencing has been widely applied for understanding structure, function and dynamics of the microbial community in health and disease. Some examples of pyrosequencing based human microbiome studies include characterization of the gut microbiome with increased energy yield capacity and its association with obesity (Turnbaugh et al., 2006), impact of antibiotic treatment on the gut microbiome (Dethlefsen et al., 2008) variation in the vaginal microbial community between healthy individuals (Ravel et al., 2011), variation within individuals of the skin microbiome across body sites (Costello et al., 2009), comparison of bacterial community profiles in health and periodontitis (Griffen et al., 2012) and characterisation of the airway microbiome in health and disease (Charlson et al., 2011; Pragman et al., 2012).
Flowgram of 454 sequencing: (A) shows the emPCR where a single DNA molecule with adaptors attached to the ends (yellow rod with blue and red end) is present per bead along with PCR reaction mixture within an oil emulsion forming individual micoreactors allowing parallel clonal amplification of the entire DNA library. Each bead contains millions of copies of a unique DNA template after amplification. The emulsion is then broken and the DNA is denatured (B) B1 Each bead carrying ssDNA clones deposited onto the fibreoptic slide. B2 smaller beads carrying immobilized enzymes required for pyrophosphate sequencing are deposited into each well. B3 shows the 454 sequencer with the fluid assembly system (a), flow cell chamber containing the sequencing slide (b), the CCD camera for imaging (c) and the computer system. B4 shows the flowgram generated from the light signal produced proportional to the incorporation of a nucleotide. Picture adapted from (Margulies et al., 2005) and (Metzker, 2010).
1.7 Aims and Objectives

Through culture based assays it is recognised that bacteria play a role in progression of COPD and are especially associated with exacerbation stages. However, limitations in the sensitivity and scope of plate techniques have not allowed the precise role of bacteria to be characterised for different COPD states (Wilson, 1998). Between exacerbations, the presence of bacteria in the airways is considered to represent colonisation with bacterial load constrained by the host immune defence to maintain an equilibrium state. Exacerbation episodes are hypothesized as events when this equilibrium is disrupted leading to an increase in the load of bacterial pathogens that further incites the host immune response (Wilson, 1998).

The overarching aim of this project was to characterise the role of bacteria in the airways using state of the art molecular techniques in a cohort of COPD subjects at different clinical stages (stable, exacerbation and post exacerbation) of disease. Within this, specific objectives were explored using two different molecular techniques applied to sputum samples from the cohort: i) Quantitative (q) PCR and ii) 454 high throughput sequencing. The different techniques were suited to examining different research questions and these are summarised below.

1.7.1 Specific objectives using qPCR

Quantitative PCR provides quantitative information about total and specific bacterial load present in a sample. The technique was therefore utilised to examine whether differences in detection rate or quantitative load in sputum of bacterial groups commonly associated with COPD pathogenesis were present during different clinical COPD states.

Quantitative PCR assays were developed and optimised to enumerate both the total bacterial load and the load of individual potentially pathogenic bacteria *H. influenzae, S. pneumoniae, S. aureus* and *M. catarrhalis*, based on previous culture based knowledge of pathogens associated with COPD.

Specific objectives were:

- To compare the qPCR method with conventional plate technique for bacterial detection and quantification.
Introduction

- To investigate if differences exist in detection rate of pathogens between stable and exacerbation.
- To examine if a significant change was measurable in the total bacterial load and individual specific bacteria between stable and exacerbation state in longitudinal samples.

1.7.2 454 Sequencing

Recent utilisation of high throughput sequencing to characterize the microbiome at various human body sites is informing the relationship between bacteria and states of health and disease. These techniques are very sensitive and provide a snapshot of the breadth of bacterial communities present within a given sample. Sequencing of samples from the lower airway tract in healthy subjects has revealed the presence of diverse microbial communities within a body compartment previously considered to be sterile (Charlson et al., 2011; Free, 2005).

454 sequencing was utilised in this project to provide an overview of the diversity and proportions of different microbial communities in sputum from the COPD cohort. The specific objectives studied were:

- To characterise the microbiome across a cohort of subjects with established COPD and determine whether key pathogens associated with COPD using plate culture techniques are major constituents of the COPD microbiome at different states.
- To classify COPD samples and COPD subjects based on their microbial flora.
- To investigate whether differences exist in the microbiome between stable and exacerbation states in longitudinal samples.
- To examine if differences in microbiome are associated with clinical phenotypes of COPD based on COPD GOLD stage or smoking status.
- To investigate if different therapy regimens for exacerbation episodes are associated with specific changes in the microbiome.
16S rDNA 454 sequencing was also performed on multiple stable visit samples on a group of subjects belonging to the above exacerbated microbiome subject cohort and single stable samples from subjects that did not exacerbate during the study period. Specific objectives in these groups were:

- To determine the stability of microbiome at stable state over longer time periods
- To examine if difference exists in microbial structure between exacerbated and non-exacerbated groups.

COPD is known as a complex and heterogeneous disease with differences between subjects recognised in disease severity (GOLD stage), clinical expression and rate of disease progression. Moreover, exacerbations are also multifactorial with varied aetiologies (infectious/non-infectious) type and severity (Anthonisen criteria). Various criteria for describing COPD phenotypes have been suggested for classification of patients into distinct prognostic and therapeutic subgroups for both clinical and research purposes (Han et al., 2010). Using inflammatory biomarker expression data, Bafadhel and colleagues at our centre have recently described COPD exacerbations phenotypes as bacterial-predominant, viral-predominant and eosinophil-predominant (Bafadhel et al., 2011). As the subjects and their exacerbations included for this project were derived from the population studied by Bafadhel, the opportunity therefore existed for this project to determine whether patterns of change in the microbiome are associated with one or more of these exacerbation phenotypes. It was hypothesized that bacteria related exacerbations would show a characteristic pattern of disturbance in the microbiome structure at exacerbation from its stable state that would be restored at recovery and be distinct from other forms of exacerbation. This hypothesis was explored as a specific objective of the project.
2  Materials and methods

2.1  Bacteria, Media and Buffers

2.1.1  Bacteria

<table>
<thead>
<tr>
<th>Bacterial species</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Escherichia coli K12</em></td>
<td>sequenced strain (widely studied enteric bacterium)</td>
<td>Laboratory stocks (NCTC 10538)¹</td>
</tr>
<tr>
<td><em>Streptococcus pneumoniae D39</em></td>
<td>sequenced strain (virulent encapsulated serotype 2 strain)</td>
<td>Laboratory stocks (NCTC 7466 )³</td>
</tr>
<tr>
<td><em>Haemophilus influenzae</em>³</td>
<td>clinical strain (respiratory patient isolate)</td>
<td>Leicester Royal Infirmary hospital (LRI)</td>
</tr>
<tr>
<td><em>Staphylococcus aureus Newman</em></td>
<td>sequenced strain</td>
<td>Laboratory stocks (NCTC 8178)³</td>
</tr>
<tr>
<td><em>Moraxella catarrhalis NCTC1120</em></td>
<td>sequenced strain</td>
<td>Health Protection Agency (HPA), UK</td>
</tr>
<tr>
<td><em>Pseudomonas aeruginosa PA14</em></td>
<td>sequenced strain (opportunistic pathogen)</td>
<td>Departmental stocks (NC_008463.1)²</td>
</tr>
<tr>
<td><em>Bacteroides fragilis</em>³</td>
<td>clinical strain (anaerobe isolated from patient stool sample)</td>
<td>Leicester Royal Infirmary (LRI)</td>
</tr>
</tbody>
</table>

¹ National collection of type culture number
² NCBI genome accession number
³ These bacteria were isolated from patient samples and only characterised till genus level.

2.1.2  Media

Luria-Bertani (LB) broth

10g Bacto-tryptone, 5g Bacto-yeast extract and 5g NaCl were dissolved in 1 litre of distilled water. The pH was adjusted to 7.0 with 1M NaOH and the solution was sterilised by autoclaving.

LB Agar

15g agar was dissolved in 1 litre LB broth before autoclaving. Autoclaved molten media was poured in petri dishes to make LB plates.

Brain Heart Infusion (BHI) Broth

37g BHI broth powder was dissolved in 1 litre of distilled water and sterilised by autoclaving.

Blood Agar
Materials and methods

40g blood agar base was dissolved in 950ml distilled water. The solution was sterilised by autoclaving. 5% v/v blood was added to the autoclaved blood agar base solution held at 40°C. The media was swirled to mix, and poured into petri dishes. Blood agar plates were kept at room temperature for 1 week or at 4°C for 4 weeks.

Chocolate Agar

5% v/v blood was added to the autoclaved blood agar base held at ~70°C to lyse the red blood cells within the medium. This media was poured into petri dishes to make chocolate agar plates, which were kept at room temperature for 1 week or at 4°C for 4 weeks.

Haemin (1mg/ml)

4ml triethanolamine and 96ml distilled water was placed in a falcon tube with 100mg haemin. This mixture was dissolved and stabilised by heating at 65°C for 30 minutes. The haemin stock was stored at 4°C.

2.1.3 Reagents and Buffers

50X TAE (Tris-acetate-EDTA)

1 litre was prepared by adding 242g Tris base, 57.1ml glacial acetic acid and 37.2g Na₂EDTA.2H₂O to 900ml distilled water, adjusting the pH to 8.5 with a few drops of 1M HCl, adding water to 1 litre and autoclaving. The TAE stock was diluted to 1X using distilled water.

EDTA (ethylenediamine tetraacetic acid), 0.5M (pH 8.0)

186.1g of Na₂EDTA powder was dissolved in 700ml of water. pH was adjusted to 8.00 with ~50ml of 10M NaOH and the volume was made up to 1 litre with water. This solution was sterilised by autoclaving.

Tris-Cl [tris(hydroxymethyl) aminomethane], 1M (pH 8.0)

121g of Tris base was dissolved in 800ml water. pH was adjusted to 8.00 with ~42ml HCl and the volume was made up to 1 litre with water. This solution was sterilised by autoclaving.
Materials and methods

**TE (Tris/EDTA) buffer**

The buffer was prepared by mixing 1mM EDTA (2ml of 0.5M EDTA) and 10mM Tris-Cl (pH 8.00) (10ml of 1MTris-Cl). The volume was made up to 100ml with sterilised water.

**Lysozyme solution**

Lysozyme solution for lysis of Gram positive bacteria was prepared by dissolving 20mg/ml lysozyme solution from a 100mg/ml lysozyme stock solution in lysozyme buffer [20mM Tris-Cl (pH8.0), 2mM EDTA, 1.2% triton].

2.2 Preparation, Cultivation and Storage of Bacteria

2.2.1 Glycerol Stocks of Bacteria

For long term storage of bacteria, overnight culture suspension of individual bacteria were mixed with equal amount of glycerol solution (65%v/v glycerol, 0.1M MgSO₄, 0.025M Tris-HCl pH 8.0) and stored at -70ºC.

2.2.2 Enumeration of Colony Forming Units (C.F.U)

Culture counts were determined using the drop-plate method modification of the Miles Misra procedure. Ten-fold serial dilutions of cells were made in fresh media. Three 20μl sample droplets from each dilution were plated on to duplicate pre-dried agar plates. Plates were sealed with paraflim, inverted and incubated at 37ºC. Dilutions yielding a final count of 10 to 100 colonies were used to calculate the number of colony forming units (C.F.U) in 1ml volume.

2.2.3 Optical Density (OD) measurements

The absorbance of 1ml of culture was measured at wavelength of 600nm using the Sanyo SP75 UV/Vis spectrophotometer (Watford,UK). For optical density readings greater than one, 100ul of culture was diluted in 900ul of fresh media prior to absorbance readings.
2.2.4 Cultivation of *E. coli, P. aeruginosa*

To culture, glycerol stocks were streaked on to LB agar and incubated at 37°C for 24 hours, until single colonies were visible. Single colonies were emulsified in 10ml LB broth and incubated at 37°C overnight by shaking at 200rpm. These culture suspensions were used for DNA extraction, C.F.U counts, OD measurements and preparing glycerol stocks. Remaining culture suspension was distributed in 1ml volumes and the pellet was stored in 20°C for future DNA extraction.

2.2.5 Cultivation of *S. pneumoniae*

To culture, glycerol stocks were streaked on to blood agar and incubated in a candle-jar at 37°C for at least a day, until single colonies were visible. *Streptococcus* genus was confirmed by their haemolytic activity and Gram staining. Single colonies were emulsified in 10ml BHI broth and grown without shaking at 37°C overnight. These culture suspensions were used for DNA extraction, C.F.U counts, OD measurements and preparing glycerol stocks. Remaining culture suspension was distributed in 1ml volumes and the pellet was stored in -20°C for future DNA extraction.

2.2.6 Cultivation of *M. catarrhalis*

*M. catarrhalis* was cultured in the same manner as *S. pneumoniae*. These culture suspensions were used for DNA extraction, C.F.U counts, OD measurements and preparing glycerol stocks. Remaining culture suspension was distributed in 1ml volumes and the pellet was stored in -20°C for future DNA extraction.

2.2.7 Cultivation of *H. influenzae*

To culture, glycerol stocks were streaked on to Chocolate agar plate and incubated in a candle-jar at 37°C for at least a day, until single colonies were visible. Single colonies were emulsified in 10ml BHI broth with 20ul NAD (100mg/ml) and 100ul haematin (1mg/ml), and grown without shaking at 37°C overnight. These culture suspensions were used for DNA extraction, C.F.U counts, OD measurements and preparing glycerol stocks. Remaining culture suspension was distributed in 1ml volumes and the pellet was stored in -20°C for future DNA extraction.

2.2.8 Cultivation of *B. fragilis*

To culture, glycerol stocks were streaked on to blood agar in and incubated in an anaerobic chamber at 37°C for at 48hr, until single colonies were visible. Single colonies were emulsified in 10ml BHI broth and grown without shaking at 37°C
overnight in anaerobic chamber. These culture suspensions were used for DNA extraction, C.F.U counts, OD measurements and preparing glycerol stocks. Remaining culture suspension was distributed in 1ml volumes and the pellet was stored in -20ºC for future DNA extraction.

2.3 Patient Samples

2.3.1 COPD subjects and sample collection

Longitudinal COPD study

Samples were obtained from a COPD prospective observational clinical study performed in Glenfield (Leicester, UK). The study was approved by local research committee and written consent obtained from all the patients. 156 COPD patients aged over 45 and mostly between COPD GOLD stage II to IV were recruited for this study and a total of ~900 sputum samples collected from 145 subjects (11 withdrew from the study). The final database with samples matched to patient number and data available on both culture and molecular bacterial load enumeration consisted of 662 samples. Subjects for this study have had between 1 and 5 exacerbations requiring antibiotics and or corticosteroids in the preceding year.

Sputum samples were taken for microbiological and other analyses at all visits. After the baseline characterisation visit, sputum samples were collected at 3-monthly follow-up visits during stable state for 18months. Sputum samples were also collected at the onset of exacerbation and 2 and 6 weeks post-exacerbation. Stable samples, post-exacerbation, were collected at least 8 weeks apart from the onset of exacerbation. Where patients were unable to provide a spontaneous sample, sputum induction was performed. Induced sputum samples were collected at the Glenfield Hospital and sent to LRI for sputum homogenization and bacteriological plate culturing. A portion of this homogenized sputum aliquot was used for our molecular work.

Full details on study design, patient enrolment criteria, patient demographics, sputum collection and measurements performed have been described (Bafadhel et al., 2011). All patients gave informed written consent and the study was approved by the local ethics committee(Bafadhel et al., 2011). Apart from demographics, history and lung function measurements, data on expression level of various immunological markers, treatment
 Materials and methods

given at exacerbation, virology detection and culture based pathogen detection and plate
counts were also collected. Molecular technique based bacteriological data was
provided from this study.

Sputum collection

Most sputum samples in this study were produced spontaneously and collected in sterile
sputum pots by the study volunteers. In cases, where the patient didn't produce
spontaneous sputum it was induced. The procedure of induced sputum collection is
described below.

Induced sputum: Sputum induction was performed as according to the guidelines set
in the Glenfield Hospital (Bhowmik et al., 1998). Briefly the following steps are
involved in the sputum induction. Prior to sputum induction baseline spirometry is
performed to record the FEV$_1$ and the FVC, and 400ug Salbutamol is administered to
increase bronchodilation and spirometry was repeated. Before proceeding, the patient
blew their nose and rinsed their mouth; this was repeated after every saline inhalation.
5ml 3% hypertonic saline is dispensed into an ultrasonic nebuliser, the output was set at
over 0.8ml/minute and this was inhaled by the patient for 5 minutes through the
mouthpiece. Any sputum produced by coughing during the procedure is collected in a
sterile container.

2.3.2 Sample preparation, Bacteriological plating and storage

A part of the collected sputum plug was sent to LRI and rest was processed for the plate
enumeration of total bacterial load (Pye et al., 1995) in Glenfield on Blood agar and
chocolate agar. This involved separation of sputum plugs from the saliva, condense the
sputum plugs by moving the entire mass around the lid with circular motions and finally
gather in to a single mass. This concentrated sputum was weighed and 4X w/v of 0.2%
dithiothreitol (DTT) in 0.1X phosphate buffered saline was added. This suspension was
aspirated with pasteur pipette, then vortexed for 15 seconds and rocked for 15 minutes
on ice to homogenise the sputum which, was filtered through 48um nylon gauze into a
clean pre-weighed 15ml centrifuge tube.
100ul of this filtrate was used for 10 fold serial dilutions \((10^{-1} \text{ to } 10^{-4})\) with D-PBS and 3X 20ul drops in each quadrant of the respective dilution were placed on blood and chocolate agar plate culture. Plates were incubated at 37°C for 24 hours in 5% CO\(_2\). After 24 hours colonies were counted and averaged from all the three drops of a dilution series, which had > 30 and < 300 colonies to calculate colony forming units/ml.

The remainder of the sputum that was sent to LRI for routine culture testing (Health Protection Agency, 2009) was diluted with 1X sputasol (0.1% DTT). This suspension was agitated for 10 seconds and incubated at 37°C for 15 minutes. This homogenized sputum was agitated again for 10 seconds. A portion of this suspension was stored at -80°C for molecular analysis and the rest was used for isolation of potential pathogenic bacteria (PPB).

The homogenized sputum was treated as the neat suspension. 1 blue loopful (10ul) of ‘neat’ sputum was added to 5ml of sterile distilled water and this is 10\(^{-3}\) dilution. This was agitated for 10 seconds and 10ul of this dilution was streaked on to chocolate agar plate with bacitracin and blood agar plates with optocin and checked for following bacterial growth: \(H.\ influenzae, S. pneumoniae, S. aureus, M. catarrhalis\) and \(P.\ aeruginosa\) as per the HPA standards (Health Protection Agency, 2009). Culture that did not show a dominant pathogenic respiratory pathogen was classified as non-significant growth.

Remaining filtrate was placed in a sterile eppendorff for our molecular analysis and stored at -80°C.

2.4 Molecular Methods

2.4.1 Real time PCR and primers
For bacterial quantification studies the real time assay chemistry, primer sequence, cycling conditions utilized and data analysis methodology are mentioned in detail in the relevant chapters ahead.

2.4.2 High-throughput parallel sequencing
16S rDNA primers complementary to conserved regions of 16S bacteria gene were utilized to capture the hypervariable region within the bacterial amplicons for studying the bacterial community composition in COPD samples. Again primer sequence,
Materials and methods

amplicon library production and data analysis methodology is described in the relevant chapter ahead.

2.4.3 Gel analysis of PCR amplicons

Pure culture DNA and PCR products were analysed by gel electrophoresis, to verify amplicon size, non-formation of non-specific PCR product and check DNA purity isolated from pure cultures. All agarose gels contained 0.5μg/ml ethidium bromide. Gel concentration of 1% to 2% was utilized for electrophoresis depending on the size of the DNA fragment to be separated. DNA was run along with 1x gel loading dye. Pure culture DNA were run along with lambda DNA as standards and amplicons with 5μg of GeneRuler 100-bp DNA ladder (Fermentes Life Sciences, York). Gel electrophoresis was performed at 90-100V for 45-60 minutes, depending on the gel size. After electrophoresis DNA bands on gel were visualized under the G:BOX UV transilluminator (Syngene, Cambridge, UK).

2.4.4 Cloning and sequence analysis of Streptococcus species

Bacterial recombinase subunit (recA) gene targeting primers, recAF - 5' GCCTTYATCGATCGBGARCA 3' and recAR- 5' GTTTCCGGRDCCRAACAT 3', specific to Streptococcus species were utilized to amplify the Streptococcus species in the COPD samples (Zbinden et al., 2011). Final PCR reaction consisted of 1μM of each of the primers (IDT), 0.2mM dNTP, Platinum high fidelity Taq DNA polymerase (1U) and 1x Platinum high fidelity Taq buffer (Life Technologies, UK). PCR were carried out in 50μl volume with 1μl template DNA using following conditions 95°C for 2 minute(initial denaturation) followed by 35 cycles of 95°C for 1 minute, 60°C for 30 seconds, 72°C for 1 minute and final extension at 72°C for 5 minutes producing a ~354bp PCR product. PCR amplicons were purified using the QIAquick PCR purification kit (Qiagen, California, USA) according to the manufacturer's protocol. Purified PCR products were eluted in 30μl molecular grade water (Sigma-Aldrich, UK). For cloning purified PCR products were ligated into pGEM-T Easy Vectors (Promega). Ligation reaction were carried out in 10μl volume containing the PCR product and pGEM T Easy vector in a 1:1 ratio along with 1μl T4 DNA ligase and 1x ligation buffer. Reactions were incubated at 4°C overnight. 2 μl of each ligation reaction was added to 50μl of the competent cells (E. coli DH5α cells), mixed by gentle flicking, and placed on ice for 20 minutes. Cells were heat-shocked in a water bath at exactly 42°C for 45-50 seconds for transformation, returned immediately to ice for 2 minutes, 950 μl
Materials and methods

of LB broth added, and the tubes incubated at 37°C with shaking at 200 rpm for 1.5 hours. 100 μl each of neat and 1:10 diluted transformation culture was plated onto duplicate LB/ampicillin (100μg/ml)/IPTG(0.5mM)/X-Gal(80μg/ml) plates. Plates were incubated at 37°C for 16-24 hours. White colonies containing inserts were randomly selected and checked for the insert by performing colony PCR using the SP6 and T7 promoter primer sequence, complementary to the vector sequence, flanking the cloning region of the vector. Plasmid extraction was performed, on clones confirmed carrying insert by colony PCR, using the GenElute plasmid miniprep kit (Sigma Aldrich) according to manufacturer's protocol.

Plasmid DNA samples were sequenced in both directions from GATC Biotech. Resulting DNA sequences were analysed by BioEdit sequence alignment (http://www.mbio.ncsu.edu/BioEdit/bioedit.html) program. DNA sequences were aligned using the ClustalW alignment, NCBI blast (http://blast.ncbi.nlm.nih.gov/Blast.cgi) was performed and a tree of sequence similarity drawn using the neighbour joining method.
3 Changes in the total and specific bacterial load at stable and exacerbation state of COPD

3.1 Introduction:
As discussed earlier (1.3.2) the presence of pathogenic bacteria in the respiratory tract is considered by some to be an opportunistic event and a secondary factor in COPD pathogenesis. Although, infection is believed to be the leading cause of exacerbation, PPB are not isolated from all exacerbation samples and their occurrence in stable COPD has led to conflicting opinions of their role in COPD. In broad terms, there are two dominant hypotheses implicating bacterial infection in the pathogenesis of COPD exacerbations:

1. *Proliferation of existing pathogenic bacteria colonising the airways.*

It has been suggested that there is a relationship between bacterial load and host immune response. Low bacterial loads are effectively cleared by host immune responses, while bacterial loads of $>10^6$ C.F.U/ml are not contained by host immunity, leading to further bacterial proliferation and exaggerated host inflammation in a vicious cycle (Abusriwi and Stockley, 2007). COPD is a disease characterised by airway inflammation and chronic injury. It is considered to provide an opportunistic environment for colonisation with respiratory pathogens. During stable COPD presence of PPB in airways is considered as colonisers and their numbers contained by the host immune system while, exacerbation episodes are hypothesized as events when this equilibrium is disrupted resulting in an increase in bacterial (pathogens) load and their dominance (Wilson, 1998). The mechanisms leading to proliferation however remain unclear.

2. *Acquisition of new pathogenic bacterial strains at exacerbation.*

Here it is hypothesised that exposure of the airways to new strains of pre-existing bacterial species stimulates a vigorous immune response, leading to exacerbation (Sethi et al., 2002).
Various groups have provided evidence for and against these hypotheses. However, most studies have been limited by the use of culture techniques alone, which lack sensitivity for detection and provide only crude estimates of the quantitative change in bacterial load. Furthermore, many studies have been cross-sectional in design and not equipped to specifically examine within-subject change between stable and exacerbation states.

To address these limitations, qPCR assays were developed in this project to study the following:

1. To explore if COPD exacerbation is associated with quantifiable changes in total bacterial load (TBL), both at population and within subject level.

2. To examine if differences exists in detection rate and quantity of some commonly encountered COPD pathogens between stable and exacerbation state. qPCR assays targeting the bacterial species *H. influenzae*, *M. catarrhalis*, *S. pneumoniae* and *S. aureus* were developed to enumerate these specific bacterial load (SBL).

3. To compare the qPCR assay results of TBL and SBL with culture detection and colony forming unit counts (C.F.U) and examine how the clinical definition of bacterial infection in COPD correlates with organisms identified using both techniques.

### 3.2 Materials and Methods:

#### 3.2.1 Bacterial strains and Study sample

DNA was extracted from the bacterial strains, for qPCR optimisation assay and standards, are listed in (2.1.1). For preliminary bacterial molecular analysis and optimisation of TBL assays, sputum collected from another COPD clinical trial at Glenfield hospital was utilised. Twenty-seven stable COPD patients, with samples collected at baseline and after treatment at 7, 14 and 28 days, were used as pilot samples. DNA from sputum collected in the longitudinal COPD study mentioned in (2.3.1) was used for exploring the differences in bacterial composition at different stages of COPD.
3.2.2 DNA Extraction

3.2.2.1 DNA extraction from sputum samples
Total genomic DNA for all sputum samples was extracted using the QIAamp DNA Mini Kit (Qiagen, California, USA). For the pilot study, due to availability of very small sample volumes, 20μl of the homogenized sputum was directly used, while for the longitudinal COPD study bacteria were harvested from 500μl of homogenized sputum. DNA isolation from the Gram positive bacteria extraction method was followed as per the Manufacturer’s protocol, which involved hydrolysis of peptidoglycan cell wall layer with 20mg/ml lysozyme and incubated at 37°C for 30 minutes. This was followed by Proteinase K digestion of contaminating proteins and nucleases released from the cells at 55°C for 30 minutes and 95°C for 15 minutes. The remainder of the extraction was done according to the "DNA Extraction from Tissue" of the manufacturer’s protocol. This involved adding the cell lysate formed from the above step with 200μl absolute ethanol to the QIAamp spin column and centrifuging briefly to adsorb the DNA optimally to the column’s silica gel membrane. This was followed by wash steps with buffer AW1 and AW2 to remove the impurities and the DNA was eluted in 200μl of DNAse, RNAse free distilled water. The DNA was stored at -20°C.

3.2.2.2 DNA extraction from Gram positive cultures
DNA was extracted from 1ml overnight cultures of *S. pneumoniae* and *S. aureus* using the QIAamp DNA mini kit following the ‘DNA extraction from Gram positive culture for bacterial cell lysis and ‘DNA extraction from Tissue’ method for the DNA adsorption, cleanup and elution as per Manufacturer’s protocol.

DNA was eluted in 200μl of DNAse, RNAse free distilled water. DNA was quantified by nanodrop and stored at -80°C.

3.2.2.3 DNA extraction from Gram Negative cultures
DNA was extracted from 1ml overnight cultures of *H. influenzae* and *M. catarrhalis* using the QIAamp DNA mini kit. Harvested bacteria were extracted following the ‘DNA isolation from bacterial cultures’ extraction method which involved lysis with proteinase K at 55°C for an hour followed by DNA adsorption clean up and elution as
Changes in the total and specific bacterial load at stable and exacerbation state of COPD

outlined in ‘DNA extraction from Tissue’ in the Manufacturer’s protocol. DNA was quantified by nanodrop and stored at -80°C.

3.2.2.4 Other DNA release and purification methods used

**microLYSIS:**

MicroLYSIS (Microzone, West Sussex) is an optimised enzyme solution for one step DNA release method from bacteria. For speed and convenience, it was used in preliminary DNA extraction optimisation assays. For pure cultures, 2μl overnight culture was mixed with 18μl of 20X microLYSIS and placed in a thermal cycler (Dyad DNA engine, MJ Research, USA). The thermal cycling profile used was as follows:

- 65°C for 5 minutes;
- 96°C for 2 minutes;
- 65°C for 4 minutes;
- 96°C for 1 minute;
- 65°C for 1 minute;
- 96°C for 30 seconds and final hold at 20°C.

After cycling, this mixture of microLYSIS DNA was directly used for PCR reactions or stored at -20°C for future use.

**MolYsis DNA isolation kit:**

The MolYsis kit (Molzym GmbH & Co.KG, Germany) includes pre-treatment steps for the removal of extracellular and eukaryotic DNA prior to column based DNA extraction from bacterial cells. The eukaryotic DNA removal element of the kit was used prior to extracting bacterial DNA from sputum samples with Qiagen DNA. The following steps were performed for enriching bacterial cells and removing human DNA from samples:

- 250μl of buffer CM was mixed with 500μl of sputum sample and incubated at room temperature for 5 minutes to lyse eukaryotic cells.
- Released DNA from cells was degraded with 250μl of buffer DB1 and 10 μl MolDNase B, incubated at 25°C for 15 minutes. Bacterial cells were then centrifuged down and the supernatant discarded and the bacterial pellet washed with 1 ml buffer RS to remove any residual DNase B and PCR inhibitors in the solution. Finally the bacterial cell was harvested by centrifugation and removing the supernatant.

After this DNA was extracted from the samples using the Qiagen kit as described in methods section 3.2.2.1
3.2.3 DNA Concentration measurement

DNA concentration was measured using NanoDrop 1000 (Thermo Scientific, UK). 1μl of water was loaded as blank for the DNA concentration measurements. 1μl of DNA sample was loaded on to the NanoDrop and absorbance was read at 260nm ($A_{260}$).

The concentration of DNA was calculated using Equation 3.1, derived from $A_{260} = 1$ being equivalent to a double stranded (ds) DNA concentration of 50μg/ml according to the Beer Lambert law (Ingle, 1988)

$$C = A_{260} \times D \times 50$$

where

- $C$ = Concentration (μg/ml)
- $A_{260}$ = Absorbance at 260nm
- $D$ = Dilution factor

3.2.4 Primers and oligonucleotide sequence used for the real time assays

Table 3.1 lists the primer names, target bacteria and gene and oligo sequences used for the qPCR assays
Table 3.1

<table>
<thead>
<tr>
<th>Primer (Name)</th>
<th>Target Organism (gene)</th>
<th>Sequence</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>16S 338F</strong></td>
<td>All Bacteria (16S rDNA)</td>
<td>5' ACTCCTACGGGNGGCNGCA 3'</td>
<td>(Free, 2005)</td>
</tr>
<tr>
<td><strong>16S 515R</strong></td>
<td>M.Nad16S F</td>
<td>5' TCCTACGGGAGGCGAGCAGT 3'</td>
<td>(Nadkarni et al., 2002)</td>
</tr>
<tr>
<td><strong>M.Nas16S R</strong></td>
<td>All bacteria (16S rDNA)</td>
<td>5' GGACTACCGGGGTATCTAATCTGT 3'</td>
<td>(Nadkarni et al., 2002)</td>
</tr>
<tr>
<td><strong>M.Nad16S R</strong></td>
<td>M.Nad16S R</td>
<td>5' FAM-CGTATTACCGCGGCTGCTGGCAC-TAMRA 3'</td>
<td>(Nadkarni et al., 2002)</td>
</tr>
<tr>
<td><strong>P6F</strong></td>
<td>Haemophilus influenzae (Outer Membrane Protein P6)</td>
<td>5' TTGGCGGWTACTCTGTGCT 3'</td>
<td>(Fang and Hedin, 2003; Strålin et al., 2005)</td>
</tr>
<tr>
<td><strong>P6R</strong></td>
<td>S.aurF</td>
<td>5' TGCAGGTTTTTCTCCACCCT 3'</td>
<td>(Fang and Hedin, 2003)</td>
</tr>
<tr>
<td><strong>S.aurF</strong></td>
<td>Staphylococcus aureus (Thermonuclease)</td>
<td>5' GCGATTTGATGGTGATACGGTT 3'</td>
<td>(Fang and Hedin, 2003)</td>
</tr>
<tr>
<td><strong>S.aurR</strong></td>
<td>S.aurF</td>
<td>5' AGCCAAGGCTTGAGACAACTAAAAGC 3'</td>
<td>(Fang and Hedin, 2003)</td>
</tr>
<tr>
<td><strong>S.pneuF</strong></td>
<td>Streptococcus pneumoniae (Pneumolysin)</td>
<td>5'-AGCGATAGCTTTTCCAAAGTG-3'</td>
<td>(Greiner et al., 2001)</td>
</tr>
<tr>
<td><strong>S.pneuR</strong></td>
<td>S.pneuR</td>
<td>5' CTTAGCACAACAAATTGCCTTACCG 3'</td>
<td>(Greiner et al., 2001)</td>
</tr>
<tr>
<td><strong>S.pneu probe</strong></td>
<td>S.pneu probe</td>
<td>5' Cy5-ACCCCAGCAATTGAGTGTTGC-3'</td>
<td>(Greiner et al., 2001)</td>
</tr>
</tbody>
</table>
Changes in the total and specific bacterial load at stable and exacerbation state of COPD

| Primer and probe sequence for TBL and SBL qPCR: Base degeneracies as follows: R= A/G, Y= C/T, M=A/C, K=G/T, S=G/C, W=A/T, H=A/C/T, B=C/G/T, V=A/C/G, D=A/G/T, N= A/C/G/T. Probe sequences have the dye name at the 5' end and the quencher name at 3' end of the sequence. |
|---|---|
| **M.catF** | **Moraxella catarrhalis** (Outer Membrane Protein CopB) |
| **M.catR** | 5' GTGAGTGCCGCTTTACAACC 3' |
| **M.cat probe** | 5' TGTATCGCTGCCAGACAA 3' |
| **BHQ2 3’** | (Greiner et al., 2003) |
| 5' JOE- TGCTTTTGCAGCTGGTAGCCAGCTAGA-TAMRA 3' |

**3.2.5 Real Time PCR**

Real Time PCR follows the general principle of polymerase chain reaction and allows detection of the amplified DNA in real time as the reaction progresses. In real time PCR quantification is achieved by detection of a fluorescent dye that accumulates in direct proportion to the yield of amplified PCR products using various assay chemistries. For this project all qPCR assays were performed based on either SYBR Green or TaqMan chemistries. The SYBR Green assay is based on the principle of high binding specificity of SYBR Green dye to dsDNA. During polymerization, multiple SYBR green molecules bind to dsDNA and emit fluorescence that is detected in directly proportion to the amount of DNA present in a sample. For the TaqMan assay, the reporter dye is present at the 5 prime ends and its fluorescence is quenched by the quencher dye present at the 3prime end of the probe. During the annealing step of the PCR cycle the 5 prime nuclease activity of Taq polymerase cleaves off the reporter dye and separates it from quencher emitting fluorescence.
3.2.5.1 DNA standards for quantification assays

Absolute quantitation on real time PCR refers to an analysis where unknown samples are compared to a standard curve. The standard is a known DNA sample whose absolute concentration is known.

DNA from pure culture of the target gene was used to prepare standards. From the concentration of pure culture DNA obtained, genomic equivalent/μl or targeted gene copies/μl, when the target gene is present in single copy, was determined using the following equation (Equation 3.2):

**Equation 3.2: Determining genome equivalent for pure culture standards**

\[
\text{Genome equivalent/μl} = \frac{(6.02 \times 10^{23} \, \text{copies/mole}) \times (\text{concentration in g/μl})}{(\text{Molecular weight (MW) in g/mole})}
\]

Where

| 6 x 10^{23} \, \text{copies/mole} | Avogadro’s constant |
| MW (g/mole) | Size of the bacterial genome (bp) x 649Dalton/bp |
| 649 | average molecular weight of a DNA base pair (g/mole) |

As multiple copies of 16S rDNA are present in many bacteria, the 16S gene copy number of individual bacteria was used to prepare DNA standards for the 16S genes targeted qPCR. For this calculation, the copy number was multiplied to the genomic equivalent/μl value obtained using equation 1.2. After determining the genome equivalent/μl of the pure culture DNA, 10-fold dilution series were prepared ranging from 10^7 to 10^2 genome/μl.
Changes in the total and specific bacterial load at stable and exacerbation state of COPD

Table 3.2: Bacterial DNA standards used for TBL and SBL qPCR

<table>
<thead>
<tr>
<th>Pure cultures used for DNA standards</th>
<th>Target Gene</th>
<th>Gene copy number</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Streptococcus pneumoniae</em> D39</td>
<td><em>Ply</em> (pneumolysin)</td>
<td>1</td>
</tr>
<tr>
<td><em>Staphylococcus aureus</em> Newman</td>
<td><em>Nuc</em> (thermonuclease)</td>
<td>1</td>
</tr>
</tbody>
</table>

3.2.5.2 Quantification Assays

Two types of real time assay chemistry were utilized for DNA quantification assays:

SYBR Green Assays (intercalating dye assay): To determine the TBL, *S. aureus* and *H. influenzae* in sputum samples individual SYBR Green quantification assays were performed targeting 16S gene with 16S rDNA broad range primers, nuc gene targeting primers and P6 gene targeting primers respectively.

PCR reactions were set up in a PCR cabinet (Scie-Plas, UK) 20 minutes after UV irradiation to the plastic wares, racks, water and pipettes. Filter tips were used for molecular work to avoid cross contamination. Reactions for the TBL assay were prepared with 12.5μl 2x Absolute QPCR SYBR Green PCR mix (ThermoFisher Scientific, UK), 1μl of 16S 338F forward primers (6.25μM) and 1μl of 16S 515R reverse primers (6.25μM). The final reaction volume of 25μl was made by adding 9.5μl water and 1μl of template DNA. *E. coli* DNA was used as the standard and 1μl of *E.coli* DNA serial dilutions from $10^7$ to $10^2$ genome/μl were prepared. Two negative controls (sterile distilled RNase DNase water instead of DNA) were also included.

Reactions for *S. aureus* and *H. influenzae* assays were prepared in the same manner as described above, with the exception that for *S. aureus* assay 1μl of S.aurF forward primer (5μM) and 1μl of S.aurR reverse primer (5μM) and for the *H. influenzae* assay 1μl of P6F forward primer (5μM) and 1μl of P6R reverse primer (5μM) were used. *S. aureus* and *H. influenzae* DNA were respectively used as standards for *S. aureus* and *H. influenzae* qPCR. The qPCR protocol began with initial denaturation at 95°C for 15
minute to achieve activation of the Taq polymerase in the PCR mix. The amplification step was repeated for 40 cycles, as follows:

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Time</th>
<th>Type of Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>95°C</td>
<td>20 secs</td>
<td>Denaturation</td>
</tr>
<tr>
<td>60°C</td>
<td>30 secs</td>
<td>Annealing</td>
</tr>
<tr>
<td>72°C</td>
<td>20 secs</td>
<td>Elongation</td>
</tr>
<tr>
<td>* 76°C/ 80°C</td>
<td>20 secs</td>
<td>Data acquisition green (470nm/510nm) channel</td>
</tr>
</tbody>
</table>

(* for the *H. influenzae* and *S. aureus* assays, data was acquired at 76°C and for the 16S assay at 80°C. 470nm - excitation wavelength; 510nm - detection wavelength)

Finally melt (T_m) curve analysis was performed from 65°C to 99°C to determine the specificity of the amplicons.

**TaqMan Assays (dual labelled probe assay)**

Reactions for total bacterial load assay were prepared with 12.5μl 2x QPCR mix, 1μl of 16S 331F forward primer (17.5μM), 1μl of 16S 797R reverse primers (17.5μM), 1μl of ‘FAM-TAMRA’ labelled 16S probe (2.5μM) and 1μl of Template DNA. The final reaction volume of 25μl was made by adding 8.5μl water.

*S. pneumoniae* and *M. catarrhalis* were quantified in sputum by individual TaqMan assays targeting the pneumolysin gene and copB gene respectively.

Reactions for *S. pneumoniae* were prepared with 12.5μl 2x Absolute QPCR mix (ThermoFisher Scientific,UK), 1μl of each S.pneuF (17.5μM), S.pneuR (12.5μM) and S.pneu probe (2.5μM) and 1μl of Template DNA. The final reaction volume of 25μl was made by adding 8.5μl water. For *M. catarrhalis* 1μl of each M.catF (2.5μM), M.catR (12.5μM), and M.cat probe (2.5μM) were used and the reaction was set up in the same manner as the *S. pneumoniae* assay. For both *S. pneumoniae* and *M. catarrhalis* assays DNA for the respective organisms was used as the standard.
The qPCR cycle began with denaturation at 95°C for 15 minutes followed by 40 cycles of amplification step as follows:

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Time</th>
<th>Type of Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>95°C</td>
<td>20secs</td>
<td>Denaturation</td>
</tr>
<tr>
<td>60°C</td>
<td>60secs</td>
<td>* Annealing and data acquisition</td>
</tr>
</tbody>
</table>

(* Data for 16S assay was acquired on green channel (470nm/510nm); *S. pneumoniae* assay on red channel (625nm/660nm); and *M. catarrhalis* on yellow channel (530nm/555nm).

### 3.2.5.3 qPCR data analysis

All real time PCR were performed on Roto-Gene 6000 (Corbett Life Science, UK). qPCR signals were analysed with the Roto-Gene 6000 Series Software1.7. The data collected was normalized to compensate for background fluorescence and a threshold level set for data analysis. The Ct-value (threshold cycle), defined as the number of cycles required for a sample to reach the threshold level, was set at a level where the rate of amplification was greatest during the exponential phase. Samples from runs with correlation coefficients (R²) above 0.98, amplification efficiencies above 0.7 and replicates with less than 1 Ct difference were used for further analysis (*Dorak, 2006*).

No more than 15% of the samples failed to meet these parameters requiring repetition of the qPCR. Of the assays conducted the highest failure rate was seen in the 16S rDNA based TBL qPCR.

Melt curve analyses performed for *H. influenzae* and *S. aureus* were used to confirm the identity of PCR products by comparing with the melting temperature (Tm) of the product from the DNA standard. Samples with Tms within the range of the Tm of the DNA standard product ±0.5°C were regarded as true.

### 3.2.5.4 Calculations of genome/ml from real time assay

For the SBL qPCR genome/ml was calculated as follows:
Genome/ml = Average copies/μl x 200 μl x 2

(200 μl is the volume of eluted DNA and 2 is the Dilution factor (DNA was extracted from 500μl sample))

For TBL qPCR genome/ml was calculated as follows:

Genome/ml = (Average copies/μl x 200μl x 2)/5

(5 was taken as an average 16S copies for mixed respiratory bacteria in sputum samples refer to result section 3.3.1.2.2)

3.2.6 Statistical analysis

All statistical analyses were performed in Prism 5.00 (GraphPad software, La Jolla California USA). qPCR readings were log transformed to achieve a normal distribution. For continuous parametric data, between group and paired sample comparisons were performed using the two tailed unpaired student t-test and paired t test respectively. Corresponding analysis of non-parametric data was performed using the Mann-Whitney test (between sample / group) and Wilcoxon signed rank test (paired samples). For categorical data significance testing was performed using the chi-square test. Pearson correlation, Bland Altman agreement and ROC curve based analyses were performed to study the relationship between data sets. A p value of <0.05 was set as the significance threshold for all analyses.

3.3 Results

3.3.1 Developing and optimizing qPCR assays for quantifying TBL and for SBL in COPD sputum

There were 2 major aspects to this:

1. Extraction of the total bacterial DNA from the sputum samples

2. Selecting primer pairs and developing the qPCR assays.

3.3.1.1 DNA Extraction Methods

There was wide variation in the volume of sputum expectorated by COPD patients, ranging from 20μl to 1ml. Sputum samples for the pilot study (obtained from a separate
clinical study, as previously described (3.2.1) used for optimizing the extraction and 16S based qPCR had a maximum volume of 20μl. Two different commercially available extraction methods, Qiagen DNA mini kit and microLYSIS reagent, were tested.

MicroLYSIS provided a single microLYSIS solution for DNA extraction from both Gram positive and Gram negative bacterial cells. Although microLYSIS produced closer 16S rDNA qPCR readings to C.F.U counts than the Qiagen method for *E. coli*, it failed to produce any detectable reading from pure cultures of *S. aureus* on multiple occasions (Table 3.3). Extraction of a 1:1 mixture of *E. coli* and *S. aureus* culture also didn't produce any readings (Table 3.3).

Qiagen DNA mini kit for DNA extraction of Gram positive bacteria includes an extra cell lysis pretreatment step with lysozyme for hydrolyzing the thicker peptidoglycan layer present in Gram positive bacteria. To be representative of all bacterial cell types (both gram negative and gram positive), the Gram Positive Bacterial DNA extraction method was the desirable method for the TBL qPCR assay.

Finally with the sputum samples using identical starting sample volumes, the Qiagen Gram positive protocol yielded a greater 16S amplicon copy number than either microLYSIS or the Qiagen Gram negative protocol (representative data Table 3.4). For these reasons, the Qiagen Gram positive extraction protocol was selected for DNA extraction from sputum samples in this study.
Table 3.3

<table>
<thead>
<tr>
<th>Pure cultures</th>
<th>Qiagen</th>
<th>MicroLYSIS</th>
<th>C.F.U</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli</td>
<td>2.144E+08</td>
<td>2.490E+09</td>
<td>1.765E+09</td>
</tr>
<tr>
<td></td>
<td>1.899E+08</td>
<td>2.011E+09</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.759E+08</td>
<td>2.595E+09</td>
<td></td>
</tr>
<tr>
<td>S. aureus</td>
<td>2.697E+08</td>
<td>3.203E+08</td>
<td>1.365E+09</td>
</tr>
<tr>
<td></td>
<td>2.858E+08</td>
<td>No Readings</td>
<td></td>
</tr>
<tr>
<td>MIX (1:1)</td>
<td>2.611E+08</td>
<td>No Readings</td>
<td>1.800E+09</td>
</tr>
<tr>
<td></td>
<td>1.907E+08</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.572E+08</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Efficiency of Qiagen and microLYSIS DNA extraction method: MIX represents an equal volume mixture of *E. coli* and *S. aureus* overnight culture. DNA was extracted in triplicates from each pure culture and the MIX. 16S qPCR assay was performed to enumerate the genome/ml readings from the extracted DNA. C.F.U/ml counts were also performed on the cultures used for DNA extraction.

Table 3.4

<table>
<thead>
<tr>
<th>Sputum samples</th>
<th>16S rDNA amplicon copies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Qiagen (+)</td>
</tr>
<tr>
<td>RF26</td>
<td>2.72E+06</td>
</tr>
<tr>
<td>Mk23</td>
<td>1.14E+06</td>
</tr>
</tbody>
</table>

Representative sputum samples comparing 16S amplicon copies using different extraction protocol: Mean 16S amplicon copies produced from DNA of the sputum samples extracted by Qiagen Gram positive method (Qiagen +), Qiagen Gram negative (Qiagen -) and microLYSIS are presented. 5μl of sample was used for each of the extraction methods. Each sample was quantified in triplicate. The no template control (NTC) reading was 8.34E+02.
3.3.1.2 Development and Optimisation of TBL and SBL qPCR

3.3.1.2.1 Optimisation of 16S rDNA based TBL qPCR

The total bacterial load of sputum samples was determined by performing two different primer pairs utilizing two different qPCR chemistries:

1. SYBR Green (dsDNA binding dye) assay with (338F and 515R) 16S universal primer pair (Free, 2005)

2. Hydrolysis based TaqMan assay with M.Nad 16S universal primer pair and probe (Nadkarni et al., 2002)

For both types of qPCR assay chemistry, commercially available pre-optimized qPCR mixes were used requiring only addition of primers and template DNA. Therefore the primer pair concentration with SYBR Green and concentrations of primer pair and probe for TaqMan were optimised for the respective assays.

SYBR Green assay primer concentration optimization: The optimization of primer concentrations was performed using various concentrations of forward and reverse primers ranging from 50nM to 250nM in increments of 50nM (Roto-Gene 6000 Manual). The dilution series of primers was tested with 1μl DNA template of a sputum sample and the 3 lowest Ct producing primer pair concentrations were assayed further to find the optimum primer pair concentration producing a low Ct value with DNA but a high Ct value with NTC (Table 3.5). A primer pair concentration of 250nM produced the lowest Ct value with the DNA template and highest Ct value with the NTC.

TaqMan Assay primer and probe concentration optimization: For the TaqMan assay, optimization was performed by testing each of the 3 different probe concentrations (100nm, 150nm and 250nm) with various concentrations of both primers ranging from 100nm to 700nm in increments of 200nm (Roto-Gene 6000 manual). The dilution series of primers was tested with 1μl DNA template of a sputum sample. The
Changes in the total and specific bacterial load at stable and exacerbation state of COPD

probe concentration of 100nm with 700nm of each forward and reverse primer produced the lowest Ct value with the DNA sample (Table 3.6).

### Table 3.5

<table>
<thead>
<tr>
<th>Primer concentration (F/R)</th>
<th>50/50</th>
<th>50/100</th>
<th>50/150</th>
<th>50/250</th>
<th>100/50</th>
<th>100/100</th>
<th>100/150</th>
<th>100/250</th>
<th>150/50</th>
<th>150/100</th>
<th>150/150</th>
<th>150/250</th>
<th>250/50</th>
<th>250/100</th>
<th>250/150</th>
<th>250/250</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ct reading (DNA Sample)</td>
<td>34.59</td>
<td>33.85</td>
<td>28.03</td>
<td>25.14</td>
<td>31.73</td>
<td>25.81</td>
<td>24.06</td>
<td>20.68</td>
<td>32.81</td>
<td>24.03</td>
<td>20.51</td>
<td>17.72</td>
<td>29.25</td>
<td>23.91</td>
<td>16.61</td>
<td>16.85</td>
</tr>
</tbody>
</table>

**Optimization of primer pair concentration for 16S SYBR Green assay:** Table (A) shows different combinations of forward (F) and reverse (R) primer concentrations tested with 1μl of a DNA sample and the cycle threshold (Ct) value of the amplification signal obtained with SYBR Green. Columns highlighted in red were producing the lowest Ct value and chosen for further testing. Table (B) shows the Ct value obtained in the DNA sample and NTC with the given primer concentration. The primer concentration 250/250 (highlighted red) was chosen for all further 16S SYBR green assays.
Changes in the total and specific bacterial load at stable and exacerbation state of COPD

Table 3.6

<table>
<thead>
<tr>
<th>Probe concentration-100nM</th>
<th>100/100</th>
<th>22.58</th>
<th>100/100</th>
<th>23.89</th>
<th>100/100</th>
<th>23.32</th>
</tr>
</thead>
<tbody>
<tr>
<td>100/300</td>
<td>23.12</td>
<td></td>
<td>100/300</td>
<td>22.97</td>
<td>100/300</td>
<td>22.9</td>
</tr>
<tr>
<td>100/500</td>
<td>21.86</td>
<td></td>
<td>100/500</td>
<td>22.8</td>
<td>100/500</td>
<td>23.19</td>
</tr>
<tr>
<td>100/700</td>
<td>22</td>
<td></td>
<td>100/700</td>
<td>24.16</td>
<td>100/700</td>
<td>22.61</td>
</tr>
<tr>
<td>300/100</td>
<td>18.78</td>
<td></td>
<td>300/100</td>
<td>20.46</td>
<td>300/100</td>
<td>19.32</td>
</tr>
<tr>
<td>300/300</td>
<td>18.94</td>
<td></td>
<td>300/300</td>
<td>19.87</td>
<td>300/300</td>
<td>19.19</td>
</tr>
<tr>
<td>300/500</td>
<td>18.71</td>
<td></td>
<td>300/500</td>
<td>19.61</td>
<td>300/500</td>
<td>18.78</td>
</tr>
<tr>
<td>300/700</td>
<td>18.54</td>
<td></td>
<td>300/700</td>
<td>19.61</td>
<td>300/700</td>
<td>18.74</td>
</tr>
<tr>
<td>500/100</td>
<td>18.01</td>
<td></td>
<td>500/100</td>
<td>19.56</td>
<td>500/100</td>
<td>19.07</td>
</tr>
<tr>
<td>500/300</td>
<td>18.32</td>
<td></td>
<td>500/300</td>
<td>19.35</td>
<td>500/300</td>
<td>18.58</td>
</tr>
<tr>
<td>500/500</td>
<td>18.67</td>
<td></td>
<td>500/500</td>
<td>18.59</td>
<td>500/500</td>
<td>18.5</td>
</tr>
<tr>
<td>500/700</td>
<td>18.01</td>
<td></td>
<td>500/700</td>
<td>18.9</td>
<td>500/700</td>
<td>18.41</td>
</tr>
<tr>
<td>700/100</td>
<td>18.88</td>
<td></td>
<td>700/100</td>
<td>19.13</td>
<td>700/100</td>
<td>19.36</td>
</tr>
<tr>
<td>700/300</td>
<td>18.06</td>
<td></td>
<td>700/300</td>
<td>18.2</td>
<td>700/300</td>
<td>18.31</td>
</tr>
<tr>
<td>700/500</td>
<td>17.91</td>
<td></td>
<td>700/500</td>
<td>17.75</td>
<td>700/500</td>
<td>18</td>
</tr>
<tr>
<td>700/700</td>
<td>16.93</td>
<td></td>
<td>700/700</td>
<td>18.33</td>
<td>700/700</td>
<td>18.21</td>
</tr>
</tbody>
</table>

Optimisation of primer pair and probe concentration for 16S TaqMan assay: Cycle threshold values (Ct) obtained with different combinations of forward (F) and reverse (R) primers at 3 different probe concentrations. The primer concentration of 700 F / 700 R with a probe concentration of 100 nM (highlighted in red) gave the lowest Ct.

3.3.1.2.2 Correction of 16S gene copy variation and the no template control (NTC) reading for enumerating TBL

One of the limitations of enumerating the TBL in samples using 16S gene was the occurrence of multiple 16S gene copies in bacteria and the variability in this copy number between different bacteria (Acinas et al., 2004; Crosby and Criddle, 2003; Farrelly et al., 1995). This meant that 16S rDNA gene based qPCR reading will not be equivalent to the bacterial load (genome/ml) present in sample.

The rRNA operon copy number ranges from 1 to 15. Nearly 40% of bacterial strains have either one or two 16S rDNA copies, while bacteria with 3 to 7 rRNA copies contribute about 46%, belonging primarily to Gammaproteobacteria and Firmicutes groups (Acinas et al., 2004; Větrovský and Baldrian, 2013). As shown in (Table 3.7)
most of the bacterial groups associated with COPD range from 4 to 7 16S copies and generally belong to the above two phyla. On this basis, an average 16S copy number /genome of five was chosen to be representative of mixed microbial populations within COPD sputum samples in the study.

Table 3.7: 16S rRNA gene copy number per genome in bacteria associated with COPD

<table>
<thead>
<tr>
<th>Respiratory organisms</th>
<th>16S rRNA gene Copy number</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Streptococcus pneumoniae</em></td>
<td>4</td>
</tr>
<tr>
<td><em>Staphylococcus aureus</em></td>
<td>5 to 6</td>
</tr>
<tr>
<td><em>Haemophilus influenzae</em></td>
<td>6</td>
</tr>
<tr>
<td><em>Moraxella catarrhalis</em></td>
<td>4</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>7</td>
</tr>
<tr>
<td><em>Pseudomonas aeruginosa</em></td>
<td>4</td>
</tr>
<tr>
<td><em>Legionella pneumophila</em></td>
<td>3</td>
</tr>
<tr>
<td><em>Chlamydia pneumoniae</em></td>
<td>1</td>
</tr>
<tr>
<td><em>Mycoplasma pneumoniae</em></td>
<td>1</td>
</tr>
<tr>
<td><em>Bacteroides fragilis</em></td>
<td>6</td>
</tr>
<tr>
<td><em>Fusobacterium nucleatum</em></td>
<td>5</td>
</tr>
</tbody>
</table>

The NTC tubes produced readings in the range of $10^2$ to $10^3$ in the 16S qPCR assay even after UV crosslinking the PCR mix. This was considered to be caused by DNA contamination present in the taq polymerase, with no easy method for its elimination (Hughes et al., 1994). A threshold of < 1% contamination, contributed by the NTC readings was therefore set to sample qPCR readings. COPD samples where NTC readings contributed more than the threshold percentage had repeat assays. 16S qPCR reads of all samples analysed had <1% contamination from NTC.
3.3.1.2.3 Comparison between SYBR Green and TaqMan assays for 16S DNA quantification

Two different qPCR assay chemistries were chosen for enumerating the TBL. Although the TaqMan assay is known to be more specific than the SYBR Green assay, two limitations were identified in this study. Firstly, the TaqMan primers used produced amplicons of 500bp size that are considered suboptimal for qPCR assays (Dorak, 2006). In contrast, SYBR Green primers gave amplicons of 200bp size. Secondly, Taqman primers and probes didn't include any degenerate bases in the primer or probe sequence.

Comparative analysis was performed on the 16S qPCR data obtained from the pilot study of 58 samples. There was a strong correlation between the data obtained using the two qPCR assay ($R^2=0.90$, $p<0.0001$) (Figure 3.1). However, genome/ml values obtained with the TaqMan assay were consistently and significantly lower than corresponding values with SYBR Green (log mean ± SEM, 8.14±0.43 vs 8.44±0.43; $p=0.002$) (Figure 3.2).

**Figure 3.1**

Correlation between TaqMan and SYBR Green 16S qPCR: Scatter plot of 16S rDNA SYBR Green qPCR and TaqMan qPCR readings from 58 samples. SYBR Green readings represented on y axis) and TaqMan readings on x axis. Linear regression identifies a tight correlation that is highly significant.
Changes in the total and specific bacterial load at stable and exacerbation state of COPD

**Figure 3.2**

**Bland Altman plot of agreement between SYBR Green and TaqMan 16S assay:** Y axis displays the difference between SYBR Green and TaqMan counts and X axis represents the mean of the two measurements when performed on the same sputum sample. Negative values indicate a higher count with SYBR. For all but one sample, counts with SYBR Green were greater than the corresponding count with TaqMan. The red dashed line represents mean difference and the red dotted lines represent the 95% C.I.

The observed difference between the assays was investigated further for differences in oligo sets utilised. Both forward and reverse primers used in the SYBR Green assay shared overlapping 16S gene regions with TaqMan forward primer and probe sequence respectively (Figure 3.3). SYBR Green primers had degenerate bases whereas for TaqMan none of the oligos used had degenerate bases. The former therefore probably allows a wider range of bacteria to be detected and quantified and this appeared to be relevant in the quantification of bacterial load in sputum. The 16S SYBR green assay was therefore selected as our standard for quantifying sputum total bacterial load in the study.
3.3.1.2.4 Optimisation and assay chemistry of SBL qPCRs

For enumeration of the individual COPD pathogens mentioned in (3.1), genes specific to each of the bacteria being quantified were targeted as 16S gene primers were non-specific at genus and species level (data not shown). *H. influenzae* qPCR (HI qPCR) and *S. aureus* qPCR (SA qPCR) were SYBR Green assays producing 296bp and 280bp amplicons respectively. *S. pneumoniae* qPCR (SP qPCR) and *M. catarrhalis* (MC qPCR) were quantified using TaqMan qPCR chemistry targeting the pneumolysin and copB genes respectively, each producing ~70bp products.

The detection limit for all four SBL was >10^2 copies. NTC copies for TaqMan based assays were mostly not detected (ND) or <10 copies and for SYBR Green based SBL ranged from most being ND to <100 copies but with non-specific melt temperature to the target amplicon. None of the four SBL primers cross reacted with the other bacteria tested except for SA qPCR and MC qPCR producing signals with *M. catarrhalis* (3-logfold lower) and *H. influenzae* (4-logfold lower) respectively (Table 3.8). This is likely to represent a contamination as it would otherwise be expected that all samples with high abundance of *H. influenzae* (>10^8) and *M. catarrhalis* (>10^7) would also be positive for *M. catarrhalis* and *S. aureus*, respectively. However, 45 sputum samples with >10^8 genome/ml of *H. influenzae* were negative for detection of *M. catarrhalis.*
Similarly, 98 sputum samples produced a negative result with SAqPCR but had >10^7 genome/ml *M. catarrhalis* detected with MCqPCR.

### Table 3.8

<table>
<thead>
<tr>
<th>Strains tested</th>
<th>Log genome/ml readings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>16S qPCR</td>
</tr>
<tr>
<td><em>H. influenzae</em></td>
<td>9.31</td>
</tr>
<tr>
<td><em>S. aureus</em></td>
<td>8.86</td>
</tr>
<tr>
<td><em>S. pneumoniae</em></td>
<td>10.10</td>
</tr>
<tr>
<td><em>M. catarrhalis</em></td>
<td>9.48</td>
</tr>
<tr>
<td><em>P. aeruginosa</em></td>
<td>9.73</td>
</tr>
<tr>
<td><em>B. fragilis</em></td>
<td>9.33</td>
</tr>
</tbody>
</table>

**Specificity of individual specific bacterial load qPCR:** The following qPCR assays 16S for enumerating TBL, HI for *H. influenzae*, SA for *S. aureus*, SP for *S. pneumoniae* and MC for *M. catarrhalis* readings are presented in columns against the bacterial groups tested. ND (not detected) had no readings or readings below 10^2 copies. In 16S rDNA qPCR readings of each bacterial group are corrected for multiple copies.

### 3.3.1.2.5 Correction of high baseline signals in SYBR Green assays

In the biomarker COPD study, 1μl and 5μl DNA template were used for TBL qPCR and SBL qPCR, respectively. For SYBR Green based assays, some samples produced high baseline signals at lag phase of the PCR amplification cycle leading to inaccurate estimation of TBL, *H. influenzae* and *S. aureus* in these samples. Each of the sputum samples showing high baseline signals showed this phenomenon with all three SYBR assays, but not with TaqMan assays (Figure 3.4). It was hypothesized that this was due to the presence of high amounts of host contaminating DNA extracted in addition to the bacterial DNA from sputum, as fluorescence with SYBR Green qPCR is non-specific, binding to any double stranded (ds) DNA. One possible way of eliminating these signals was pre-treating the samples before bacterial DNA extraction to remove human
Changes in the total and specific bacterial load at stable and exacerbation state of COPD DNA. To test the hypothesis, MolYsis kit was used for removal of contaminating human DNA in blood samples. The kit consists of a chaotrophic solution to release DNA specifically from eukaryotic cells, followed by a DNAase treatment to remove the extracellular DNA before centrifuging down the bacterial pellet in the sample. 9 samples producing high baseline signals were treated with MolYsis human DNA removal enzymes and then bacterial DNA was extracted using Qiagen method. Although MolYsis treated samples eliminated the high baseline signals, there was also a significant loss ($p=0.0032$) in TBL compared to non-treated MolYsis sample. This was probably due to cell free DNA arising from weakening of bacterial cell wall with MolYsis treatment in frozen samples.

**Figure 3.4**

**SYBR green qPCR producing high baseline signals:** Raw qPCR amplification signals for two samples from all 5 assays. In the SYBR green assays these samples during lag phase produced signals higher than rest of the samples. This is not seen in TaqMan assay.

High baseline samples showed up to a 2.5-log fold difference in quantification reading between two different background fluorescence normalisation methods, Dynamic tube
normalisation (DTN) and Noise slope correction (NSC) (Figure 3.5). Background fluorescence with DTN is determined by taking an average from cycle 1 up to the starting amplification cycle number of each sample. In contrast, NSC uses a line-of-best-fit for normalisation. For high baseline samples DTN overestimated the quantification reading and in some cases produced false positive results while NSC underestimated the readings compared to readings produced by 1:10 dilution of the same sample. When SYBR Green assays were performed with 1:10 diluted samples the background high baseline signals disappeared and there was ≤ 0.1 log fold difference between the readings obtained from the two normalisation methods described (Figure 3.5). To maintain a balance between sensitivity of the qPCR along with reducing the number of samples producing high baseline signal, 1μl template DNA volume was used for all five qPCR assays. Samples (~80) still producing high baseline signal were re-assayed with 1:10 dilution.

**Figure 3.5**

Difference in quantification readings for high baseline samples with different normalization methods used: Scatter plot shows the log_{10} (genome /ml) readings of *H. influenzae* using DTN and NSC normalisation methods for samples producing high baseline signal with 5μl neat DNA and corresponding readings when 5 μl from 1:10 diluted of these samples were used. The figure shows near complete agreement between normalisation methods for diluted samples (overlapping crosses and triangles) but wide variability for neat samples (squares and diamonds)
3.3.2 Comparison between culture and qPCR for changes in TBL and SBL in stable and exacerbation visit

The longitudinal design of the COPD biomarker study provided an opportunity to examine bacterial populations in stable COPD over time and to identify changes in bacterial load and composition occurring during exacerbations. In this study, samples were also tested in parallel using bacterial culture techniques for pathogen and total bacterial load, allowing direct comparison of culture and molecular techniques.

3.3.2.1 Detection rate of individual bacteria tested by qPCR and its concordance to culture result

Culture and qPCR were performed on 662 sputum samples from 145 COPD patients at one or more time-points. Cross-sectional analysis of all available samples was performed to compare the two techniques for pathogen detection. In both methods *H. influenzae* was the most frequently isolated bacteria followed by *M. catarrhalis, S. pneumoniae* and *S. aureus*. Quantitative PCR was significantly more sensitive than routine culture method for detecting the specific respiratory pathogens. In culture 21% of the samples tested positive for at least one of these four pathogens and 2% of the samples had two of these pathogens simultaneously, while by qPCR 86%, 54% and 23% of the samples were positive for at least one, two and three pathogens, respectively. The detection rate ratio defined as the ratio of the detection rate with qPCR: detection rate with culture was between 5 and 9 for the four specified pathogens (Table 3.9). The ratio was comparable for *M. catarrhalis, S. pneumoniae* and *S. aureus* but lower for *H. influenzae*, primarily due to a higher rate of positive culture for this pathogen. Although qPCR detects specific organisms more frequently, the pattern of detection for different organisms is similar for the two techniques (Figure 3.6). Only four samples were identified that were positive for a pathogen with culture but not detected with qPCR (Table 1.9). One possibility for this finding would be the presence of inhibitors within these samples that interfered with the qPCR assay. Therefore, these samples were examined further to explore this possibility.
Table 3.9

<table>
<thead>
<tr>
<th></th>
<th>HI</th>
<th>MC</th>
<th>SP</th>
<th>SA</th>
</tr>
</thead>
<tbody>
<tr>
<td>qPCR positive</td>
<td>443</td>
<td>334</td>
<td>254</td>
<td>63</td>
</tr>
<tr>
<td>culture positive</td>
<td>89</td>
<td>38</td>
<td>28</td>
<td>7</td>
</tr>
<tr>
<td>Detection rate ratio</td>
<td>5.0</td>
<td>8.8</td>
<td>9.1</td>
<td>9.0</td>
</tr>
</tbody>
</table>

**Numerical summary of pathogen positive samples detected by both methods:** summarizes the number of COPD sputum samples that were associated with the detection of specific organisms with qPCR and culture. HI – *H. influenzae*; MC – *M. cattarhalis*; SP – *S. pneumoniae*; SA – *S. aureus*. Detection rate ratio = number of samples qPCR positive: number of samples culture positive.

Figure 3.6

**Detection rate of pathogens and comparison of the detection pattern between culture and qPCR:** Proportion (%) of COPD sputum samples positive for each of the pathogens tested with qPCR and culture. HI – *H. influenzae*; MC – *M. cattarhalis*; SP – *S. pneumoniae*; SA – *S. aureus*. Although qPCR was associated with more frequent detection, the pattern of detection of different organisms was similar using both techniques.

To test the hypothesis of inhibition of the qPCR assay in the 4 samples that were culture positive but qPCR negative, samples were spiked with a fixed amount of *M. tuberculosis* (Mtbc). Both SYBR Green and TaqMan assays were performed on each
Changes in the total and specific bacterial load at stable and exacerbation state of COPD

sample employing primer and probes specific for targeting the Mtb gene (Cheah et al., 2010). Water spiked with the Mtb DNA was used as a positive control.

For all the samples the Ct value of Mtb DNA in the qPCR assays showed <1 Ct difference with the MTB positive control indicating no inhibition in amplification of target DNA (Table 3.10). These results did not support the idea of inhibition as a factor in the observed absence of pathogen detection using qPCR.

**Table 3.10**

<table>
<thead>
<tr>
<th>Sample type</th>
<th>Sample id</th>
<th>SYBR</th>
<th>TAQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mtb positive control</td>
<td>Mtb PC</td>
<td>20.49</td>
<td>22.48</td>
</tr>
<tr>
<td><em>H. influenzae</em></td>
<td>B46</td>
<td>19.89</td>
<td>22.62</td>
</tr>
<tr>
<td>qPCR-ve/culture+ve</td>
<td>D31</td>
<td>20.74</td>
<td>22.69</td>
</tr>
<tr>
<td><em>S. pneumoniae</em></td>
<td>I6</td>
<td>20.49</td>
<td>22.71</td>
</tr>
<tr>
<td>qPCR-ve/culture+ve</td>
<td>AA19</td>
<td>20.02</td>
<td>22.21</td>
</tr>
<tr>
<td>control samples</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(qPCR +ve/ culture +ve)</td>
<td>B58</td>
<td>19.69</td>
<td>22.16</td>
</tr>
<tr>
<td></td>
<td>D50</td>
<td>20.20</td>
<td>22.31</td>
</tr>
</tbody>
</table>

DNA inhibition for samples identified to be qPCR negative but culture positive: SYBR (SYBR green qPCR) and TAQ (TaqMan qPCR) columns show the Ct values obtained for the positive control (Mtb PC) and tested samples for both SYBR and TAQ assays at 1μl template volumes. For all the tested samples, Ct values obtained were within ± 0.5 of the Ct value with positive control. This suggests that inhibition is not occurring.

### 3.3.2.2 Comparing qPCR counts of SBL between culture positive and culture negative samples

To explore samples that were positive for pathogen with qPCR but negative in culture, quantitative readings of culture positive and culture negative samples for individual pathogens was compared. Culture positive samples had significantly higher qPCR counts than culture negative samples for each of the individual pathogens (p=0.0003) for *M. catarrhalis* and (p <0.0001), *H. influenzae*, *S. pneumoniae* and *S. aureus* (Figure 3.7). Although the qPCR positive, culture negative group had approximately a 3 log10
fold lower median genome/ml count compared with the culture positive group, there
was considerable overlap in quantitative readings between the two groups. To
investigate if qPCR thresholds for predicting a positive culture could be determined,
receiver operated characteristics (ROC) curve analysis was performed on qPCR
readings between culture positive and negative samples for stable and exacerbation visit
in each of the 4 bacterial groups (Table 3.11). All four specific pathogens showed >
70% sensitivity and > 80 % specificity for being culture positive using a qPCR
threshold between $10^6$ and $10^7$ genome/ml.

**Figure 3.7**

Comparison of qPCR counts between culture positive and culture negative samples for individual
pathogen: Y axis represents the the qPCR count (genome/ml) in logarithmic scale and X axis represents
the samples of the 2 groups of samples that were qPCR + culture positive(+) and qPCR + culture
negative(-) for each of the pathogens A - *H. influenzae*, B- *S. pneumoniae*, C- *M. catarrhalis*, D- *S. aureus*
.Culture(+) had 88, 25, 37 and 7; Culture(-) had 355, 229, 297,56 samples in A, B, C, D respectively. The
red line represents the median and the values are given for each group in the textbox and the t test (Mann-
Whitney) values.
Table 3.11

<table>
<thead>
<tr>
<th>Bacteria</th>
<th>Area under ROC curve</th>
<th>Threshold for culture positive (genome/ml)</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Number of samples</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>H. influenzae</em></td>
<td>0.926</td>
<td>$\geq 6.82$</td>
<td>95</td>
<td>85</td>
<td>310</td>
</tr>
<tr>
<td><em>S. pneumoniae</em></td>
<td>0.814</td>
<td>$\geq 6.52$</td>
<td>72</td>
<td>84</td>
<td>188</td>
</tr>
<tr>
<td><em>M. catarrhalis</em></td>
<td>0.944</td>
<td>$\geq 7.29$</td>
<td>100</td>
<td>83</td>
<td>253</td>
</tr>
<tr>
<td><em>S. aureus</em></td>
<td>0.925</td>
<td>$\geq 6.2$</td>
<td>100</td>
<td>68</td>
<td>52</td>
</tr>
</tbody>
</table>

ROC curve table based on quantitative readings for individual pathogen between culture positive and culture negative samples: The lowest $\log_{10}$ genome/ml qPCR readings for each pathogen that has optimum sensitivity and specificity for predicting a culture positive sample.

3.3.2.3 Change in the TBL of COPD samples between stable and exacerbation states

Total bacterial load was measured using both culture (C.F.U) and qPCR techniques in patients that provided samples at both stable (n= 315) and exacerbation (n= 116) states. Quantitative PCR was associated with a significantly higher 1.5 to 2 log-fold higher measured TBL than culture (p < 0.0001). There was no difference in the TBL between stable and exacerbation visits with culture (p=0.585) or qPCR (p=0.0882) at population level. Furthermore, within-subject paired analysis identified no difference in TBL between stable and exacerbation states in 56 longitudinal samples with qPCR (p=0.37) or culture (p=0.44) (Figure 3.8).
Figure 3.8

The box whisker plots represent the mean (line inside the box), standard deviation (box) and minimum and maximum counts (bars) of total bacterial load in the groups shown. Y axis represents in log$_{10}$ scale C.F.U/ml for culture and genome/ml reading for qPCR.

3.3.2.4 Comparison of detection rate and quantitative readings of individual SBL between stable and exacerbation state

To investigate whether there were detectable differences between stable and exacerbation samples for the commonly associated COPD pathogen both the detection rate of the four measured pathogens and changes in their quantitative readings were compared. In the 65 paired stable and exacerbation samples available for analysis on specific pathogens, *H. influenzae* was the most frequently observed bacteria in both states followed by *S. pneumoniae* and *M. catarrhalis*. Using qPCR, all the pathogens were mostly present at both stable and exacerbation state, with higher detection rate of the pathogens at stable state. Using culture, *S. pneumoniae* and *M. catarrhalis* showed a higher rate of detection in exacerbation state (Table 3.12). Analysis of quantitative data identified significantly higher qPCR counts at stable state compared with exacerbation.
Changes in the total and specific bacterial load at stable and exacerbation state of COPD

state for *S. aureus* in paired samples (p=0.046) (Figure 3.9). Although *M. catarrhalis* at population level showed significantly higher bacterial load in exacerbation (p= 0.0076), this was not seen in matched stable exacerbation samples.

**Table 3.12: Detection rate of pathogens in paired stable and exacerbation samples by qPCR and culture**

<table>
<thead>
<tr>
<th>Method of detection</th>
<th>Bacterial groups</th>
<th>Total number of positives</th>
<th>% Exacerbation positive only</th>
<th>% Stable positive only</th>
<th>% positive in both</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>qPCR</strong></td>
<td><em>H. influenzae</em></td>
<td>58</td>
<td>18.97</td>
<td>24.14</td>
<td>56.90</td>
</tr>
<tr>
<td></td>
<td><em>S. pneumoniae</em></td>
<td>44</td>
<td>18.18</td>
<td>34.09</td>
<td>47.73</td>
</tr>
<tr>
<td></td>
<td><em>M. catarrhalis</em></td>
<td>48</td>
<td>25.00</td>
<td>27.08</td>
<td>47.92</td>
</tr>
<tr>
<td></td>
<td><em>S. aureus</em></td>
<td>13</td>
<td>23.08</td>
<td>61.54</td>
<td>15.38</td>
</tr>
<tr>
<td><strong>Culture</strong></td>
<td><em>H. influenzae</em></td>
<td>12</td>
<td>25.00</td>
<td>41.67</td>
<td>33.33</td>
</tr>
<tr>
<td></td>
<td><em>S. pneumoniae</em></td>
<td>8</td>
<td>50.00</td>
<td>25.00</td>
<td>25.00</td>
</tr>
<tr>
<td></td>
<td><em>M. catarrhalis</em></td>
<td>11</td>
<td>63.64</td>
<td>18.18</td>
<td>18.18</td>
</tr>
<tr>
<td></td>
<td><em>S. aureus</em></td>
<td>3</td>
<td>0.00</td>
<td>66.67</td>
<td>33.33</td>
</tr>
</tbody>
</table>
Comparison of the qPCR counts of pathogens between longitudinal stable and exacerbation samples: compares the qPCR counts between paired stable and exacerbation samples for A: *H. influenzae*, B: *S. pneumoniae*, C: *M. catarrhalis*, D: *S. aureus*. Y axis represents the qPCR counts in logarithmic scale (log\(_{10}\)) and x axis the samples belonging to the two groups stable and exacerbation. Samples with bacteria below detection limit were given a constant of log 2. The red line represents the median and the values are given for each group in the textbox.

### 3.4 Discussion

#### 3.4.1 Developing and optimizing qPCR assays for quantifying TBL and for SBL in COPD sputum

#### 3.4.1.1 DNA Extraction Methods

Several methods exist for extracting bacterial DNA from samples. The objective was to choose an extraction method that would:
1. Allow isolation of DNA from most bacterial cell types, thus representing closely the entire bacterial population in sputum samples.

2. Be quick and convenient for routine extraction of a large number of samples

3. Enable storage of the extracted DNA so that it may be utilised downstream to various molecular methods.

In this study, two DNA extraction methods (QIAmp and microLYSIS) were evaluated for purpose in a pilot sample set. Important differences between these kits were identified that informed the choice of extraction method for the study set.

Among the various column based DNA extraction commercial kits available, the QIAamp kit is extensively used for extraction of DNA from biological samples (Gao et al., 2007; Hendolin et al., 2000; Ravel et al., 2011). These kits involve enzymatic lysis of cells to release DNA and purification of column bound DNA from proteins, and other cell products, with the advantages of being quicker, having less contamination risk, ease of automation and avoiding organic PCR inhibitors from conventional phenol-chloroform extraction method (Tan and Yiap, 2009).

MicroLYSIS is a one tube extraction method that has several potential advantages over the Qiagen method including being quicker, technically easier, with less contamination risk. It has also been used effectively for isolating DNA from tough bacterial cells such as M. tuberculosis in our lab (Cheah, May 2010). However, in this study, microLYSIS failed to extract DNA from S. aureus culture on several occasions. Interestingly, the 1:1 culture mix of E. coli and S. aureus extracted by microLYSIS didn't yield any qPCR readings for the latter. In contrast, Qiagen DNA extraction was reliable, giving similar readings to C.F.U counts for both organisms. One possible explanation for not obtaining reads with microLYSIS extracted samples is the release of PCR inhibitory substances during S. aureus cell lysis. The method does not include a purification step following DNA extraction, leading to bacterial proteins and cell debris released during cell lysis and inhibitors present in the clinical sample interfering with the outcome of quantification assays as well as making extracted samples unsuitable for storage (Wilson, 1997). Using a smaller sample volume compared to Qiagen implied a reduction in sensitivity, and with >10% of the sample read being contributed by NTC, there was consequent loss of specificity with the use of this extraction method. The
significant limitations with microLYSIS directed the decision to use QIAmp as the DNA extraction method of choice.

### 3.4.1.2 Development and Optimisation of TBL and SBL qPCR

Although molecular based enumeration of total bacterial load is much more sensitive than the routine plate technique, there are potential drawbacks. In particular, careful consideration of technical issues related to the development and optimization of quantitative assays is needed and results should be interpreted taking these into account. As commercially optimized qPCR mixes were utilized for all assay chemistries, only the primer concentrations and annealing temperature were optimized to reduce non-specific reaction and formation of primer dimers, without losing sensitivity.

Due to the variable 16S rDNA copy numbers among bacterial groups in a mixed microbial sample, quantified 16S copies would not be equivalent to the number of genomes present (Crosby and Criddle, 2003). This issue was managed by using a correction factor based on an estimated average of 16S rDNA copies contributed by bacterial groups commonly identified in COPD sputum. At a population level this is likely to be reasonable but for individual samples dominated by one or a few bacterial groups with 16S rDNA copy numbers significantly different to the correction factor, misleading overestimates or underestimates of genome readings may arise. In addition, biases are introduced with replicating bacteria having more than one copy of certain chromosomal portions leading to more 16S rDNA copies compared to an equivalent non-replicating cell (Klappenbach et al., 2001; Nadkarni et al., 2002).

A number of issues were also identified with the choice of reagent for qPCR analysis. There were advantages and limitations with both reagents, SYBR Green and TaqMan, evaluated in this study.

TaqMan assays are highly specific due to the specific binding of the probe to the template and a single fluorescent molecule binding to each amplicon. However, significant limitations compared to SYBR Green were identified. Longer amplicon product and reduced coverage of the TaqMan probe compared to SYBR Green 16S rDNA primers are factors that are likely to have contributed to lower qPCR readings.
with the TaqMan assay, suggesting loss of readings from bacterial population present.

In contrast for SYBR Green the use of universal primers and non-specific binding of SYBR Green to dsDNA meant detection of qPCR readings close to DNA standard concentrations of NTC (Bottger, 1990; Rand and Houck, 1990). This decreased the sensitivity of the TBL assay which was overcome by use of higher sample volumes for DNA extraction. All samples produced > 10^5 copies/μl 16S rDNA readings with < 1% readings contributed from NTC readings. Binding of multiple molecules of SYBR Green to a single amplicon can lead to bias introduced by the length and the GC content of the amplicon (Giglio et al., 2003). However, this is unlikely to be significant as 16S rDNA GC content varies only slightly even among diverse bacterial groups (Belozersky and Spirin, 1958; Muto and Osawa, 1987) compared to the wide variation, from 25% to 75%, in the total GC content (Sueoka, 1962).

As discussed in section 3.3.1.2.5 SYBR Green assays produced high baseline signals with few samples during the lag phase of PCR cycling leading to erroneous quantification readings. Furthermore, on performing Mtb spiked inhibition assay described in Table 3.10 on these samples showed inhibition in the PCR reactions. The most likely reason was the co-purification of human DNA during extraction and binding of SYBR Green to this. However, pretreatment methods to eliminate eukaryotic DNA lead to loss of bacterial DNA from frozen samples with fragile bacterial cells such as H. influenzae and S. pneumoniae. Performing qPCR assays with diluted and lower DNA template volumes eliminated the appearance of the background high baseline signals. Dilution or decreasing the sample DNA volume in PCR reaction to dilute out PCR inhibitory and interfering substances has been suggested and used in other qPCR assays(Bach et al., 2002; Wilson, 1997).

3.4.2 Comparison between culture and qPCR for changes in TBL and SBL in stable and exacerbation visit

3.4.2.1 Detection rate of individual bacteria tested by qPCR and its concordance to culture result

Both culture and qPCR showed a similar pattern of detection of the tested pathogens with H. influenzae being the most frequently isolated bacteria followed by S. pneumoniae and M. catarrhalis in all visit samples. This was in concordance with
earlier studies that also found these bacteria to be the most prevalent at stable and exacerbation states (Monsó et al., 1999; Sethi and Murphy, 2001). *H. influenzae* was also found to be the most prevalent bacteria by another qPCR study (Garcha et al., 2012). *P. aeruginosa*, also considered as a frequent pathogen in COPD was isolated in very few samples by culture and wasn't tested by qPCR.

Although the pattern of expression of the different bacterial groups was broadly similar, the proportion of samples positive with qPCR was significantly higher for each of the individual pathogen. Culturing is performed at 10⁵ dilution of samples (Health Protection Agency, 2009) while qPCR could detect upto 10³ genome/ml. ROC analysis to determine a threshold qPCR reading that was predictive for a positive culture was comparable between the different specific bacteria, ranging from 10⁶ – 10⁷ genome/ml. Overall, ~ 75% of culture negative qPCR positive for a given pathogen were contributed by < 10⁶ genome/ml.

A proportion of samples with >10⁶ qPCR counts were culture negative. This may reflect a contribution from dead bacteria, bacteria in non-culturable state or non-specific amplification by PCR primers. Post et.al group working on otitis media showed that samples spiked with heat-killed cells or purified DNA did not detect any DNA from these cells by day 3 (Post et al., 1996). This suggests the rapid deterioration of cell free DNA and of DNA in compromised cells. The DNA extraction method used in this study involved pelleting down the sample and discarding the supernatant which should have also got rid of any extracellular DNA present. It is therefore unlikely that dead cells were a major contributing factor. Various groups have reported failure to detect *H. influenzae* in sputum cultures despite detection by molecular technique (Groeneveld et al., 1990; Murphy et al., 2004). In a study of 93 effusion samples collected from subjects with otitis media with effusion a significant proportion of *H. influenzae* culture negative samples showed positive results with PCR and mRNA reverse transcriptase-PCR (RT-PCR) assays suggesting the presence of metabolically active but non-culturable cells (Rayner et al., 1998). This might be due to reversion of bacterial cells to forms that are not easily culturable (Roberts et al., 1984), formation of biofilm (Murphy and Kirkham, 2002; Post, 2001) or intracellular localisation of bacteria in host cells (Forsgren et al., 1996; Morey et al., 2011).
In contrast to this, four discordant results were identified that were culture positive but qPCR negative. No evidence for PCR inhibition was identified in any of these four samples. Three of the samples were S. pneumoniae culture positive but failed to detect it in qPCR. S. pneumoniae species and other viridans group bacteria such as Streptococcus mitis and Streptococcus oralis are known for intergenomic variations due to natural genetic recombination events (Hakenbeck et al., 2001). It is therefore possible that such recombination events might have altered the pneumolysin gene nucleotide sequence that led to mismatches with the S. pneumoniae targeting primers used in this study producing false negative results.

As mentioned earlier, ROC analysis to evaluate the relationship between qPCR reads and culture outcome was very similar between the different bacteria in respect of qPCR read thresholds predictive of a positive culture result. However, it is notable that the relationship between qPCR and culture outcomes was poorer for S. pneumoniae, with lower values obtained for both sensitivity and AUC (Table 3.11). A possible explanation for this could be that S. pneumoniae is a viridans group Streptococcus (VGS) and is closely related to S. mitis and S. oralis of the VGS group (Whatmore et al., 2000; Whiley and Beighton, 1998). These VGS group member constitute the normal oral flora (Mager et al., 2003; Whiley and Beighton, 1998) and sputum samples can be easily contaminated with oral bacteria. Due to natural transformation capability there is widespread exchange of genetic information, including pneumococcal virulence genes among VGS group members (Cvitkovitch, 2001; Jefferies et al., 2007).

This has led to considerable debate over the choice of primers and gene targets to distinguish S. pneumoniae from other VGS group members (Johnston et al., 2010; Kawamura et al., 1999). Although the primers used were highly specific for S. pneumoniae (Greiner et al., 2001), extensive homology of pneumolysin gene with S. mitis and S. pseudopneumoniae (atypical pneumonia) might have led to non-pneumococcal amplification products (Johnston et al., 2010; Whatmore et al., 2000).

### 3.4.2.2 Change in TBL and SBL between stable and exacerbation samples

Both 16S qPCR and plate counts showed no significant difference in bacterial load between stable and exacerbation samples either at population level or in paired samples. Also the commonly associated COPD pathogen tested H. influenzae, S. pneumoniae and M. catarrhalis didn't show difference in number between stable and exacerbation visit.
Furthermore, *H. influenzae, S. pneumoniae* and *M. catarrhalis* were all detected more frequently with qPCR in stable samples compared to exacerbation samples. These observations do not support the hypothesis that exacerbations are associated with a consistent and detectable rise in bacterial load.

Isolation or increase in loads of *H. influenzae* and *S. pneumoniae* has been associated with exacerbation in various culture studies (Bandi et al., 2003; Gump et al., 1976; Wilkinson et al., 2006). A recent qPCR based COPD study by Garcha et al showed higher prevalence rate of *S. pneumoniae* and *M. catarrhalis* at exacerbation compared to stable state in matched samples (Garcha et al., 2012). Garcha and co-workers also showed an increase in the bacterial load of these pathogens at exacerbation state compared to their load in stable state (Garcha et al., 2012). There are likely to be a number of reasons for the observed differences in our study. Firstly, the cohort of COPD subjects in this study were more severe with lower baseline lung function (Bafadhel et al., 2011) than those of the study by Garcha et al (2012) . Increased severity is associated with chronic colonisation by common pathogens at stable state (Wilkinson et al., 2003) and this was frequently observed in the sample set.

Sethi et al showed an association between strain change in *H. influenzae, S. pneumoniae* and *M. catarrhalis* with exacerbation but no increase in bacterial concentrations or prevalence rate of these bacteria at exacerbation (Sethi et al., 2002; Sethi et al., 2007). Clinical DNA sample apart from the possibility of generating non-specific amplicon products from other bacteria might also contain co-colonising strains of a bacterial species which can produce overlapping profiles. Therefore, molecular typing is normally performed on DNA from pure culture isolates, as presence of which are not routinely maintained in our lab.

Finally, it is also likely that a significant proportion of COPD exacerbations are not due to bacterial infection. Heterogeneity in the mechanism of COPD exacerbation will also reduce the possibility of detecting a significant change between visits.

The frequent identification of pathogens in stable COPD sputum supports the view that these may be involved in COPD progression. However, this finding is in contrast with study that found no significant difference between healthy subjects and COPD in 16S assay based TBL (Erb-Downward et al., 2011). However, TBL (16S) was significantly higher in samples that were either culture positive or had $>10^7$ cfu/ml detected.
(Bafadhel et al., 2011) suggesting the role of commonly tested pathogens in pathogenicity in COPD.

### 3.5 Conclusion

In previous studies use of culture methods and lack of longitudinal samples have produced contradictory outcomes to support the theory of increased pathogenic bacterial load in exacerbation. QPCR methods developed in this study showed 98% agreement with the culture positive result and similar detection pattern for the pathogens tested. The sensitivity of qPCR was at least 2 log fold higher than routine culture methods, leading to a greater than 5 fold higher detection rate for all 4 specific organisms. All the SBL assays showed >70% sensitivity and >82% specificity for predicting a COPD culture positive samples. This indicates the efficacy of the qPCR assays and the detection of live bacteria in most cases. The higher sensitivity, quantification of pathogen and quicker turnaround time using qPCR methods are all valid reasons for considering its wider implementation in practice or at least for considering inclusion in addition to routine culture for routine bacteriological testing in COPD.

All COPD sputum samples tested had a considerable bacterial community (TBL >10^6 genome/ml) in their airways with a mean abundance > 10^5 genome/ml for each of the bacteria quantified. Among the bacteria tested *H. influenzae* was detected in over 50 % of the samples while *M. catarrhalis* and *S. pneumoniae* were identified in approximately 50% of the samples. These observations suggest *H. influenzae, M. catarrhalis* and *S. pneumoniae* are dominant constituents of COPD sputum samples.

This study showed no significant change in the total or specific bacterial load and no difference in specific pathogen detection rates between longitudinal stable and exacerbation samples using either qPCR or culture methods. These results suggest that an increase in bacterial load may not be required to implicate bacterial exacerbation. It is possible that disturbance in the equilibrium among bacterial constituents of the COPD microbial community might be associated with exacerbation episodes. This hypothesis is amenable to study using DNA sequencing techniques that characterise the entire microbiome.
4  16S rDNA based COPD bacterial community analysis

4.1  Introduction

In previous chapter sputum analysis using qPCR techniques identified no quantitative
difference in total bacterial load and specific respiratory pathogens within subjects
between stable and exacerbation states. This was concordant with the plate count
results. However, qPCR techniques are limited to the quantitative study of the bacteria
being tested, most commonly based on the information acquired from culture results. As
discussed previously, standard culturing techniques employed in clinical practice are
limited in sensitivity and the spectrum of bacteria cultured. It is quite plausible that
uncultivated bacteria such as ‘not yet cultured’ (NYC) bacteria, anaerobes or obligate
intracellular bacteria are associated with exacerbations. The potential importance of this
is illustrated by the discovery of *Helicobacter pylori*, an organism not previously
detected in standard laboratory practice due to longer growth times. The work of
Marshall and Warren associating *H. pylori* infection with chronic gastritis and peptic
ulcers revolutionised our understanding and management of these conditions (Marshall

The use of DNA fingerprinting and more recently high throughput 16S rDNA
sequencing techniques to study the microbiome of the gastrointestinal tract (GIT) have
found >50% of the bacterial community to be constituted of NYC bacterial populations
(Aas et al., 2005; Eckburg et al., 2005). These molecular profiling techniques are
beginning to reveal the existence of a complex and diverse array of bacterial
communities within various human organ systems and increasingly the importance of
interactions between these communities and the human host health and disease is being
understood (Dethlefsen et al., 2007; Foxman et al., 2008). In a healthy host, it is
hypothesised that a dynamic equilibrium is established within the microbiome
constituents that stops proliferation of potentially pathogenic organisms and disruption
in this balance can lead to a diseased state. For example, in the gut microbiome there is
a body of evidence supporting the changes in the relative abundance of Firmicutes and
Bacteroidetes to be a determinant of metabolic status and obesity (Ley et al., 2006;
Turnbaugh et al., 2009; Turnbaugh et al., 2006), while Saini and co-workers have shown that dental caries are associated with a shift in the oral microbiome from one dominated by anaerobic Gram positive cocci in health to anaerobic Gram positive bacilli (Saini et al., 1999). It is unclear why changes in the microbiome occur and the extent that these are cause or effect in different diseases. Changes in the microbiome in disease might be a result of synergistic, antagonist or additive interactions between microbes in a given environment that may be influenced by the presence of certain organisms (Brogden et al., 2005; Peters et al., 2012). Dental caries and bacterial vaginosis are polymicrobial diseases in which synergistic interactions between microbes for metabolite co-dependencies have been predicted (Palmer et al., 2001; Saito et al., 2008a; Srinivasan et al., 2012). Antagonistic interactions between S. pneumoniae and S. aureus and more recently between S. pneumoniae and H. influenzae have been reported in the pathogenesis of upper respiratory tract (URT) infections (Bogaert et al., 2004; Pettigrew et al., 2008), while additive interactions within the microbiome of the lower respiratory tract have been reported in cystic fibrosis (Rogers et al., 2003). Viruses have been suggested as predisposing factors in polybacterial infection of otitis media (Hament et al., 1999).

The URT is known to harbour a number of commensal bacteria, along with the occasional presence of S. pneumoniae and H. influenzae (Austrian, 1968; Beasley et al., 2012), while the lower respiratory tract (LRT) has been considered to be sterile until recently. However, recent molecular profiling studies have shown otherwise (Charlson et al., 2011; Erb-Downward et al., 2011; Hilty et al., 2010). Charlson and co-workers generated 16S rDNA sequence profiles from various URT (oropharyngeal and oralwash) and LRT (BAL samples) from healthy subjects and showed that the lung microbiome is not distinct from the URT and is colonised by similar bacterial groups like Veillonella, Prevotella and Streptococcus, though in lower abundance (Charlson et al., 2011). To date, microbiome studies in COPD have used bacterial community directed sequencing analysis utilising 16S rDNA based 454 sequencing, T-RFLP, phylochip and microarray techniques to compare the microbial community of healthy smokers and non-smokers with COPD patients; analyse differences between subgroups of COPD severity; and evaluate the microbiome during severe exacerbation episodes. A summary of these study designs and outcomes is presented in Table 4.1. These studies
have been largely limited by low numbers and cross-sectional design. No study has yet looked at longitudinal changes in the microbiome during exacerbation events.
Table 4.1: Overview of microbiome studies to date in subjects with COPD

<table>
<thead>
<tr>
<th>Reference</th>
<th>Outcomes</th>
<th>Study objectives</th>
<th>Study design</th>
<th>Study population</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Huang et al., 2010)</td>
<td>Identified a total of 3 phyla and 140 distinct families. With the increase in the number of incubation days there was a decrease in the richness of the bacterial community. The microbiota composition changed from Firmicutes and Bacteroidetes (Pseudomicrobiota) with increase in incubation days.</td>
<td>To determine if more diverse bacterial community is present at exacerbation and treated on antibiotics</td>
<td>16S rRNA genebased phylogenetic microarray analysis for determining bacterial diversity and relative abundance of these bacteria. The qPCR was used to validate the phylotype that could best determine the changes in the relative abundance of bacteria.</td>
<td>8 COPD exacerbation subjects admitted to ICU and were from a cohort of P. aeruginosa study in intubated patient, California.</td>
</tr>
<tr>
<td>(Hilty et al., 2010)</td>
<td>Analyzed outcomes from ~5000 16S sequences. They showed Prevotella species to be predominant in COPD and asthma. Nasal microbiota clustered together from the other 2 phenotypes and was most distant from the other 2 respiratory location samples. The OP and UL microbiota of COPD were clustered together.</td>
<td>To characterize the bacterial community between asthmatic COPD, and healthy controls</td>
<td>16S RNA clone library from pulmonary samples of nose and oropharynx (OP) and left upper lobe (LUL) gpcr to determine bacterial load.</td>
<td>5 COPD, 11 asthma and 8 healthy controls, France.</td>
</tr>
<tr>
<td>(Enn-Downward et al., 2011)</td>
<td>Healthy smokers, non-smokers and mild COPD tend to have a more diverse microbial community. Contrast to moderate and severe COPD microbiota. COPD is dominated by fewer bacterial groups. There was no significant difference in the microbiota between the three study groups. No change in bacterial load between 3 groups.</td>
<td>To explore the differences in the lung microbiomes of healthy smokers, non-smokers, and COPD</td>
<td>16S rDNA, 454 sequencing of V3 and V4 region. Taxonomic and phylogenetic based gpcr analysis.</td>
<td>7 healthy smokers, 3 never smokers and 4 COPD subjects, USA.</td>
</tr>
<tr>
<td>(Sei et al., 2012)</td>
<td>T-RFLP showed presence of distinct communities between COPD group with significant increase in Lactobacillus genus in severe COPD lungs. Similar microbial diversity indices were observed between the 3 groups.</td>
<td>To characterize the lung microbiome in severe and mild COPD</td>
<td>Lung tissue samples used for quantifying bacterial load using qPCR and T-RFLP (V1 to V3) to characterize the microbiome in the different groups.</td>
<td>8 non-smokers, 8 severe COPD and 8 cystic fibrosis positive control. Canada.</td>
</tr>
<tr>
<td>(Fragman et al., 2012)</td>
<td>Showed overlap in the microbial constituents between the 3 groups. Actinobacteria</td>
<td>To characterize the lung microbiome in severe and mild COPD</td>
<td>454 sequencing of BAL samples</td>
<td>22 COPD moderate and severe subjects and 10 controls.</td>
</tr>
<tr>
<td>healthy subject, America</td>
<td>lung microbiome</td>
<td>moderate COPD</td>
<td>was one of the dominant groups in COPD and contradictory to Erb-Downward et al. control samples were less heterogeneous compared to COPD microbiota.</td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>----------------</td>
<td>---------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>8 stable subjects with moderate disease and not had any exacerbation for a year preceding the study and no antibiotic treatment, Spain</td>
<td>454 sequencing (V1 to V2) of 4 samples from each individual: sputum, bronchial aspirate, BAL and bronchial mucosa</td>
<td>To identify lower-airway bacteria unrecognized through culture in COPD, in the absence of signs and symptoms of bronchial infection and to examine the complexity of microbial flora in these patients infection.</td>
<td>Streptococcus, Prevotella, Moraxella, Haemophilus, Acinetobacter, Fusobacterium, and Neisseria were the most commonly identified groups. Sputum samples showed significantly lower diversity than the other three sample types. (Cabrera-Rubio et al., 2012)</td>
<td></td>
</tr>
<tr>
<td>Bronchoscopy and BAL samples obtained from 9 COPD and 9 healthy subjects, Germany</td>
<td>T-RFLP and cloning</td>
<td>To examine the pulmonary microbial communities in COPD patients as compared to healthy individuals</td>
<td>Highly diverse microbial community was observed with bacteria from Prevotella, Sphingomonas, Pseudomonas, Acinetobacter, Fusobacterium.,(Zakharkina et al., 2013) Megasphaera, Veillonella, Staphylococcus, and Streptococcus constituting the major part of the core microbiome in both groups. Occurrence of Pseudomonas associated with reduction in microbial diversity. (Zakharkina et al., 2013)</td>
<td></td>
</tr>
<tr>
<td>10 severe to very severe and 9 moderate subject’s sputum samples</td>
<td>454 sequencing V1-V3</td>
<td>Compare the microbiota in two groups of patients with COPD of different severity in order to detect potential microbiological markers</td>
<td>The most abundant phylum found in the samples was Firmicutes, followed by Proteobacteria, Actinobacteria and Bacteroidetes. Alpha diversity indices was significantly higher in moderate compared to severe COPD. Prevalence of Actinomyces was significantly higher in moderate group. Microbial composition among moderate samples was more stable compared to microbial composition among severe COPD samples. (Galiana et al., 2013)</td>
<td></td>
</tr>
</tbody>
</table>

*ICU - intensive care unit, *ETA - endotracheal aspirates
16S rDNA based COPD bacterial community analysis

References cited in Table 4.1:


We hypothesized that differences between the COPD clinical states might also be related to changes in microbial community membership and structure that might not have been represented by enumeration of bacterial loads.

To achieve this 16S rDNA sequence based bacterial community profiles were generated using the high throughput 454 sequencing method and microbiome analysis was performed to investigate the following aims:

1. To characterise the microbial community in COPD patients at the four visit times S, E, F and R and explore any differences between visits focussing on changes related to exacerbation.

2. To examine the consistency of the microbiome in individuals over multiple stable time points.

3. To determine any association between the microbiome and pre-defined clinical phenotypes based on smoking status, GOLD stage severity and treatment received at exacerbation.

4. To determine if differences exist in stable COPD microbiome between exacerbated and non-exacerbated subjects.

4.2 Materials and Methods

4.2.1 454 sequencing samples
From the initial 662 samples collected at various time points from 145 subjects three sets of 454 samples were assembled. These are described in detail in (section 4.3.1.1). A total of 151 samples from 39 subjects were studied.

For Set 1 120 samples from 30 COPD subjects with each subject comprising of sequential stable (S), exacerbation (E), follow-up or 2 weeks post exacerbation (F) and recovery or 6 weeks post exacerbation (R) state samples were chosen. Henceforth these four COPD states will be referred to as visit types.
Set 2 included longitudinal stable samples from three stable time points for each of the 11 subjects with a total of 33 samples. One out of three stable samples for each subject was also present in above Set1.

Set 3 consisted of a single baseline sample from 9 non-exacerbated individuals in the study.

4.2.2 Production of 16S rDNA bacterial amplicon library

For 454 sequencing of the bacterial community in COPD samples one way/unidirectional reads amplicon sequencing method was chosen and Lib-L based method for amplicon library production was followed according to "Amplicon Library Preparation Method Manual, GS FLX Titanium Series (October 2009)". Following steps were involved in the production of 16S rDNA bacterial amplicon library:

4.2.2.1 Primer design and sequence

Primers for 454 sequencing called as fusion primers consisted of following four parts

Adaptor sequence - Both forward and reverse primer have an adaptor sequence starting at the 5' prime end of the primers called as Lib-L/A and Lib-L/B respectively. These sequences allow binding of single stranded (ss) DNA to Lib-L capture beads and subsequent annealing to emulsion PCR (emPCR) and sequencing primers.

Key sequence - This is a 4 base sequence "TCAG" present in both forward and reverse primer in the 5'-3' orientation downstream of the adaptor sequence. It is used as calibrator for the signal intensity produced during the sequencing flowgram cycle.

Multiplex identifier (MID) sequence - For each 454 run all the samples were tagged with a unique 10 bp sequence acting as a barcode for sample identifying in downstream sequence analysis. These unique MIDs are introduced in the primer sequence immediately downstream to the key sequence. Since the 454 run performed was unidirectional sequencing only the forward primer had a MID sequence. These MIDs were chosen from the Roche's 454 set of designed MIDs for Genome Sequencer FLX titanium series (Using Multiplex Identifier (MID) Adaptors for the GS FLX Titanium Chemistry - Extended MID Set, April 2009).

Template specific primer sequence - Primers targeting the eubacterial16S rDNA in COPD sputum samples were chosen based on the following criteria
i. The 16S oligonucleotides (oligo) were conserved amongst most eubacterial groups and at the same time non-specific to eukaryotic DNA.

ii. Aim was also to include as many 16S hypervariable region for bacterial community discrimination within the optimal size PCR product for 454 sequencing.

For all 454 sequencing universal 16S primers 926F (Muyzer et al., 1995) and 1391R (Lane et al., 1985) were utilised for amplification of hypervariable regions V6 to V8. The 16S oligo sequence are given below in Table 4.2

Table 4.2: 16S rDNA sequence utilised for building the 454 bacterial library

<table>
<thead>
<tr>
<th>16S oligo name</th>
<th>16S oligo sequence (5’- 3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>926F</td>
<td>AAACTCAAAGAATTGACGG</td>
</tr>
<tr>
<td>1391R</td>
<td>GACGGGCCTGGTGTRCA</td>
</tr>
</tbody>
</table>

In total 31 fusion primers consisting of 30 MID tagged 454 forward primers and a single 454 reverse primer were designed (Figure 4.1) and were sourced from Sigma-Aldrich (Dorset, UK).

Figure 4.1

454_926F (Forward primer) :
5’-CCATCTCATCCCTGCCGTGTCTCGAGACTCAG-MID-AAAACCAAKGAATTGACGG-3’

454_1391R (Reverse primer):  
5’-CTATCCCTCTGTGTGAGCTCTGACGGCGCGGGGTGTRCA-3’

454 fusion primer design and sequence: Adaptor sequence is represented in Blue, Key in red and 16S specific sequence in green. 30 forward primers were designed each with a unique decamer MID sequence.

4.2.2.2 PCR conditions, Gel run and purification

After optimisations of PCR components yielding PCR product from most samples and the different barcoded primers, final PCR reactions were done in 50μl volume containing 1x High-Fidelity PCR buffer, 0.8μM of each forward and reverse primer,
2.5mM MgCl₂, 0.25mM of each deoxyribonucleotide triphosphate (dNTP) (Promega, WI, USA), 0.8M Betaine HCl (Sigma-Aldrich, Dorset, UK), 0.08mg/ml BSA (NEB, UK) and 0.5μl(1U) Phusion High-Fidelity DNA Polymerase (Finnzymes, Finland). 1μl DNA template was used for PCR reaction. Each PCR batch had a negative control with DNA replaced by 1μl molecular grade water. PCR cycle was performed in Dyad DNA engine involving an initial denaturation at 98°C for 5 minutes and 28 cycles of 98°C for 40s, 58°C for 40s, 72°C for 20s with a final extension at 72°C for 4 minutes producing approximately 570bp long amplicon.

PCR product size and purity were checked by performing gel electrophoresis as described in section (2.4.3). Most samples produced a single specific product with exception of few samples that had some non-specific products amplified as well. This might be due to the broad range of the primers utilised, producing extraneous products in presence of low amount of target gene and/or high host DNA contamination. For these samples gel purification was performed on the gel cut of the amplicon of interest using QIAquick Gel Extraction kit (Qiagen) following manufacturer's protocol. Rest of the PCR reactions were cleaned of PCR constituents and primer dimers using the Agencourt® AMPure® XP magnetic bead purification system (Beckman Coulter, USA) according to manufacturer’s instructions.

4.2.2.3 DNA quantification, standardization and pooling

Subsequent quantification was via the Quant-iT™ PicoGreen ® (Molecular Probes Inc., Invitrogen, USA) assay technique as per the manufacturer’s instructions. Fluorescence of samples was assessed in duplicate along with 2 fold serially diluted standard DNA ranging from 100ng/μl to 1.56ng/μl, at 480 nm excitation and 520 nm emission detection, using a FluoStar Omega Spectrophotometer (BMG Labtech, UK). Concentration (ng/μl) of dsDNA in samples was extrapolated from the standard curve and used to determine the concentration in molecules/μl. Each sample was then standardized with 1 x TE buffer to 10⁹ molecules/μl. Pooled amplicon libraries were prepared using 5 μl of the standardized amplicons from each sample allocated in a 454 quarter run. Purity of the pooled amplicon library was verified using the Agilent 2100 Bioanalyzer (Agilent Technologies, UK) high-sensitivity dsDNA kit before they were dispatched for sequencing utilising the Genome Sequencer FLX Instrument titanium series (454 Life Sciences, Roche Diagnostics, UK) to Liverpool (Center for Genomic Research, Liverpool, UK).
4.2.3 Sequence data analysis using QIIME (Quantitative insights into microbial ecology)

Sequencing data was available from each 454 quarter run in the form of three files: standard flowgram format (.sff) files containing the binary coded pyrosequencing result of 454 outputs, sequence files (.fasta) and (.qual) files encoding the quality of the sequence reads generated. Sequence processing and microbial community analysis was done using the python based QIIME (quantitative insights into microbial ecology) pipeline version 1.3.0 and 1.5.0 (Caporaso et al., 2010b). This open source software provides a complete pipeline with various programs, developed by other groups, integrated and required for sequence curation and microbiome analysis, along with sophisticated graphical and visualization tools.

4.2.3.1 Sequence processing and curation

Initial sequence processing involved demultiplexing to assign reads to individual samples. Low quality reads were removed based on trimming, primer matching and quality parameters and are described in detail in results section 4.3.1.2. DeNoiser algorithm was utilised to correct for sequencing artefact (Reeder and Knight, 2010). Curated sequence files generated from the four regions of the 454 plate corresponding to the four visit group in Set 1 were concatenated for further analysis at this point. Similarly, cleaned sequence files for Set 2 and Set 3, generated from two 454 quarter runs of the chip, were concatenated together for downstream analysis. Although primers utilised for 454 sequencing were eubacteria specific, the low stringency PCR conditions applied to obtain fair representation of bacterial community and PCR yield from all samples allowed some non-specific amplification of host DNA. Human sequences were eliminated by performing blast against the human genomic and transcript blast database downloaded from NCBI (National centre for biotechnology information; ftp://ftp.ncbi.nlm.nih.gov) and then aligned (PyNAST; (Caporaso et al., 2010a) against the Greengenes core reference alignment (DeSantis et al., 2006). UCHIME algorithm (USEARCH 4.2.66;(Edgar et al., 2011) installed as a standalone program on windows, due to significantly shorter processing time and better sensitivity compared to chimera slayer (Haas et al., 2011) present in QIIME, was used for reference database (‘gold’; http://sourceforge.net/projects/microbiomeutil/files/) based chimera removal. Before moving on to microbial community analyses samples with <500 reads also were removed.
4.2.3.2 Bacterial community classification and diversity analysis

**Taxonomic classification:** Ribosomal Database Project (RDP) classifier integrated in QIIME was used for taxonomic classification. RDP classifier (v2.2) utilises a naïve Bayesian classifier algorithm based on word size of 8 and returning the highest scoring taxonomy to classify the sequence reads into taxonomic groups from phylum to genus level at 80% bootstrapped confidence interval (CI) (Wang et al., 2007). Phylum and genus level microbial characterisation were used for further statistical analyses and graphic outputs of the bacterial community.

**Operational taxonomic unit (OTU) classification and microbial diversity analyses:** For generation of OTUs reads were grouped at 97%, assumed to represent species level (Drancourt et al., 2000) sequence similarity using uclust (Edgar, 2010); *de novo* clustering method and picking a representative sequence (seed) for each cluster/OTU which, are then mapped onto a taxonomic group. Phylogenetic tree (FastTree 2; Price et al., 2010), required for phylogenetic distance measures utilising the evolutionary relationship between bacterial groups, on aligned sequence was built. Final output is an assembled OTU table with taxonomic assignments for all the samples. ANOVA test in QIIME was used to compare difference in OTU abundance between groups.

Based on OTU (0.03) groups, sample rarefaction curves were generated and used to determine the even sequencing depth of samples for microbial community diversity analyses (Andrew and Mark, 1993; Heck et al., 1975). For determining within sample diversity (alpha diversity) Chao1 (Chao, 1984) estimating species richness based on ratio of singletons to doubletons in samples, Shannon-Weiner (H) (Shannon, 1948) and Simpson (1-D) (Simpson, 1949) index derived based on the number of OTUs and their distribution and Shannon equitability (E_H) (Shannon, 1948) index representing only the dispersion pattern of OTUs were calculated. Among H and 1-D index Shannon index is more sensitive to increases in species richness and gives more weightage to rarer groups by log transformation of relative abundance while, Simpson diversity index squares the relative abundance giving more weightage on abundant communities in a sample and thus can sometimes yield contradictory results (Harini, 2002). Differences in microbial community members (OTUs present) and community structure (relative abundance of members of OTUS) between samples (beta diversity) were deduced using UniFrac distance measure (Lozupone and Knight, 2005) that takes into account the
phylogenetic relationship of OTUs between samples. Based on this distance measure, principal coordinate analysis (PCoA) was performed and 3-D plots (Chen et al., 2009) were used to visualise the clustering of samples. QIIME was used to perform non-parametric two-way comparisons of UniFrac distance measure between visit times, smoking status, COPD GOLD stage and treatment groups employing 999 permutations. A Bonferroni correction was applied for multiple comparisons.

A summary flowgram of bioinformatics analysis pipeline is shown in Figure 4.2  

**Figure 4.2**

Flowgram of microbial community 16SrDNA sequence analysis: Sequence data curation, classification and diversity analysis were performed in QIIME. Square brackets outline the algorithm and program names utilised at various stages of data analysis.

### 4.2.4 Statistical analysis and graphical output using other programs

Microsoft excel (2007) was used to represent microbiome through bar charts and pie charts at phylum and genus level.

All univariate statistical analyses, unless otherwise indicated, were performed in Prism v5.0 (GraphPad Software, San Diego, CA) and SPSS v 18(2009, Chicago; SPSS Inc). Pearson correlation test was used to study relationships and associations between
groups. For categorical data significance test was performed using chi-square test. For continuous parametric data unpaired student t-test and one way analysis of variance (ANOVA) was used for between groups comparisons two and more than 2 groups respectively at population level. For within-group longitudinal comparisons, the paired t-test (comparison across two visits) and Repeated measures anova (RM-ANOVA, for comparison across more than two visits) were used. For ANOVA and RM-ANOVA Tukey's HSD procedure was used to determine statistical significance of pair-wise comparisons. Equivalent testing of non-parametric data was performed using the Mann-Whitney U-test for 2 groups; the Kruskal Wallis test for more than two groups at population level in cross-section; and the Wilcoxon signed rank test (n=2) and Friedman test (n>2) for longitudinal data.

Multivariate analyses other than the ones performed in QIIME included hierarchical clustering, principal components analysis (PCA) and permutation multivariate analysis of variance (PERMANOVA). 'R', v2.12.2, a statistical computing language (R Development Core Team, 2008)\(^{(http://www.R-project.org.)}\) was used to draw heatmaps, performing hierarchical clustering and principal component analysis (PCA). PERMANOVA (Anderson, 2001), installed as a standalone program in windows, is based on the principle that no significant difference in the F-statistic between the actual and shuffled data will exist if the null hypothesis is true. To determine if changes in the microbiome were associated with visit times, PERMANOVA was performed on distance measure using 999 permutations.

A p value of <0.05 was taken as a statistically significant value for all analyses.
4.3 Results

4.3.1 Sequencing plate design, patient characteristics and pre-processing of reads

4.3.1.1 Study design and sample sets for 454 sequencing:

Three sample sets were used for microbial community analysis and are summarised in Figure 4.3. A total of six 454 runs were performed utilizing 1.5 454 chips.

Set 1- Sequential stable, exacerbation, follow-up and recovery samples from individuals (SEFR set) (30 subjects, 120 samples): To characterise and explore changes in the microbiome at the four visit times representing four different clinical stages of COPD, 120 samples obtained from 30 exacerbating COPD subjects providing samples at each stage of their exacerbation were sequenced. This will be henceforth called the SEFR set.

For the SEFR sample set a single 454 chip was partitioned into quarters using mechanical barriers and each quarter was assigned all samples from a visit. Thirty unique barcoded forward primers were designed and were used to tag the samples in each quarter such that all visits from a subject had the same identifier sequence.

Samples from each subject were assigned a subject number between 1 and 30, preceded by S, E, F or R to denote the visit type of the samples.

Set 2 - Sequential stable samples of exacerbating individuals (SSE set) (11 subjects, 33 samples): To examine whether the microbiome varied longitudinally in COPD subjects at stable state, three samples collected at different stable visits from each of 11 subjects, were sequenced. The subjects were a subset of the SEFR study and the sample-set will henceforth be called the SSE set.

Set 3- Non-exacerbating individuals stable samples (NES set) (9 subjects, 9 samples): A proportion of subjects included in the clinical study did not exacerbate. To examine whether the microbiome at stable state differed in subjects that did and did not exacerbate during the clinical study period, baseline samples from 9 non-exacerbated subjects were sequenced and formed set 3, henceforth called the non-exacerbating stable sample NES set.
Two 454 runs utilising half of a chip with 21 unique barcoded forward primers for multiplexing samples in each run were used for sequencing the 42 samples of sets 2 and 3. Samples of the SSE group were assigned a subject number from 1 to 11, preceded by A, B or C to denote serial stable samples in chronological sequence. The 9 baseline samples from non-exacerbated subjects were labelled as NES1 to NES9.

Patient characteristics for all three sets are summarised in Table 4.3

<table>
<thead>
<tr>
<th>Patient demographics</th>
<th>SEFR (n=30)</th>
<th>SSE (n=11)</th>
<th>NES (n=9)</th>
<th>Anova</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (Mean ± SEM)</td>
<td>70.04 ± 1.62</td>
<td>69.9 ± 3.09</td>
<td>72.89 ± 2.94</td>
<td>0.694</td>
</tr>
<tr>
<td>BMI (Mean ± SEM)</td>
<td>25.77 ± 0.87</td>
<td>23.99 ± 1.32</td>
<td>28.52 ± 1.74</td>
<td>0.117</td>
</tr>
<tr>
<td>post-bronchodilator FEV1 predicted (Mean ± SEM)</td>
<td>0.52 ± 0.04</td>
<td>0.57 ± 0.06</td>
<td>0.51 ± 0.06</td>
<td>0.70</td>
</tr>
<tr>
<td>FEV1/FVC (Mean ± SEM)</td>
<td>0.49 ± 0.02</td>
<td>0.50 ± 0.04</td>
<td>0.50 ± 0.04</td>
<td>0.938</td>
</tr>
<tr>
<td>Number of exacerbation in previous year (Mean ± SEM)</td>
<td>4 ± 0.50</td>
<td>5 ± 1.07</td>
<td>2.44 ± 0.89</td>
<td>0.162</td>
</tr>
<tr>
<td>Gender (%male)</td>
<td>80</td>
<td>81.8</td>
<td>66.67</td>
<td>0.658</td>
</tr>
<tr>
<td>Smoking status (% current smoker)</td>
<td>30</td>
<td>9.09</td>
<td>22.22</td>
<td>0.377</td>
</tr>
<tr>
<td>GOLD stage %</td>
<td>2 (moderate)46.66</td>
<td>54.54</td>
<td>44.44</td>
<td>0.987</td>
</tr>
<tr>
<td></td>
<td>3 (Severe)40</td>
<td>36.36</td>
<td>44.44</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4 (very severe)13.33</td>
<td>9.09</td>
<td>11.11</td>
<td></td>
</tr>
</tbody>
</table>

Patient characteristics for the three sets of 454 sequencing samples: BMI- body mass index. FEV1/FVC determines the extent lung function decline, Post- bronchodilator FEV1 predicted and GOLD stage determines the severity stage of COPD.

### 4.3.1.2 Raw sequence read depth and pre-processing of raw sequence files

**SEFR set:** The raw sequencing depth obtained from each of the visit group runs, with the mean sequence length is shown (Table 4.4). Samples from subjects 21 and 23 did not generate reads for three or more of their four visit samples and were excluded from sequence analysis.
Raw sequence run files were demultiplexed and sequence reads assigned to individual samples based on mapping file. To ensure that sequence reads represent the true microbial community rather than PCR or sequencing artifacts, sequence reads meeting the following criteria: > 25 sequence quality read (Phred score of >99% for DNA base calling accuracy), containing no ambiguous bases, with < 8 homopolymers, < 2 bases primer mismatch and > 200 bases read length were selected and used for further analysis (Schloss et al., 2011). It has been shown that 454 sequencing can generate low quantities of sequences that are variations of the original sequence. In downstream analysis these may inappropriately increase the number of sequence clusters and therefore OTU groups (Reeder and Knight, 2010). To adjust for this, the DeNoiser algorithm was performed on the filtered sequences. This puts the variant sequence reads back into the original sequence read cluster. Host sequences and sequences that did not satisfy the blast or alignment parameters set as well as chimeric sequences were also removed (Schloss et al., 2011).

At the end of this process, approximately 41% of the total numbers of sequence reads with a mean sequence length of 491 (SD ± 10) base pairs were retained for taxonomic classification and phylogenetic analysis (Table 4.5). Of the initial 120 samples from SEFR study, 101 samples were left for bacterial community analysis with 25, 23, 26 and 27 samples from S, E, F and R visit groups respectively.

Table 4.4

<table>
<thead>
<tr>
<th>Longitudinal sample groups</th>
<th>Total number of sequence reads (30 samples)</th>
<th>Samples with no reads</th>
<th>Mean Sequence length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stable</td>
<td>103678</td>
<td>S21, S23</td>
<td>430</td>
</tr>
<tr>
<td>Exacerbation</td>
<td>178766</td>
<td>E1, E21, E23</td>
<td>472</td>
</tr>
<tr>
<td>Follow up</td>
<td>153731</td>
<td>F21, F23</td>
<td>459</td>
</tr>
<tr>
<td>Recovery</td>
<td>150360</td>
<td>R21</td>
<td>446</td>
</tr>
</tbody>
</table>

SEFR 454 raw sequence file statistics: Each row represents a quarter run of a 454 chip and gives the total number of reads generated from 30 samples. Eight samples across the four runs didn't produce any reads.
16S rDNA based COPD bacterial community analysis

Table 4.5

<table>
<thead>
<tr>
<th>Filters</th>
<th>SEFR set</th>
<th>NES set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filtering and Denoising</td>
<td>63019</td>
<td>32492</td>
</tr>
<tr>
<td>Human sequence filtration</td>
<td>40621</td>
<td>33850</td>
</tr>
<tr>
<td>Alignment</td>
<td>40569</td>
<td>88247</td>
</tr>
<tr>
<td>Chimera filtration</td>
<td>33858</td>
<td>87119</td>
</tr>
<tr>
<td>Remove Unwanted taxa</td>
<td>33850</td>
<td>61865</td>
</tr>
<tr>
<td>Remove samples &lt; 500 reads</td>
<td>32492</td>
<td>58727</td>
</tr>
<tr>
<td>Total reads</td>
<td>392548</td>
<td>240203</td>
</tr>
<tr>
<td>Sequence reads removed (%)</td>
<td>33.08</td>
<td>1.31</td>
</tr>
</tbody>
</table>

Number of sequence reads removed at each filtering step in SEFR set: Column headers represent sequential filtration steps applied to sequence files. Last row gives the percentage of reads lost at each filtration step from its preceding step.

SSE and NES set: After applying all the filtration steps to eliminate low quality, contaminating and erroneous reads only 39% of the reads with mean length 477 (SD ± 37) of were retained from the initial sequence reads (Table 4.6). Out of 42 samples five (A6, B6, C6, C9 and B10) did not generate any reads and two more samples after all the filtration steps had <500 reads and therefore further bacterial community analysis was performed on 35 samples.

Table 4.6

<table>
<thead>
<tr>
<th>Filters</th>
<th>Run 1</th>
<th>Run 2</th>
<th>Total reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw reads</td>
<td>197581</td>
<td>167593</td>
<td>365174</td>
</tr>
<tr>
<td>Filtering and denoising</td>
<td>95786</td>
<td>99536</td>
<td>195322</td>
</tr>
<tr>
<td>Remove Human sequence</td>
<td>82937</td>
<td>87733</td>
<td>170670</td>
</tr>
<tr>
<td>Alignment</td>
<td>82244</td>
<td>86271</td>
<td>168515</td>
</tr>
<tr>
<td>Remove Chimeras</td>
<td>68727</td>
<td>73553</td>
<td>142280</td>
</tr>
<tr>
<td>Remove unwanted taxa</td>
<td>68570</td>
<td>73534</td>
<td>142104</td>
</tr>
<tr>
<td>Remove samples &lt; 500 sequence reads</td>
<td>68505</td>
<td>73533</td>
<td>142038</td>
</tr>
<tr>
<td>Sequence reads removed (%)</td>
<td>46.51</td>
<td>12.62</td>
<td>0.05</td>
</tr>
</tbody>
</table>

132
Number of sequence reads removed at filtration steps in SSE and NES set: Rows represent read numbers retained after each filtration step for the two 454 quarter runs and the percentage of reads lost across both runs.

A summary of samples included for 454 sequencing in each set and the outcomes of processing is presented (Figure 4.3)

**Figure 4.3: Schematic representation of the sample set used for 454 sequencing**

![Diagram of sample set](image)

4.3.2 Taxonomic Characterisation of the bacterial community

Taxonomic characterisation was performed at the phylum and genus level by assigning 16S sequence reads to bacterial taxonomic groups through RDP classifier, using naïve Bayesian classifier algorithm at 80% CI (Wang et al., 2007).

4.3.2.1 Phylum level classification

**SEFR set:** In all, fifteen bacterial phyla were identified from samples across the four visit groups. There were 12, 11, 14 and 10 phylum groups identified respectively for samples of S, E, F and R visit groups. The four most prevalent phyla (mean relative abundance ± standard deviation) in all visit types were: Proteobacteria (0.39 ± 0.38),
Firmicutes (0.47 ± 0.33), Bacteroidetes (0.07 ± 0.08) and Actinobacteria, (0.06 ± 0.09). There was no significant difference in the average relative abundance of phyla across visit groups (p=0.994, chi square test) (Figure 4.4). Fusobacteria were also observed frequently in samples across the 4 visit types but had <1.5% relative abundance in all samples (Figure 4.5). Together, bacteria of the phyla Firmicutes and Proteobacteria constituted approximately 80% of sequence reads in most samples (Figure 4.6).

**SSE and NES set:** The relative abundance and frequency of occurrence of bacterial phyla in samples of the SSE and NES groups was similar to the SEFR group. In both sample sets Proteobacteria, Firmicutes, Bacteroidetes and Actinobacteria were the four most dominant phyla (Figure 4.7 and Figure 4.8). There was no significant difference in the average relative abundance of different phyla between samples of the SSE and NES groups (p=0.393), with Proteobacteria and/or Firmicutes constituting >80% of the community in 32 out of 35 samples (Figure 4.9).

**Figure 4.4**

Bacterial community structure at phylum level in each visit type of the SEFR set: Pie charts represent the mean proportion of bacterial phyla in samples of each visit type. All phyla present at <1.5% average relative abundance were grouped together in the group labelled "Others". Labels: S - stable visit; E - exacerbation visit; F - follow-up visit; and R - recovery visit.
**Figure 4.5**

Prevalence of bacterial phyla in samples of the SEFR set: The prevalence is defined as the percentage of samples of each visit type in which each phylum was detected. Prevalence_S, Prevalence_E, Prevalence_F, Prevalence_R stand for the prevalence in samples from S, E, F and R visits respectively.

**Figure 4.6**

Dominance of Proteobacteria and Firmicutes in individual samples: heatmap presents the bacterial composition of each sample as the relative proportion of each phylum identified. This is colour coded (colour key and histogram) with white denoting bacterial groups not detected and increasing bacterial...
proportions represented as colours trending from yellow to violet. Each row represents a bacterial group and each column a sample. The multicolour row above the heatmap shows the visit type of a sample and the number row below the heatmap indicates the subject number of samples. Available S, E, F, R visit samples for each subject is presented sequentially, with red vertical lines separating samples from different subjects.

**Figure 4.7**

**Bacterial community structure at phylum level in SSE and NES sets:** Data for each of the three stable samples obtained in the SSE group is presented on a separate pie chart (labelled A to C as previously described). All the phyla present at <1.5% average relative abundance were included together as "Others". Legend key shows the colour coding of the bacterial groups.
Prevalence of bacterial groups in SSE and NES sets: Prevalence_A, Prevalence_B and Prevalence_C denote the 3 stable samples for the SSE group and Prevalence_NES denotes the prevalence of phyla in NES group.

Bacterial composition and their relative abundance in individual samples in SSE and NES set samples: heatmap presents the bacterial composition of each sample (column) as the relative proportion
of each phylum (row) identified. The colour scheme is explained in more detail with the legend for Figure 4.6.

4.3.2.2 Genus level classification

As phylum level classification includes a broad range of bacterial groups within each phylum it was decided to interrogate the bacterial composition at genus level.

**SEFR set:** A total of 165 groups were present across the 101 samples analysed at genus level. In keeping with the phylum bacterial composition, the dominant groups at genus level were *Streptococcus* (phylum Firmicutes); *Haemophilus* (phylum Proteobacteria) and *Moraxella* (phylum Proteobacteria), contributing 36%, 29% and 7% of the total reads respectively (Figure 4.10 and Figure 4.11). *Streptococcus* was present in all the samples. *Neisseria* was detected in about 70% of the samples with a mean relative abundance of 5%. Other genera observed in over 80% of samples but in very low proportions were *Actinomyces, Prevotella,* and *Veillonella* (Figure 4.10 and Figure 4.11). There was considerable variability between samples in the number of bacterial groups identified at genus level at all visit times, with a range of 9 to 45, 11 to 54, 9 to 48, and 12 to 53 in S, E, F and R visit samples.

**SSE and NES set:** Out of 148 genera identified in the SSE and NSE sample sets, *Streptococcus* was the most dominant bacterial group identified, present in all the samples and contributing about half of the total reads sequenced. *Haemophilus, Neisseria, Pseudomonas* and *Moraxella* from Proteobacteria were the other dominant groups (Figure 4.12). Figure 4.13 shows the bacterial composition in each sample at genus level.
Average relative abundance, Prevalence and Total abundance in SEFR set: Total Abundance (green bars) indicate the absolute proportion of sequence reads represented by a bacterial group while Average relative abundance (yellow bars) is the mean proportion of a bacterial group across 101 samples. The 20 genera presented in the graph are sorted in the descending order of total abundance from left to right and together constituted 96% of the total abundance. All 20 genera had >0.04% average relative abundance, >50% prevalence and >20% total abundance.
Figure 4.11

Bacterial composition at genus level in individual samples in SEFR set: heatmap shows the relative proportion of 49 genera (rows) in individual samples (columns) with the highlighted in red genera being the dominant across the samples. Genera present at $\geq 0.02\%$ total abundance and $\geq 1\%$ maximum relative abundance in at least one sample, in each sample are shown. Heatmap colouring scheme is explained in Figure 4.6.
**Figure 4.12**

Average relative abundance, Prevalence and Total abundance in SSE and NES sets: Green bars - Total abundance; Yellow bars – Average relative abundance. The 20 genera presented in the graph constituted 93% of the total abundance.
**Bacterial composition at genus level in individual samples in SSE and NES sets:** Heatmap shows the relative proportion of genera (rows) in individual samples (columns). Genera present at $\geq 0.02\%$ total abundance and $\geq 1\%$ maximum relative abundance in at least one sample, in each sample are represented. Heatmap colouring scheme is explained in Figure 4.6.

### 4.3.2.3 Concordance between the 454 sequencing, qPCR and culture results for the 3 dominant genera

The bacterial composition of COPD samples identified with 454 sequencing showed dominance of *Streptococcus*, *Haemophilus* and *Moraxella*. This was in keeping with the findings of qPCR and positive cultures. The association of 454 sequencing with qPCR and culture results was examined in samples of the SEFR set also analysed using these methods. For qPCR, the log counts for *H. influenzae*, *M. catarrhalis* and *S. pneumoniae* were compared with the percentage of sequence reads for *Haemophilus*, *Moraxella* and *Streptococcus* obtained with 454 sequencing at the genus level. There was a significant
linear correlation between the two methods for the bacterial groups studied, except *Streptococcus* (Table 4.7). The relationship between 454 sequencing and culture detection was analysed by comparing the percentage sequence reads for *Haemophilus*, *Moraxella* and *Streptococcus* between samples that were respectively culture positive and negative for *H. influenzae*, *M. catarrhalis* and *S. pneumoniae*. Culture positive samples expressed significantly higher % sequence reads, compared with culture negative samples for bacterial groups other than *Streptococcus* (Table 4.7).

### Table 4.7

<table>
<thead>
<tr>
<th>Comparison groups</th>
<th>qPCR vs 454 (Linear Regression)</th>
<th>Mean proportion 454 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2$</td>
<td>Sig</td>
</tr>
<tr>
<td><em>H. influenzae</em> &amp; <em>Haemophilus</em></td>
<td>0.62</td>
<td>$&lt;0.001$</td>
</tr>
<tr>
<td><em>M. catarrhalis</em> &amp; <em>Moraxella</em></td>
<td>0.59</td>
<td>$&lt;0.001$</td>
</tr>
<tr>
<td><em>S. pneumoniae</em> &amp; <em>Streptococcus</em></td>
<td>0.05</td>
<td>0.13</td>
</tr>
</tbody>
</table>

**Relationship between 454 sequencing, qPCR and culture:** Each of the rows shows the result of the bacteria from qPCR and culture that was compared with its genus group reads from 454. Linear regression between the qPCR log counts and percentage sequence reads in each of the three comparison groups was performed. Between-group comparisons of culture positive and culture negative samples were performed using the student t-test.

### 4.3.2.4 Changes in the bacterial community at different visit stages in SEFR set

A key objective of this project was to determine whether longitudinal changes in the bacterial composition of sputum were detectable during different clinical phases of COPD. In the following sections the results of analyses comparing the bacterial composition at different levels using 454 sequencing data from the SEFR are presented.

Hierarchical cluster analysis using the Euclidean distance method was performed at phylum level on all available samples of the SEFR set. It was hypothesized that a cluster model by visit type would be generated if this was a significant determinant of
differences in bacterial composition for the sample set. Three clusters were identified with no association between microbiota characterized at the phylum level and the study visit group. These clusters were rather formed on the basis of the proportional representation within samples of the 2 dominant groups, Proteobacteria and Firmicutes (Figure 4.14). The clusters could be respectively identified as a group with a high abundance of Proteobacteria sequences, a group with a high abundance of Firmicutes sequences and a group with a mixture of Proteobacteria and Firmicutes sequences at moderate abundance (Figure 4.14).

These observations were supported by findings at genus level. Principal component analysis (PCA) identified 3 groups that were characterized on the basis of the abundance of genera belonging to the two dominant phylum groups (Figure 4.15). *Haemophilus*, *Streptococcus* and *Moraxella* were the major determinants at genus level of sample grouping with PCA. The visit type did not differ across the groups (Figure 4.15).

It is possible that important but subtle differences contributed by low abundance groups to clustering at genus level were masked by the dominant bacterial groups. To examine this, hierarchical clustering was performed on logit transformed relative abundance data at genus level. Using this approach, two groups were identified characterized by differences in the diversity of detected bacterial groups. No association with the visit type was observed (Figure 4.16).

To explore the variability in the microbiome between visits, analysis was performed at genus level in the 19 subjects of the SEFR set with 454 sequence reads obtained for all 4 visits. Differences in the relative abundance of genera across visits were investigated using PERMANOVA (Anderson, 2001) on the Bray Curtis dissimilarity measure (Bray and Curtis, 1957). No significant between-visit difference in the microbial community was identified at genus level (p=0.4668). Exacerbation and stable visit types exhibited greater within-visit than between-visit dissimilarity between samples (Table 4.8). There was no significant difference between (p=0.2094) or within subjects (p=0.2995) for the observed number of genera at 4 visit types.
**Table 4.8**

<table>
<thead>
<tr>
<th></th>
<th>Stable</th>
<th>Exacerbation</th>
<th>Follow-up</th>
<th>Recovery</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stable</strong></td>
<td>67.428</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Exacerbation</strong></td>
<td>66.497</td>
<td>68.167</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Follow-up</strong></td>
<td>65.814</td>
<td>66.357</td>
<td>66.390</td>
<td></td>
</tr>
<tr>
<td><strong>Recovery</strong></td>
<td>67.471</td>
<td>66.502</td>
<td>68.395</td>
<td>65.548</td>
</tr>
</tbody>
</table>

Bacterial community analysis for difference between COPD visit types: The Bray Curtis dissimilarity measure is derived on the basis of relative abundance reading and whether bacterial groups are present or absent in samples. The quantitative dissimilarity index is expressed as a number between 0-100 with a higher number indicating higher dissimilarity between samples.

**Figure 4.14**

Samples form three major cluster based on Proteobacteria and Firmicutes dominance rather than SEFR visit types: Clustering was done using Euclidean distance and complete linkage method. Each row represents a phylum group in a sample and each column a sample. The heatmap is based on the proportional abundance of microbiota constituents at phylum level. The top single row above the heatmap represents the four study visit group with green colour indicating stable, red - exacerbation, yellow- follow up and blue- recovery visits.
Separation of SEFR samples influenced by dominance of three genera rather than visit types:
Principal Component Analysis (PCA) plot is based on the principal components of 1 and 2 co-ordinates.
The dominant bacterial groups were determined from the variable loading's coordinates.
16S rDNA based COPD bacterial community analysis

**Figure 4.16**

Low diversity and bacterial richness in *Haemophilus* dominant samples: Logit transformation of the relative abundance (RA) data was done by log ((RA/(1-RA))+0.00001) where 0.00001 is added as a constant for bacterial groups not detected. Groups not detected represented in white; -10, -5, 0 and 5 on colour key histogram represents approximately following bacterial proportions 5E-05, 5E-03, 0.5 and 0.99. Bacterial groups represented are as explained in Figure 4.11.
4.3.2.5 Stability of microbial pattern in longitudinal stable samples in SSE set

Longitudinal stability of the microbiome in sputum was analysed using samples of the SSE set. Similar to analyses in the SEFR set, hierarchical clustering at genus level of the SSE set formed 2 clusters characterised on the basis of relative abundance of bacterial groups from the Proteobacteria and Firmicutes. There was an overall tendency for longitudinal stable samples from the same subject to cluster closely together (Figure 4.17).

Comparison of NES set stable samples with first stable sample of each individual from SSE set showed no difference between stable samples from exacerbated and non-exacerbated COPD subjects (Figure 4.18).

**Figure 4.17**

Hierarchical clustering of the COPD microbiota in longitudinal stable samples: heatmap shows the clustering of the longitudinal stable samples of SSE set. Each individual is given a unique colour to
represent its samples in the top single row above heatmap. Bacterial groups represented are as explained in Figure 4.13.

**Figure 4.18**

Comparison of bacterial community between exacerbated and non-exacerbated stable samples: heatmap shows the clustering of the longitudinal stable samples from NES set with the "A" stable samples from SSE set. Bacterial groups represented are as explained in Figure 4.13

### 4.3.3 Sequence similarity based classification and diversity analysis of COPD microbiota

The bacterial community of each sample was classified further into operational taxonomic units (OTUs) based on sequence similarity at the 0.03 cut-off level (97% similarity threshold), using the uclust method (Edgar, 2010). OTUs were mapped onto taxonomic groups and phylogenetic trees to calculate bacterial diversity metrics and for comparative analysis. A total number of 616 OTUs were identified in the SEFR set and 443 OTUs together in the SSE and NES sets.
4.3.3.1 Alpha diversity measurement

SEFR set: There was considerable heterogeneity in species richness of samples for all 4 visit types (Figure 4.19). Most of the samples had < 2000 reads and in stable visit all but one sample <1350 reads. As most samples in rarefaction curves approached an asymptote after 500 sequence reads indicating fewer additional OTUs accumulating with increasing sequencing depth, an even sampling depth of 600 reads was chosen to avoid further loss of longitudinal samples for analysis. The mean observed species richness was comparable for the 4 visit types, although a higher variance was observed between samples of the exacerbation visit (Figure 4.20).

In addition to the observed species ($S_{\text{obs}}$), Chao1 ($S_{\text{est}}$) was calculated for each sample to estimate the total number of bacterial groups that would be present. Though exacerbation visit samples had the highest mean for observed and estimated richness, there was no significant difference in species richness or Chao1 estimates across the 4 visit types at either population level ($p=0.9575$ ($S_{\text{obs}}$), $p= 0.7981(S_{\text{est}})$) or within subject ($p=0.4962$ ($S_{\text{obs}}$), $p= 0.6803(S_{\text{est}})$) (Table 4.9).

Although $S_{\text{obs}}$ and $S_{\text{est}}$ metrics inform the number of bacterial groups that are observed or predicted to exist, they do not provide information about the dispersion of bacterial groups within samples. The Shannon index ($H$) and Simpson index ($1-D$) that take into account the relative abundance of bacterial communities as well as richness within a sample, were used to deduce dispersion indices. Overall, recovery samples were the least diverse communities but no significant difference in bacterial diversity was identified across the 4 visit types either within subjects ($p=0.1084(H)$, $P=0.0819 (1-D)$) or between subjects ($p= 0.6679 (H)$, $p= 0.3510 (1-D))$ (Table 4.9).

Shannon's equitability ($E_{H}$) is a test that can isolate the relative abundance component of diversity from species richness. This was performed to compare the samples based only on distribution pattern of bacterial communities. No between-visit differences in this aspect of diversity were identified (Table 4.9).
Rarefaction plot present a wide variation in exacerbation samples among the 4 visit type: Line plot of number of OTUs as a function of the number of reads sub sampled in each of the visit types at 0.03 level sequence similarity.
No difference in mean observed species richness across the 4 visit types: Line plot of mean OTU richness with standard deviation bars at different sequence read sub-sampling levels for the 4 visit types.

Table 4.9

<table>
<thead>
<tr>
<th>Visit types</th>
<th>Sobs</th>
<th>Eest</th>
<th>F</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>α diversity indices</td>
<td>Population level (n= 28 subjects; 101 samples)</td>
<td>p (Anova)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sobs</td>
<td>30.14 ± 3.49</td>
<td>31.98 ± 5.36</td>
<td>29.58 ± 4.79</td>
<td>29.04 ± 3.19</td>
</tr>
<tr>
<td>Eest</td>
<td>43.81 ± 5.00</td>
<td>48.64 ± 7.31</td>
<td>40.8 ± 3.19</td>
<td>44.59 ± 4.72</td>
</tr>
<tr>
<td>H</td>
<td>2.12 ± 0.21</td>
<td>2.08 ± 0.31</td>
<td>2.29 ± 0.22</td>
<td>1.88 ± 0.22</td>
</tr>
<tr>
<td>1-D</td>
<td>0.55 ± 0.05</td>
<td>0.51 ± 0.07</td>
<td>0.59 ± 0.05</td>
<td>0.47 ± 0.05</td>
</tr>
<tr>
<td>E_H</td>
<td>0.43 ± 0.03</td>
<td>0.40 ± 0.05</td>
<td>0.48 ± 0.03</td>
<td>0.38 ± 0.03</td>
</tr>
</tbody>
</table>

Comparison of species richness, diversity and evenness index between four visit types: For each visit group, alpha diversity indices are presented as mean ± SEM.

SSE and NES sets: Compared with the SEFR set, samples from SSE set showed significant accumulation of bacterial OTU groups beyond a sampling depth of 500 reads (Figure 4.21). With the exception of one sample each from the SSE and NES sets all the samples had >1700 reads. In most samples a \( S_{obs} \) plateau was reached at between 1500 and 2000 reads and therefore 1710 reads was used for even sampling depth.

There was considerable overlap between the mean \( S_{obs} \) curves for the three stable time points in the SSE set, indicating similar species richness over time in clinically stable COPD (Figure 4.22). Diversity and richness measure between longitudinal stable time points in SSE set had similar diversity and richness measures (Figure 4.23).

Compared with the SSE set, samples from the NES set generally showed lower mean species richness (Figure 4.22) but neither richness nor diversity were significantly different from SSE (Figure 4.23).
Figure 4.21

![Rarefaction curves of stable samples in SSE and NES sets](image1)

**Rarefraction curves of stable samples in SSE and NES sets:** Line plot of number of OTUs as a function of the number of reads sub-sampled in each of the stable groups at 0.03 level sequence similarities.

Figure 4.22

![Bacterial richness trend comparison](image2)

**NES set showed a lower bacterial richness trend compared to SSE set stable samples:** line plot of mean species richness with standard deviation bars at different sequence read sub-sampling levels for the 3 longitudinal stable time points in SSE and NES stable samples.
No significant difference in richness and diversity measures at multiple stable points in SSE and NES stable samples: Bars represent the mean along with standard error for richness, diversity measures.

4.3.3.2 Beta diversity
While alpha diversity characterizes a sample based on the number and dispersion of bacterial groups present in a sample, beta diversity quantifies differences in microbial community and abundance between samples as a distance measure. The UniFrac distance used in this analysis incorporates the phylogenetic tree along with the OTUs, providing between-sample comparisons, with weightage applied to phylogenetic distance (Lozupone and Knight, 2005).

SEFR set: The unweighted UniFrac distance measure gives qualitative diversity indices, while the weighted UniFrac also incorporates abundance into the calculation of the diversity measure. Both distances were used on OTUs (0.03) at an even sampling depth of 600 reads to examine differences in the bacterial composition between samples.
Principal co-ordinate analysis (PCoA) plots using both unweighted and weighted UniFrac distance were constructed for samples of the SEFR set. There was no distinct clustering of samples by visit type for either plot (Figure 4.24 and Figure 4.25). At the genus level, the cluster structure of diversity in these samples was dominated by *Streptococcus*, *Moraxella* and *Haemophilus*, using both weighted and unweighted UniFrac distance (Figure 4.24 and Figure 4.25). Other genera belonging to the phyla Actinobacteria and Bacteroidetes have a smaller but detectable influence. In the PCoA plots (Figure 4.24 and Figure 4.25) these other genera group more closely with *Streptococcus*, with less diversity evident for *Haemophilus* and *Moraxella* dominant samples.

**Figure 4.24**

Negative association between *Haemophilus* and other dominant bacterial groups: A- PCoA plot of all SEFR samples coloured by visit group: green for S, red for E, orange for F and blue for R visit group sample shows no difference in bacterial structure between the four visit types; B- Influence of top 10 dominant genera in different areas of the plot with diameter of each OTU representing it’s total abundance across all samples; C to F- plots of samples from each visit group S, E, F and R respectively.
16S rDNA based COPD bacterial community analysis

**Figure 4.25**

PCoA plot on unweighted UniFrac distance in SEFR visit types: PCoA plots based on the unweighted UniFrac distance groups. Plots A to F represent the same orientation as described in Figure 1.23.

Beta diversity box plots of within-visit diversity were constructed and compared across visits, using non-parametric tests in QIIME. Overall, the magnitude and spread of diversity using the weighted UniFrac distance was lower for all samples. As this analysis corrects for the effects of abundance, it suggests this is an important contributor to observed diversity using the unweighted UniFrac distance. For both, diversity is greatest across samples of the exacerbation visit and lowest for the follow-up visit. Pairwise comparisons identified significant variation in bacterial community membership (unweighted UniFrac analysis; \( p=0.045 \)) and structure (weighted UniFrac analysis; \( p=0.045 \)) in the exacerbation group samples compared to stable and follow-up group samples, respectively (Figure 4.26). However no significant difference was observed in either microbial membership (\( p=0.1490 \); unweighted UniFrac) or microbial structure (\( p=0.3676 \); weighted UniFrac) between visit types compared to within visit types using PERMANOVA analysis.
Boxplots of phylogenetic distance measure within visit type in SEFR. Based on samples from each visit types, box plots with median (red), interquartile range (IQR) (blue box) and IQR*1.5 (whisker) of between sample diversity distance measures within each visit type.

Within-subject change in beta diversity across visits (S vs E, S vs F, S vs R, E vs F, E vs R, F vs R) was analysed. Overall, these analyses demonstrated wide variation between subjects in the magnitude of diversity change across visits, with some subjects showing large changes in bacterial composition, while others were quite stable (Figure 4.27). For individual subjects, changes in microbiome membership (presence and absence of bacterial groups) rather than microbial composition (relative abundances of bacterial groups) were observed between the four visits.
Box plot of distance measure between visits in each individual: Box plots, of each individual showing between visit beta diversity distance, with median (red), interquartile range (IQR) (blue box) and IQR*1.5 (whisker). Red diamond dots represent the distance between stable and exacerbation sample in each subject. X axis represents subject numbers and Y axis represents distance measure.

SSE and NES set: PCoA plots based on the weighted UniFrac distance (OTUs at even sampling depth of 1710) of samples in the SSE set showed longitudinal stable samples from individual subjects to cluster together in most instances (Figure 4.28). Samples A8, B4, B5 and C11 were considerably different from other longitudinal samples for these subjects. At the genus level, clusters were dominated by either Streptococcus or Haemophilus (Figure 4.28).

There was no difference in patterns of diversity identified for stable samples from exacerbated and non-exacerbated individuals (Figure 4.29).

In a separate analysis the repeatability of 454 sequencing was tested by measuring the change in PCoA plots of UniFrac distance for stable samples that were analysed as part of both the SEFR and SSE sets. This was performed using Generalised Procrustes analysis and demonstrated good agreement in diversity patterns from two different sequencing reactions for the repeated samples (Figure 4.30).
16S rDNA based COPD bacterial community analysis

**Figure 4.28**

**PCoA plot of SSE set based on phylogenetic distance measure:** Top PCoA plot shows the clustering of SSE set samples and samples are coloured based on their subject numbers. The lower plot shows the top 10 abundant genera influence on the samples.

**Figure 4.29**

**Comparison of beta diversity pattern of NES samples and SSE samples:** NES samples are represented in red and are compared with only first stable sample ("A"), shown in green, of each individual from SSE set.
**4.3.4 Metadata based bacterial community analysis**

To explore whether differences in microbial communities are associated with clinically defined phenotypes of COPD, comparison was made between pre-defined groups of the SEFR set based on smoking status and GOLD stage. Bacterial community analysis was performed at phylum level and diversity measures compared at OTU (0.03) level, normalised to a sequencing depth of 600. Between group analyses were performed at phylum level across all 4 visits and sample diversity measures in stable and exacerbation visits.

**4.3.4.1 Bacterial Community comparison between smokers and ex-smokers**

There were 21 ex-smokers and 9 smokers in the SEFR set. Table 4.10 shows the number of samples available for between smoking status across visit types and within subject comparisons.

There were no significant differences at the phylum level between smokers and ex-smokers. Within each smoking status, no significant change across the 4 visit types was identified (Figure 4.31).
Changes in relative abundance of phyla across SEFR visit type between smokers and ex-smokers:
Graph represents the mean relative proportion and SD for each of the visit type sample grouped based on their smoking status. "Others" include all the minor phylum groups together.

S_{est} and H diversity metrics were calculated in stable and exacerbation visit samples and compared between smokers and ex-smokers. There was no difference in bacterial diversity or bacterial richness between smokers and ex-smokers at stable or exacerbation visit type. Neither smokers nor ex-smokers had a significant change in bacterial diversity or in bacterial richness from stable to exacerbation visit at an individual or population level (Table 4.10). PCoA plots based on weighted UniFrac distance showed no clustering of samples according to smoking status at stable or exacerbation visit. Between-smoking status comparison of the median weighted UniFrac distance within each visit identified significantly less (p=0.0048) dissimilarity of exacerbation samples in current smokers compared with ex-smokers (Figure 4.32). However, no significant between-smoking status difference was identified in the within-subject change of bacterial composition from stable to exacerbation state (Figure 4.32).
Table 4.10 Comparison of alpha diversity metrics based on smoking status:

<table>
<thead>
<tr>
<th>Smoking status</th>
<th>N=25</th>
<th>Sest</th>
<th>H</th>
<th>N=23</th>
<th>Sest</th>
<th>H</th>
<th>Between visit type comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smokers</td>
<td>n=7</td>
<td>40.77 ± 7.96</td>
<td>1.88 ± 0.42</td>
<td>n=7</td>
<td>47.27 ± 9.85</td>
<td>2.02 ± 0.49</td>
<td>0.617 0.821</td>
</tr>
<tr>
<td>Ex-smokers</td>
<td>n=18</td>
<td>46.5 ± 6.52</td>
<td>2.29 ± 0.25</td>
<td>n=16</td>
<td>47.97 ± 10.83</td>
<td>2.07 ± 0.44</td>
<td>0.904 0.642</td>
</tr>
<tr>
<td>Between smoking group comparisons</td>
<td>0.621</td>
<td>0.383</td>
<td>0.967</td>
<td>0.956</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Smoking status</th>
<th>Subjects</th>
<th>N=19</th>
<th>Sest</th>
<th>H</th>
<th>Sest</th>
<th>H</th>
<th>Within subject comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smokers</td>
<td>n=6</td>
<td>37.64 ± 8.66</td>
<td>1.72 ± 0.47</td>
<td>42.31 ± 10.07</td>
<td>2.07 ± 0.58</td>
<td>0.7166 0.521</td>
<td></td>
</tr>
<tr>
<td>Ex-smokers</td>
<td>n=13</td>
<td>46.32 ± 7.30</td>
<td>2.31 ± 0.27</td>
<td>45.44 ± 10.79</td>
<td>2.05 ± 0.48</td>
<td>0.9218 0.522</td>
<td></td>
</tr>
</tbody>
</table>

Comparison of alpha diversity metrics based on smoking status: Table shows the mean and SEM for alpha diversity indices measured. Table A represents between smoking status comparisons (p value) at population level with N = the total number of samples in a visit group and n = the number of samples in a metadata category and Table B within subject comparisons (p value) with N= number of subjects and n = subjects in each metadata.
Comparison of between sample diversity measures based on smoking status: A - PCoA plot of smokers and ex-smokers based on weighted UniFrac distance measure for stable and exacerbation visit samples. B - Boxplots representing median, IQR with min, max whisker of between sample distance within visit type and within individual distance between stable and exacerbation sample segregated based on smoking status.

4.3.4.2 Bacterial community changes based on COPD severity

As previously discussed in (Table 1.1) COPD severity is classified on the basis of lung function into four GOLD stages. The SEFR set comprised 4, 12 and 14 subjects with GOLD stage 4 (very severe), 3 (severe) and 2 (moderate) COPD. Between severity stage comparison at phylum level identified only a significant difference in the "Others" bacterial group between severe and very severe COPD (p=0.015) (Figure 4.33). There was no significant change across visits at phylum level across visits in any of COPD severity stages.
### Figure 4.33

Changes in relative abundance of phyla across SEFR visit types between COPD severity stages:

Graph represents the mean relative proportion and SD for each of the visit type sample stratified based on their COPD severity. "Others" include all the minor phylum groups together.

$S_{est}$ and Shannon alpha diversity metrics were calculated for stable and exacerbation visit samples. Exacerbation samples from subjects with severe COPD had fewer bacterial groups and less diversity but this was not significantly different from the moderate and the very severe stages (Table 4.11). No difference was identified either between the severity stages at stable state or within a severity stage for change in bacterial diversity and richness between stable and exacerbation visit types (Table 4.11). Within-subject analysis showed a significant increase ($p=0.043$) in estimated bacterial richness from stable to exacerbation visit in GOLD stage 4 COPD.

PCoA plots based on weighted UniFrac distance showed no clustering of samples according to COPD severity at stable or exacerbation visit (Figure 4.34). There was also no significant between-severity stage difference for change in bacterial composition from stable to exacerbation visit (Figure 4.34).
### Table 4.11

**Comparison of alpha diversity metrics based on COPD severity stages:** see legend for Table 4.10 for formatting details

<table>
<thead>
<tr>
<th>A</th>
<th>Metadata based analysis</th>
<th>Stable</th>
<th>Exacerbation</th>
<th>Between visit group comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>N=25</td>
<td>n=12</td>
<td>n=11</td>
</tr>
<tr>
<td></td>
<td>GOLD stage* (1 subject excluded as mild)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Moderate</td>
<td>S&lt;sub&gt;α&lt;/sub&gt;</td>
<td>44.57 ± 7.78</td>
<td>51.32 ± 13.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>H</td>
<td>2.26 ± 0.32</td>
<td>1.89 ± 0.55</td>
</tr>
<tr>
<td></td>
<td>Severe</td>
<td>N=10</td>
<td>43.28 ± 8.56</td>
<td>35.97 ± 8.104</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.96 ± 0.32</td>
<td>1.81 ± 0.42</td>
</tr>
<tr>
<td></td>
<td>Very severe</td>
<td>n=3</td>
<td>45.52 ± 8.35</td>
<td>66.61 ± 7.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.26 ± 0.67</td>
<td>3.15 ± 0.26</td>
</tr>
<tr>
<td></td>
<td>Between GOLD stage group comparisons</td>
<td>0.989</td>
<td>0.83</td>
<td>0.356</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B</th>
<th>Metadata based analysis</th>
<th>Subjects</th>
<th>Stable</th>
<th>Exacerbation</th>
<th>Within subject comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>N=20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GOLD stage</td>
<td>Moderate</td>
<td>n=10</td>
<td>39.39 ± 8.26</td>
<td>48.74 ± 14.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.98 ± 0.32</td>
<td>1.90 ± 0.60</td>
</tr>
<tr>
<td></td>
<td>Severe</td>
<td>n=7</td>
<td>50.40 ± 11.14</td>
<td>31.57 ± 7.85</td>
<td>0.121</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.19 ± 0.44</td>
<td>1.76 ± 0.46</td>
</tr>
<tr>
<td></td>
<td>Very severe</td>
<td>n=3</td>
<td>42.52 ± 8.34</td>
<td>60.54 ± 5.93</td>
<td>0.043</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.26 ± 0.67</td>
<td>2.97 ± 0.27</td>
</tr>
</tbody>
</table>

**Legend:**
- S<sub>α</sub>: value of α diversity
- H: value of Shannon index
- Within subject comparisons: statistical significance of differences within subjects
Comparison of between sample diversity measures based on COPD severity: A- PCoA plots of three COPD severity groups based on weighted UniFrac distance measure for stable and exacerbation visit samples. B- Boxplots representing median, IQR with min, max whisker of between sample distance within visit types and within individual distance between stable and exacerbation sample segregated based on COPD severity.

4.3.4.3 Treatment

At exacerbation, patients were treated following the Anthonisen criteria (Table 1.2). In the SEFR set, 5 subjects received steroids alone, 6 antibiotics alone and the remainder both antibiotics and steroids. For the 19 subjects with samples for all four visits 3 received steroids alone, 5 antibiotics alone and the remainder had both antibiotic and steroid treatment. Between-group comparison of bacterial communities at the phylum level were analysed between the three treatment groups across all four visits. Within and between-treatment group changes in bacterial communities after treatment were compared at OTU (0.03) using data from exacerbation and recovery samples.

Subjects treated with both antibiotic and steroid had a higher proportion of Proteobacteria compared with the other treatment groups and a significantly lower proportion of Firmicutes compared to subjects treated with antibiotics only (p=0.039).
Within treatment groups, no significant change in microbial composition at phylum level was observed across visits (Figure 4.35).

**Figure 4.35**

Changes in relative abundance of phyla across SEFR visit types between treatment groups: Graph represents the mean relative proportion and SD for each of the visit type sample grouped based on their treatment received at exacerbation. "Others" include all the minor phylum groups together.

Exacerbation samples receiving both antibiotics and steroids were characterised by lower bacterial richness and uneven abundance compared with exacerbation samples of the other 2 treatment groups. Recovery samples had similar bacterial richness and dispersion pattern in all the groups (Table 4.12).

The weighted UniFrac distance measure identified antibiotic treated recovery visit type samples to be significantly less dissimilar in bacterial composition than recovery samples for the other two treatment groups (p = 0.0079). However there was no significant between-group difference for the within-subject change in bacterial composition from exacerbation to recovery (Figure 4.36).
### Table 4.12

#### A

<table>
<thead>
<tr>
<th>Metadata based analysis</th>
<th>Exacerbation</th>
<th>Recovery</th>
<th>Between visit group comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N=23</td>
<td>$S_{\text{est}}$</td>
<td>H</td>
</tr>
<tr>
<td>Steroids only</td>
<td>n=4</td>
<td>63.87 ± 2.86</td>
<td>8.94</td>
</tr>
<tr>
<td>Antibiotics only</td>
<td>n=5</td>
<td>67.1 ± 2.76</td>
<td>25.32</td>
</tr>
<tr>
<td>Antibiotics and steroid</td>
<td>n=14</td>
<td>37.69 ± 1.61</td>
<td>7.01</td>
</tr>
<tr>
<td>Between treatment group comparisons</td>
<td></td>
<td>0.176</td>
<td>0.179</td>
</tr>
</tbody>
</table>

#### B

<table>
<thead>
<tr>
<th>Metadata based analysis</th>
<th>Subjects</th>
<th>Exacerbation</th>
<th>Recovery</th>
<th>Within subject comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N=22</td>
<td>$S_{\text{est}}$</td>
<td>H</td>
<td>$S_{\text{est}}$</td>
</tr>
<tr>
<td>Steroids only</td>
<td>n=4</td>
<td>63.87 ± 2.86</td>
<td>8.94</td>
<td>0.34</td>
</tr>
<tr>
<td>Antibiotics only</td>
<td>n=5</td>
<td>67.10 ± 2.76</td>
<td>25.32</td>
<td>0.91</td>
</tr>
<tr>
<td>Antibiotics and steroid</td>
<td>n=13</td>
<td>39.49 ± 1.74</td>
<td>7.32</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Comparison of alpha diversity metrics based on treatment received post exacerbation: refer to legend for Table 4.10 for formatting details. Alpha diversity indices were compared in exacerbation and recovery visit type.
Comparison of between sample diversity measures based on treatment received at exacerbation: A- PCoA plots of three treatment groups based on weighted UniFrac distance measure for exacerbation and recovery visit samples. B- Boxplots representing median, IQR with min, max whisker of between sample distance within visit group and within individual distance between exacerbation and recovery sample segregated based on treatment groups.

4.3.5 *Streptococcus* OTU discrepancy - an observed artefact of the denoising algorithm

The UniFrac distance (section 1.3.3.2) is based on theory that phylogenetically similar groups have shared functions and therefore similar impact on the local environment. It is the standard distance measure metric reported in the literature (Eckburg et al., 2005; Ursell et al., 2012b). However, other distance metrics are available that compare the structure of the microbiome between samples, without making this assumption.

Two such distance metrics, the Bray-Curtis and Morisita-Horn distance measure (Horn, 1966) were utilised to construct PCoA plots. These identified Firmicutes dominant
clustering of samples for the exacerbation visit (Figure 4.37). This was attributable to the presence of a different *Streptococcus* OTU group at exacerbation (Table 4.13). OTU 275, the representative sequence of the exacerbation *Streptococcus* dominant OTU cluster, was present at a significantly higher proportion in exacerbation samples, while OTU 474 (representative sequence of the non-exacerbation *Streptococcus* dominant OTU cluster) was present at a higher proportion in non-exacerbation samples (Table 4.13).

**Figure 4.37**: Beta diversity PCoA plots of SEFR group using Bray-Curtis and Morisita-Horn

![PCoA plots](image)

**Table 4.13**: Comparison of OTUs between visit groups

<table>
<thead>
<tr>
<th>OTU</th>
<th>prob</th>
<th>Bonferroni_corrected</th>
<th>FDR_corrected</th>
<th>stable_mean</th>
<th>exacerbation_mean</th>
<th>followup_mean</th>
<th>Recovery_mean</th>
<th>Consensus Lineage</th>
</tr>
</thead>
<tbody>
<tr>
<td>275</td>
<td>0.0000</td>
<td><strong>0.0000</strong></td>
<td>0.0000</td>
<td>0.0003</td>
<td><strong>0.3326</strong></td>
<td>0.0015</td>
<td>0.0020</td>
<td>Firmicutes; <em>Streptococcus</em></td>
</tr>
<tr>
<td>474</td>
<td>0.0000</td>
<td><strong>0.0017</strong></td>
<td>0.0008</td>
<td>0.3736</td>
<td><strong>0.0045</strong></td>
<td>0.3378</td>
<td>0.2851</td>
<td>Firmicutes; <em>Streptococcus</em></td>
</tr>
</tbody>
</table>

To compare the sequence difference between these two OTUs, all the unique sequence reads and their proportional contribution to these two OTUs were examined. For exacerbation *Streptococcus* OTU cluster, OTU 275 was 1 of 7 unique reads and constituted 97% of the reads for that OTU, while three dominant reads constituting 98% of reads were identified for non-exacerbation *Streptococcus* dominant OTU cluster (Table 4.14). ClustalW alignment of the single dominant sequence of OTU 275 with the
three dominant reads for OTU 474 identified 4, 2 and 1 nucleotide differences respectively. All of the mismatches were in the V6, V7, V8 variable region but this was insufficient to form two OTUs at 0.03 cut-off level for ~500bp sequence. Performing blast on the representative sequences of the *Streptococcus* OTU from the non-exacerbation and exacerbation group came up with different *Streptococcus* members, belonging to the mitis group, as best matches. For non-exacerbation group *S. pneumoniae* was the best match, while *S. sanguinis* and *S. parasanguinis* produced the best matches for the exacerbation group.

Table 4.14: Number of unique reads and their distribution in the dominant *Streptococcus* OTUs

<table>
<thead>
<tr>
<th>Streptococcus OTU 474(Dominant in all visits except exacerbation)</th>
<th>Streptococcus OTU 275(exacerbation dominant)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique sequence read group</td>
<td>Representative sequence number</td>
</tr>
<tr>
<td>1</td>
<td>F1_37734</td>
</tr>
<tr>
<td>2</td>
<td>R18_27360</td>
</tr>
<tr>
<td>3</td>
<td>S7_23502</td>
</tr>
<tr>
<td>4</td>
<td>E14_29249</td>
</tr>
<tr>
<td>5</td>
<td>F2_25780</td>
</tr>
<tr>
<td>6</td>
<td>F19_71699</td>
</tr>
<tr>
<td>7</td>
<td>R11_66143</td>
</tr>
<tr>
<td>8</td>
<td>E10_10479</td>
</tr>
<tr>
<td>9</td>
<td>R10_48434</td>
</tr>
<tr>
<td>10</td>
<td>F9_68682</td>
</tr>
</tbody>
</table>

The two OTUs were formed as a consequence of extensive mismatches in the reverse primer and the 454 adapter region in 3’ end of the reads (Figure 4.38). Clustering of exacerbation samples did not occur when the sequencing error correction was not applied, suggesting this was probably an artefact generated from the application of the DeNoiser algorithm.

Two approaches were undertaken to examine if this was an artefact. To examine if *Streptococcus* species present were different in stable and exacerbation samples five paired stable and exacerbation samples from the SEFR sample set were utilised for cloning and comparing the sequence reads of *Streptococcus* specific *recA* (bacterial recombinase subunit) gene. Samples chosen had > 0.5 relative abundance of
S*reptococcus*, from microbiome analysis, in both stable and exacerbation samples. Plasmid DNA was extracted from ten clones per sample that were confirmed carrying the PCR product in the vector and sent for sequencing. Blast and DNA sequence similarity based tree results of these sequences showed no distinct difference in *Streptococcus* species between stable and exacerbation sample.

Secondly, a different but more computationally demanding sequencing error correction algorithm called AmpliconNoise was applied to stable and exacerbation visit samples. The remainder of the sequence analysis pipeline was kept identical to the previous analysis. PCoA plots based on Bray-Curtis or Morisita-Horn distance measure didn't show separate clustering of exacerbation samples after AmpliconNoise correction (Figure 4.39).
Alignment of the dominant reads from the two OTU groups: The four dominant sequence reads from the two *Streptococcus* OTU representative number and sequence reads alignment is shown in the above figure. Red highlighted sequence number is from the exacerbation dominant *Streptococcus* OTU. Sequence region highlighted in yellow is the reverse primer region.
Figure 4.39

PCoA plots using non-phylogenetic distance measure on Ampliconoise corrected samples: in both PCoA plot the red are exacerbation samples and blue are stable samples.

4.4 Discussion:

4.4.1 Sequencing plate design and pre-processing of reads

The 454 (FLX titanium series) produces an estimated 1 million reads on a whole chip. Partitioning the plate into quarters reduces this to half a million reads but the advantage is the ability to repeat the same set of unique barcodes across each of the four regions within the chip for tagging all the samples and thus, reducing the cost of the primers and the difficulty in standardising and optimising the PCR conditions with a greater number of unique primers.

For the first 454 chip, samples were divided into 4 regions with each region including all samples of a given visit time (S, E, F or R). Inter region variability, especially in the amount of reads produced, can be introduced due to small differences in normalisation and pooling of samples and amplifying these minor changes in the subsequent amplicon library PCR clonal step. Although 454 sequencing is largely used to study the relative abundance of bacterial groups in samples rather than absolute quantitative changes, in the following 454 plate longitudinal samples were placed in the same region to avoid inter-region variability factor between the longitudinal samples.
Out of 120 samples in the SEFR study, 11 samples produced either no sequence reads or <100 reads, suggesting a high probability that any reads generated were the result of sequencing errors. These were removed from sequence analysis. A further eight samples with <500 reads after sequence processing were also excluded from analyses of microbiome classification and diversity due to inadequate capture of microbial diversity and richness in these samples. The same exclusion criteria were also applied to the SEE and NES study samples. In the sequence read files about ~30% reads were contributed by host sequences and same proportion of reads did not meet the quality filtering requirement. This might be due to use of low stringency PCR conditions used and the primers not being eubacteria specific.

4.4.2 Taxonomic Characterisation of the bacterial community

At phylum level Proteobacteria, Firmicutes, Bacteroidetes and Actinobacteria formed the core microbiome of COPD samples in the study. This is in keeping with findings reported by other COPD microbiome studies (Cabrera-Rubio et al., 2012; Erb-Downward et al., 2011; Galiana et al., 2013; Pragman et al., 2012). Bacterial members of the Proteobacteria and the Firmicutes were most abundant across all the visit samples with no clear microbiome pattern differentiating the COPD samples between the visit types. Pragman and co-workers observed Actinobacteria as a dominant phylum with a mean relative abundance of 50% and contributing ≥ 40% of sequence reads in 64% of their COPD samples, while Sze et al reported Bacteroidetes group constituents to compose 25% of the COPD microbiome population in their samples (Pragman et al., 2012; Sze et al., 2012). These results are in contrast to the relatively low average relative abundance of these phyla (~5%) observed in this study. These differences may in part be attributable to the earlier studies using sputum samples from stable, non-exacerbating COPD subjects. However, we found no difference in microbiome characteristics for samples from non-exacerbating subjects or samples obtained at stable visits in exacerbating subjects. It is likely therefore that the microbiome in COPD is diverse at a population level and between-study differences only capture outcomes from small and selected COPD cohorts.

Although >100 genera were identified from the COPD samples >90% of the sequence reads were contributed by only 20 genera with Streptococcus, Haemophilus, Neisseria, Actinomycetes, Veillonella, Prevotella Pseudomonas, Lactobacillus, Gemella, Granulicatella, Fusobacteria being present in nearly 50 % of the samples. Most of these
genera have also been reported in previous studies as core components of the COPD microbiome (Cabrera-Rubio et al., 2012; Erb-Downward et al., 2011; Hilty et al., 2010). *Streptococcus* was present in all of the COPD samples and was the most abundant genus followed by *Haemophilus* and *Moraxella* (sporadically abundant component) from the Proteobacteria group. Sze et al (2012) reported significantly higher representation of *Lactobacillus* in GOLD stage 4 COPD patients compared to individuals without COPD. *Lactobacillus* was not a prominent bacterial group. In this study only four of the 110 successfully sequenced samples had *Lactobacillus* present at >10%, but these samples were not from GOLD stage 4 COPD subjects (three GOLD stage 3 and one GOLD stage 2).

Dominance of *Streptococcus*, *Haemophilus* and *Moraxella* in the sputum microbiome was consistent with earlier studies using culture- and PCR based methods that have shown bacterial species belonging to these groups to be frequently isolated in high numbers from COPD (Ball, 1995; Garcha et al., 2012; Sethi and Murphy, 2001). Correlation of sequence-based results for *Haemophilus* and *Moraxella* with the pathogen-specific qPCR and culture results for *H. influenzae* and *M. catarrhalis*, respectively suggests dominance of these individual species for the two Proteobacteria dominant genera in the sputum samples studied. In contrast there was little correlation between reads assigned to *Streptococcus* with *S. pneumoniae* culture and PCR results. In fact COPD samples showed a higher mean proportion of *Streptococcus* sequence reads in culture negative compared to culture positive samples for *S. pneumoniae*. This suggests that sequences represent *Streptococcal species* other than *S. pneumoniae* in COPD samples. Several Streptococcal species such as *S. mitis* and *S. oralis* form constituents of the oral microflora (Mager et al., 2003) and it is probable that COPD sputum samples had some of these oral contaminants represented in the sequencing results. This idea is supported by the results of hierarchical clustering of samples from SEFR set based on logit transformed microbial data. This gives more weightage to the less abundant and rarer constituents of the microbiome and showed *Streptococcus* dominant samples to be more diverse with comparatively higher abundance of bacterial species belonging to healthy oral (Aas et al., 2005) and respiratory microbiome (Charlson et al., 2011), compared with *Haemophilus* dominant samples. Whether greater diversity with oral streptococcal species represents salivary contamination of
sputum or translocation of these organisms into the airway in COPD, cannot be
determined.

Significant changes or differences in microbial composition at either phylum or genus
level were not observed in our COPD cohort, either for subjects during the time course
of an exacerbation in the SEFR set or over longer time periods comparing multiple
stable samples in the SSE set. In the latter, a relatively constant microbiome was
demonstrated for the subjects over periods of at least 6 months with most samples
clustering together within subjects than between subjects. The lack of an observable
shift in the microbiome during exacerbations is contrary to the idea that bacterial
infection plays an important role in the pathogenesis of COPD exacerbations. There are
a number of reasons for this finding. Firstly, COPD exacerbations are likely to be
heterogeneous, having both infectious and non-infectious causes (Gao et al., 2013; Papi
et al., 2006; Sethi et al., 2009). At a population level, a significant change in the
microbiome among the subgroup with exacerbation due to bacterial infection may not
be detected. It is also possible that analysis at the phylum or genus level may not be
sensitive enough to detect subtle changes and bacterial infection caused by species of
the same phylum or genus that is dominant at stable state will not be detected.

### 4.4.3 Sequence similarity based classification and diversity analysis of
COPD microbiota

Differences in microbiome composition within and between samples were determined
using bacterial groups (OTU) formed at a threshold of 0.03% dissimilarity for sequence
reads. At this threshold the microbiome is classified closer to species level (Drancourt et
al., 2000). Analysis of microbial composition at this level is likely to be more sensitive
for detecting changes in longitudinal composition. Diversity indices have been shown to
increase in proportion with the sequencing depth of a sample (Andrew and Mark, 1993).
To eliminate this bias, a threshold for equal sampling depth was set at a level such that
little additional information would be acquired beyond the chosen sequencing depth,
while retaining as many samples possible for analysis.

Comparison of alpha diversity measures based on species incidence between the four
visit times showed sputum collected at the onset of exacerbation to have the highest
species richness, although this didn't reach significance due to high standard deviation
within the group. Rarefaction curves of individual samples at exacerbation also revealed
considerable heterogeneity, with samples from some subjects showing little accumulation of species while other samples had steep increase in number of bacterial groups with increasing depth. Two way comparisons of the UniFrac distance measure by visit type showed significant dissimilarities between samples within the exacerbation visit compared with between sample dissimilarity within the remaining visit types. These results are in keeping with the idea that the microbiome at exacerbation is heterogeneous and coherent with our earlier published work on the heterogeneity of COPD exacerbations and their stratification based on clinical biomarkers (Bafadhel et al., 2011).

The alpha diversity indices H, 1-D and E_H that are based on bacterial dispersion pattern with or without species richness showed recovery visit samples to have the lowest indices, suggesting dominance by a fewer bacterial groups in most samples. Eleven out of 27 samples analysed from this visit showed >90% relative abundance of members of *Haemophilus, Moraxella* or *Streptococcus*. Ten of the eleven samples were dominated by *Haemophilus*. High abundance of a COPD pathogen in such large number of recovery samples 6 weeks post treatment might reflect the effect of antibiotic therapy at the time of exacerbation to lower overall bacterial diversity, although the lack of effect of such treatment on the dominant pathogen may indicate treatment failure.

Two samples or groups can have similar bacterial richness and dispersion patterns but consist of completely different bacterial constituents. To explore this between sample microbial differences PCoA plots based on abundance corrected and non-corrected for phylogenetic distance measures were used. These analyses didn't show any distinct microbiome pattern differentiating the samples on the basis of visit type. This was supported by PERMANOVA analysis which confirmed no significant difference between visit groups. These results are concordant with a recent cross sectional study that analysed 11 sputum samples at stable state and 10 samples at exacerbation state in COPD, identifying no significant difference in the microbiome between these two states (Tunney et al., 2013). In this study, abundance based PCoA plot revealed grouping samples based on *Streptococcus* or *Haemophilus* dominance with higher diversity in *Streptococcus* dominant samples. Higher diversity in Firmicutes (*Streptococcus*) dominant samples, compared with Proteobacteria (*Haemophilus, Pseudomonas*) dominant samples has also been reported by other COPD microbiomic groups (Erb-Downward et al., 2011; Huang et al., 2010; Sze et al., 2012).
Multiple stable samples collected at three time points showed similar bacterial richness and dispersion indices both at population and within subject analysis. In general, most individuals showed little change in microbiome composition across the three stable time points suggesting consistency of the microbiome over longer time periods and resistance to external factors. Dominance of the same bacterial group over longer periods is consistent with previous reports of persistence of bacterial groups (Murphy et al., 2004). It is also possible that the host may be prone to acquisition of a new strain or genus of the dominant bacterial group (Sethi et al., 2002). Persistence of *H. influenzae* and *S. pneumoniae* strains for long periods in COPD even after antibiotic treatment due to variation in antigenic composition of bacteria and their internalization into host cell has been reported (Groeneveld et al., 1990; Murphy et al., 2004; van Alphen et al., 1995). In *H. influenzae* variation in the antigenic composition of the outer membrane and in *S. pneumoniae* variation in capsular polysaccharide antigenic properties can affect the efficacy of antibody-mediated defense mechanisms against the bacteria (van Alphen et al., 1995). Biofilm formation has been suggested as another possible mechanism of persistence of COPD pathogens and evading antimicrobials (Legnani, 2009; Murphy and Kirkham, 2002).

Stable samples from non-exacerbating individuals were compared with stable samples from subjects that did exacerbate in the study to explore whether differences in the microbiome may be associated with risk of exacerbation. Although stable samples from non-exacerbating subjects had lower bacterial richness and diversity, this was not significant.

### 4.4.4 Metadata based bacterial community analysis

After finding no apparent changes in the microbiome across the four visit types across the study cohort, bacterial communities were characterised in stable and exacerbation samples in COPD subgroups stratified by smoking status and GOLD stage severity. Smoking is considered the most important factor for developing COPD. This may be, in part, due to changes in the airway microbiome induced by cigarette smoke (Sethi and Rochester, 2000). Miravitlles and colleagues showed smoking history was significantly associated with detection of PPB in COPD (Miravitlles et al., 2010). Alterations in the airway microbiome associated with ceasing smoking were therefore studied. At exacerbation there was a higher between sample microbiome dissimilarity within ex-smokers compared to smokers. This was due to higher number of *Haemophilus*
dominant samples in ex-smokers compared to smokers. Although, not significant, this trend was also seen for stable visit samples between the subgroups. In contrast with these findings, Charlson and co-workers investigating the microbiome of naso-oropharynx samples in healthy smokers and non-smokers observed greater heterogeneity in the microbiome of the smokers (Charlson et al., 2010). It should be noted that the smokers in that study did not have COPD. It is likely that the chronic airway injury and damage in COPD will influence the microbiome. The association between previous smoking and dominance of Proteobacteria and its biological relevance is also not clear and needs further investigation in a larger number of samples.

Previous studies investigating the relationship between the COPD microbiome and COPD severity have reported contradictory results. On the one hand Pragman et al (Pragman et al., 2012) and Sze et al (Sze et al., 2012) reported a higher alpha diversity index of the microbiome in severe COPD, compared with moderate COPD and healthy controls. In contrast, Hilty et al (Hilty et al., 2010) and Galiana et al. (Galiana et al., 2013) reported lower microbiome diversity in similar comparisons of severe COPD with moderate COPD and healthy controls.

In this study there was no consistent relationship of bacterial diversity with lung function. Neither stable nor exacerbation samples showed significant differences in alpha diversity between the three GOLD stages investigated. However, within subject changes in alpha diversity from stable to exacerbation stratified by COPD GOLD stage showed a significant increase in diversity for subjects with GOLD stage 4 disease. In contrast, subjects with GOLD stage 2 and 3 COPD showed a trend in the opposite direction with a lower bacterial diversity at exacerbation, although this was not statistically significant.

Looking at constituents of the microbiome according to COPD severity, no clear grouping of stable or exacerbation samples by severity group was observed in the study. Galiana and co-workers reported a significantly higher prevalence of Actinomyces in moderate, compared with severe COPD (Galiana et al., 2013). Although this study reported a similar pattern, with 21 of 23 (91%) samples from moderate COPD compared with 10 of 18 (55%) samples from severe COPD having Actinomyces, no association with disease severity was apparent as all samples from subjects with very severe COPD were positive. Pragman et al. reported clear differences in constituents of
the microbiome in different COPD severity stages. Firmicutes dominance was seen in severe COPD and Proteobacteria were dominant in moderate disease (Pragman et al., 2012). In this study no trends with respect to Proteobacteria or Firmicutes dominance were identified across the three GOLD stages. Overall, the results of this study add to the heterogeneity of outcomes for studies reporting on the COPD microbiome and no association between COPD severity and the microbiome can be made with the available evidence.

To investigate if the microbiome at exacerbation could be associated with the type of treatment received and changes in the bacterial composition post treatment, exacerbation and recovery samples were stratified by post-exacerbation treatment type. At exacerbation, samples receiving steroid alone showed high bacterial richness and least bacterial dissimilarity between samples. There was lower Proteobacteria abundance compared to the groups that received antibiotics alone and antibiotics along with steroid treatment. These observations suggest that exacerbations treated with steroid alone have a more homogenous microbiome. Whether this influences the clinical presentation at exacerbation (as episodes were treated according to Anthonisen criteria) requires further study in a larger number of samples. In this study, subjects treated with antibiotics had no significant change in bacterial richness or dispersion pattern at follow-up and recovery visits. Interestingly, recovery samples from subjects treated with antibiotics alone showed a significant decrease in bacterial community dissimilarity between samples suggesting a shift with treatment to a more homogenous Streptococcus dominant microbiome compared to other two treatment groups. The effect of antibiotic therapy to lower bacterial diversity has been reported by others (Dethlefsen et al., 2008; Huang et al., 2010). Huang et al. have reported a low detection rate of COPD pathogens such as H. influenzae and M. catarrhalis in 8 mechanically ventilated COPD exacerbating subjects with antibiotic treatment (Huang et al., 2010). This would imply effectiveness of antibiotics in COPD exacerbation while antibiotics treatment have been associated with various disease conditions, such as obesity, asthma, diabetes and certain forms of cancer (Beaugerie and Petit, 2004; Blaser, 2011).

Dethlefsen and co-workers in their gut microbiome study have highlighted on the negative effects that antibiotics can have on the protective commensals of the gut with some of these beneficial taxa failing to re-colonise long after the antibiotic treatment stopped (Dethlefsen et al., 2008). Whether the effects of antibiotics on the microbiome
observed in this study are beneficial is not clear. The effect of concomitant steroid therapy appears to reverse the effect of antibiotic treatment alone and this requires further study.

4.4.5 *Streptococcus* OTU discrepancy - an observed artefact of the denoising algorithm

Even after removal of reads with anomalous bases and quality filtering there are still many low abundance sequence reads that are a result of sequencing errors. OTU clustering thresholds of \( \geq 97\% \) can greatly overestimate the number of OTUs present (Kunin et al., 2010; Quince et al., 2009). QIIME provides two different algorithms: AmpliconNoise, an improved version of PyroNoise that utilises a probabilistic iteration based clustering on flowgram distance (Quince et al., 2011) and DeNoiser (Reeder and Knight, 2010), a modification of PyroNoise utilising an agglomerative clustering algorithm on flowgram distance for the sequencing artefact correction. DeNoiser, was chosen for correction of sequencing error (Reeder and Knight, 2010) in this study for being significantly less time and usable with the processing power of available desktop computers.

The average read lengths produced were 490 bases with most reads containing the reverse primer and the adaptor sequence but with lots of mismatches. Unlike AmpliconNoise where all read flowgrams are truncated at 400 bases prior to sequencing correction, DeNoiser algorithm allows the entire sequence read length to retain more information. For sequence correction the raw flowgram are utilised and therefore it is not compatible with the reverse primer truncated sequence read file and requires a more complicated truncation of flowgrams for generating output without the reverse primer sequence. Also till QIIME version1.4 there was no standalone script available for truncating reverse primer sequence in denoised reads. Generation of two *Streptococcus* OTUs (0.03) dominant between exacerbating and non-exacerbating was a result of the DeNoiser algorithm, with extensive mismatches in the adaptor and reverse primer sequence.

The *Streptococcus* OTU results indicate the need to be cautious of the pitfalls of microbiomic analysis. Incorrect inferences may be drawn if proper consideration is not given to the effects of using different pipelines for microbiome analysis, the impact of non-removal of reverse primer and adaptor sequence at 3’ end.
4.5 Conclusions

Compared to the gut and oral microbiome which have been extensively studied in health and disease, characterisation of the airway microbial community and its association with COPD disease has been explored only in relatively small samples and either at stable or exacerbation state. To date, this is the largest study of the airway microbiome in COPD with sample collection longitudinally in subjects at stable state and at the time of exacerbation. This has enabled examination of microbiomic changes that occur at exacerbation and post exacerbation with treatment. COPD samples in this study showed high prevalence of genera belonging to the following phyla: Firmicutes, Proteobacteria, Bacteroidetes and Actinobacteria with *Streptococcus*, *Haemophilus* and *Moraxella* being the most abundant groups. Good correlation between the qPCR and culture results of *H. influenzae* and *M. catarrhalis* with microbiome abundance of *Haemophilus* and *Moraxella* reinforces the likely association of these two Proteobacteria species with the airway microbiome of COPD, as previously reported using culturing techniques.

Although overall diversity of the microbiome was greatest at exacerbation, no distinct patterns in constituents of the microbiome were attributable to at any of the four visit times (S, E, F, R). There was no clear evidence to support a clear shift in the microbiome during exacerbation or after treatment. Clinical metadata based on GOLD stage severity and smoking status also failed to reveal any consistent features either for the population as a whole or for sequential analyses of individuals. Overall, wide variability between individuals in the extent of microbiome change across the four visit times accounted for the absence of any significant findings and supports the view that there is considerable heterogeneity in the pathogenesis of COPD exacerbations.
5 Comparison of SEFR samples based on the relative abundance of the 2 dominant phyla

5.1 Introduction:

In chapter (4) I report on the bacterial community profile in sputum at different clinical states of COPD. Both quantitative and qualitative methods confirmed abundance of *Haemophilus, Moraxella and Streptococcus* associated species across different clinical states, with no significant between-visit difference or within-subject change in either bacterial load or composition across the four visits. In particular, the lack of difference between exacerbation and stable state is contrary to the hypothesis that a measurable difference in the microbiome would be observed in bacterial exacerbations of COPD.

One possible explanation for this might be due to the heterogeneous nature of COPD exacerbations (Bafadhel et al., 2011; Gao et al., 2013; Papi et al., 2006). In our earlier paper, none of the host inflammatory markers tested showed increase in their level between stable and exacerbation (Bafadhel et al., 2011). This was considered a result of the heterogeneous nature of COPD and highlighted by performing cluster analysis on biomarker expression profile. COPD exacerbation phenotypes related to bacterial-predominance, viral-predominance and eosinophil-predominance were discovered based on this expression profile (Bafadhel et al., 2011).

Bacterial infections are likely to account for only a proportion of exacerbations. Furthermore the type of bacterial infection may vary between subjects. At a population level, this variability may mask the ability to identify any significant changes in the microbiome at the time of bacterial exacerbation. This has been illustrated by Carmody and co-workers in a Cystic Fibrosis study that characterized and compared the microbiota in 68 paired stable and exacerbation samples. The study was performed to identify changes in microbiome at exacerbation and the bacterial groups associated with this change. The authors didn't find any microbial markers of change at exacerbation when all the samples were included together in the analysis. However, associations between the microbiome and exacerbation state were identified when samples were stratified by *Pseudomonas* dominance (Carmody et al., 2013).
To address the heterogeneity of COPD exacerbations, subgroups were characterised according to their sputum microbiome at each of the three visits of an exacerbation event and the preceding stable visit. It was hypothesised that exacerbations associated with bacterial infection would exhibit greater change in microbial composition across visits. Two approaches were used to classify exacerbations into subgroups:

1. Objective / unbiased approach: Cluster modelling of samples on the basis of their Proteobacteria:Firmicutes (P:F) ratio:
   Previous studies have demonstrated the value of characterising microbiomic data on the basis of the relative proportions of dominant groups. This may be a useful approach to data reduction, enabling meaningful analysis of the association between the microbiome and a biological outcome. One example is a study by Ley and colleagues in which the authors showed that alteration in the relative composition of Firmicutes and Bacteroidetes, two dominant phyla in gut, are key determinants of obesity (Ley et al., 2006). In this study, Proteobacteria and Firmicutes were the dominant phyla identified in the airway microbiome of COPD. It was hypothesized that a shift in the relative proportions of these two phyla may be representative of between-visit shifts in the microbiome during exacerbation, removing noise arising from minority communities.

2. Subjective approach: Grouping on the basis of within-subject change in the microbiome between visits:
   In this approach, subjects were grouped according to the patterns and magnitude of change in their microbiome PCoA pattern across stable, exacerbation and recovery visits.

Groups characterised by these two methods were compared with available clinical markers of bacterial infection to see whether observed group-specific patterns of microbiome changes related to clinical changes. In particular, with the group showing disturbance in microbiome between stable and exacerbation and restoration of this change at recovery, in this regard.
5.2 Materials and Methods

5.2.1 Study Sample
In the SEFR microbiome set 19 subjects with all four visit samples microbiomic profiles available were used for the above mentioned grouping analysis. QPCR based enumeration of Proteobacteria and Firmicutes copies were performed on all 120 samples from SEFR set.

5.2.2 qPCR based enumeration of Proteobacteria and Firmicutes

5.2.2.1 DNA standards for quantification assays
Pure culture DNA from *H. influenzae* and *S. pneumoniae* were used as standards for Proteobacteria and Firmicutes qPCR assay, respectively. The number of 16S gene copies/μl present in the pure culture DNA was determined by multiplying 6 and 4 to the genome equivalent /μl for *H. influenzae* and *S. pneumoniae*, respectively. Finally 10 fold serial dilutions were made ranging from 10⁷ to 10² genome/μl for both bacteria.

5.2.2.2 Primers
Development and optimisation of the qPCR assays to determine the P:F ratio is described in detail in results section 5.3.2.1. Briefly, all available 16S rDNA based oligo sequences (Table 5.3) utilised for enumerating the two phyla in previous studies were selected and analysed by in-silico PCR (Table 5.4) to choose the primer pairs with the best combination of sensitivity and specificity values based on reference database matches. Sensitivity and specificity of the primer pairs selected were also was further examined by performing qPCR assays in varying proportions of pure culture mixtures of *Streptococcus* and *Haemophilus*, the two dominant genera from the two phylum groups.

The DNA sequence of the primers used for quantifying 16S rDNA copies of Proteobacteria and Firmicutes in samples are presented in Table 5.1. To construct 16S rDNA primers conserved across the entire Proteobacteria phylum and at the same time non-specific to other phyla was difficult. Within the Proteobacteria phylum, the dominant bacterial groups belonged to the Gammaproteobacteria class therefore, primers targeting this class were utilised for the Proteobacteria enumeration assay.
Table 5.1: 16S rDNA primers utilised for Proteobacteria and Firmicutes qPCR

<table>
<thead>
<tr>
<th>Primer</th>
<th>Target Organism (gene)</th>
<th>Sequence</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y871F</td>
<td>Gammaproteobacteria</td>
<td>5′TAAGTHGACCGCCTGGGGGAGT 3′</td>
<td>(Mühling et al., 2008)</td>
</tr>
<tr>
<td>Y1202R</td>
<td>targeting 16S gene</td>
<td>5′CGTAAAGGGGCATGAIG 3′</td>
<td>(Bacchetti De Gregoris et al., 2011)</td>
</tr>
<tr>
<td>Firm928F</td>
<td>Firmicutes group</td>
<td>5′TGAAACTYAAAGGAAATTGACG 3′</td>
<td>(Bacchetti De Gregoris et al., 2011)</td>
</tr>
<tr>
<td>Firm1040R</td>
<td>targeting 16S gene</td>
<td>5′ACCATGCACCACCTGTGC 3′</td>
<td>(Bacchetti De Gregoris et al., 2011)</td>
</tr>
</tbody>
</table>

5.2.2.3 Quantification Assays
SYBR green assay chemistry was utilised to determine Proteobacteria and Firmicutes copies. Reactions for both assays were prepared with 12.5μl 2xAbsolute qPCR SYBR Green mix, 9.5μl water with 1μl of each Yproteo871F and Yproteo1202R primers at a final concentration of 200nM for Proteobacteria and 1μl of each Firm928F and Firm1040R primers at a final concentration of 150nM for Firmicutes. 1μl of template DNA was added making the final reaction volume to 25μl.

Samples were quantified in duplicates simultaneously for both bacterial groups on Roto-Gene 6000 (Corbett Life Sciences, UK). Each run also contained both sets of standards and negative controls. The qPCR protocol began with initial denaturation at 95°C for 15 minute followed by 40 cycles of amplification step, which was as follows

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Time</th>
<th>Type of Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>95°C</td>
<td>20secs</td>
<td>Denaturation</td>
</tr>
<tr>
<td>60°C</td>
<td>30secs</td>
<td>Annealing</td>
</tr>
<tr>
<td>72°C</td>
<td>20secs</td>
<td>Elongation</td>
</tr>
<tr>
<td>78°C</td>
<td>20secs</td>
<td>Data acquisition green</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(470nm/510nm) channel</td>
</tr>
</tbody>
</table>

Finally melt (T_m) curve analysis was performed from 65°C to 99°C to determine the specificity of the amplicons.
Comparison of SEFR samples based on the relative abundance of the 2 dominant phyla

All real time PCR were performed on Roto-Gene 6000 (Corbett Life Science, UK). qPCR signals were analysed with the Roto-Gene 6000 Series Software1.7. The data collected was normalized to compensate for background fluorescence and a threshold level set at which the data is analysed. Samples from runs with correlation coefficients ($R^2$) and amplification efficiencies above 0.98 and 0.7 and less than 1 Ct difference between replicates were used for further analysis (Dorak, 2006).

5.2.3 Statistical analyses for sequencing and qPCR data
R statistical programming language was used for performing clustering analysis, producing heatmaps, bootstrap analysis on cluster, Rand index analysis and PCA analysis. Hierarchical clustering using Euclidean distance measure and average clustering method was performed for clustering of samples based on Proteobacteria and Firmicutes ratio. Bootstrapping analysis to test the significance of membership to each cluster was done using Pvclust (Suzuki and Shimodaira, 2006) in R. Rand index (Rand, 1971) analysis was performed to test the agreement of cluster structure and membership between the two methods.

As described in section (4.2.4) paired and unpaired student t test; Anova RM-Anova for parametric continuous data were utilised for between and within group analysis followed by Tukey's HSD procedure for pair-wise comparisons between the groups in Prism. 2-D and 3-D PcoA plots of samples based on OTU (0.03) classification were made in Prism and QIIME respectively.

5.3 Results

5.3.1 Stratification of samples based on two dominant phyla

5.3.1.1 454 sequencing derived Proteobacteria and Firmicutes P:F ratio cluster model
As reported in earlier chapter (4), bacterial community analysis did not identify any consistent association between the microbiome and change in COPD state of exacerbating subjects. Most of the COPD samples were dominated either by Firmicutes or Proteobacteria. Therefore variation in relative abundance of Proteobacteria and Firmicutes was analysed to explore changes in the pattern of relative dominance of these two phylum groups. Hierarchical clustering utilising the P:F ratio was performed
on the 19 COPD subjects that had sequencing profile at all four visit times from the SEFR set.

Three clusters were identified (Figure 5.1):

Cluster 1: High Firmicutes (HF)

Subjects in this cluster were characterised by samples dominated by bacteria of the Firmicutes phylum.

Cluster 2: Proteobacteria-Firmicutes group (PF)

Subjects in this cluster were characterised by samples with a mixed but significant representation for both phyla.

Cluster 3: High Proteobacteria (HP)

Subjects in this cluster were characterised by samples dominated by bacteria of the Proteobacteria phylum.

Bootstrapping, an iteration based re-sampling technique, was performed to validate the cluster model (Suzuki and Shimodaira, 2006). There was strong agreement in cluster structure and membership with the exception of 2 subjects that were high Proteobacteria outliers. These were not included in further analysis (Figure 5.2).
Comparison of SEFR samples based on the relative abundance of the 2 dominant phyla

**Figure 5.1**

Classification of samples by ratio of relative abundance of Proteobacteria to Firmicutes: The heatmap is based on the log$_2$ transformed ratio of relative abundance of Proteobacteria to Firmicutes. Transition of colours from blue to yellow represents change from a higher proportion of Firmicutes to a higher proportion of Proteobacteria. Rows represent samples from a visit type and columns samples from each individual. The hierarchical cluster model is generated using Euclidean distance measure.

**Figure 5.2**

The three P:F clusters and their membership confirmed by bootstrapping test: analysis performed by pvclust package in R. Red colour number on the clusters are AU (approximately un-biased) p values.
Comparison of SEFR samples based on the relative abundance of the 2 dominant phyla for each cluster. Clusters with AU larger than 95% are highlighted by rectangles, which are strongly supported by data.

5.3.1.2 Comparison of the P:F ratio between clusters in individual visit groups

Between cluster comparisons for individual visit group were performed to examine differences in the P:F ratio and dissimilarity in microbiome composition between samples at OTU (0.03) level at each visit. With the exception of the follow-up visit, the P:F ratio differed significantly between the three clusters at each visit (stable p=0.022, exacerbation p < 0.0001, recovery p < 0.0001) Figure 5.3. Overall, variance of P:F ratios was greatest across groups for samples of the exacerbation visit. The HF group was characterised by a low a P:F ratio for samples at all visits. This was significantly lower than the PF cluster at stable (p<0.05), exacerbation (p<0.05) and recovery (p<0.001) visits and also significantly lower than the HP (p<0.0001) for exacerbation samples. Exacerbation visit was characterised by a very high P:F ratio in HP cluster and in contrast, the PF cluster exhibited a significantly higher P:F ratio at recovery, compared with the other two clusters.

These findings at phylum level were compared with grouping of samples based on dissimilarity in microbiome composition between samples at OTU (0.03) level at each visit. PCoA plots using the weighted Unifrac distance of OTUs at each visit demonstrated concordance with the patterns of clustering using the P:F ratio (Figure 5.3). Microbiome richness (Chao) or distribution pattern (H) didn't show any significant difference between the three clusters.
Comparison of SEFR samples based on the relative abundance of the 2 dominant phyla

**Figure 5.3**

![Graph showing comparison of SEFR samples based on relative abundance of 2 dominant phyla](image)

**Changes in P:F ratio between clusters and their concordance with OTU (0.03) level microbial composition:**

- **A** - shows the mean (±SD) of individual clusters within each visit of SEFR samples.
- **B** - shows the PCoA plot of each visit type derived from weighted UniFrac distance measure of OTUs (0.03 level). Samples closer to PC2 axis represent the Firmicutes dominant and samples moving upwards along the PC1 axis are Proteobacteria dominant samples. In exacerbation visit PCoA plot, HP cluster samples (highlighted with dotted white circle) are tightly clustered near the Proteobacteria dominant region separate from the other two Firmicutes dominated cluster while in recovery it is the PF cluster samples that is closer to Proteobacteria dominant region.

5.3.1.3 **Within-Cluster changes in the microbiome across visits**

There was no significant change in the P:F ratio across the four visit types within subject but the three P:F clusters exhibited different patterns of change across visits during exacerbation. For the HP (p=0.021) and PF (p=0.015) clusters, the P:F ratio changed significantly across the visits. In the HP cluster, exacerbation samples had a significantly higher P:F ratio than samples of the follow-up (p<0.05) and recovery (p<0.05) visits. In the PF cluster, the P:F ratio of samples at the recovery visit were significantly higher than for exacerbation (p<0.01) and follow-up (p<0.05) visits. In
contrast, the HF cluster showed no significant change in P:F across any visits (Figure 5.4).

**Figure 5.4**

![Bar chart showing change in P:F ratio across 4 visits within each cluster](image)

*Change in the P:F ratio across 4 visits within each cluster:* chart represents mean and ±SD in each visit group within P:F clusters

Corresponding subject specific plots at OTU level identified similar patterns of change in the microbiome structure across visits for each cluster (Figure 5.5). As observed for the P:F ratios, samples from subjects of the HF cluster grouped closely together across the four visits, except for samples E12 and F16. In contrast, there were much larger shifts in the microbiome at OTU level across quadrants defined by the PCoA axes, for subjects of both the PF and HP clusters. For the PF cluster, recovery samples grouped closely while exacerbation samples were similarly close in the HP cluster. Recovery samples of the PF cluster showed a significantly lower Shannon diversity index than exacerbation samples of that cluster but otherwise diversity was not significantly different across visits within each cluster.
Comparison of SEFR samples based on the relative abundance of the 2 dominant phyla

5.3.1.4 Baseline clinical characteristics of the three P:F clusters

There were no significant differences in any baseline clinical characteristics of COPD subjects clustered by the P:F ratio. This is in part likely to be due to the small sample size included. Overall, subjects of the HP cluster had a lower BMI and higher percent sputum neutrophil count compared with the other two clusters (Table 5.2).

Interestingly, although treatment given at exacerbation showed no difference between clusters, all subjects not prescribed antibiotics at exacerbation were members of the PF cluster and showed an increase in their P:F ratio at recovery.
### Table 5.2

<table>
<thead>
<tr>
<th>Metadata Groups</th>
<th>HP (N=3)</th>
<th>PF (N=8)</th>
<th>HF (N=6)</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GOLD Stage</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1 (33.33)</td>
<td>3 (37.5)</td>
<td>4 (66.67)</td>
<td>0.87</td>
</tr>
<tr>
<td>3</td>
<td>2 (66.67)</td>
<td>2 (25)</td>
<td>2 (33.33)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>3 (37.5)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male / N (%)</td>
<td>3 (100)</td>
<td>7 (87.5)</td>
<td>4 (67)</td>
<td>0.41</td>
</tr>
<tr>
<td><strong>Smoking status</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ex-smoker / N (%)</td>
<td>3 (100)</td>
<td>5 (62.5)</td>
<td>4 (67)</td>
<td>0.46</td>
</tr>
<tr>
<td><strong>Type of treatment</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antibiotic &amp; Steroid</td>
<td>2 (67)</td>
<td>4 (50)</td>
<td>3 (50)</td>
<td>0.27</td>
</tr>
<tr>
<td>Antibiotic alone</td>
<td>1 (33)</td>
<td>1 (12.5)</td>
<td>3 (50)</td>
<td></td>
</tr>
<tr>
<td>Steroid alone</td>
<td>0</td>
<td>3 (37.5)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><strong>BMI kg/m²</strong></td>
<td>22.68</td>
<td>25.44</td>
<td>25.90</td>
<td>0.83</td>
</tr>
<tr>
<td>GM % sputum eosinophils</td>
<td>0.7</td>
<td>0.8</td>
<td>1.0</td>
<td>0.93</td>
</tr>
<tr>
<td><strong>% sputum neutrophils</strong></td>
<td>86.33</td>
<td>65.25</td>
<td>67.22</td>
<td>0.23</td>
</tr>
<tr>
<td>Exacerbations in previous 12 months</td>
<td>4.33</td>
<td>3.00</td>
<td>5.86</td>
<td>0.283</td>
</tr>
<tr>
<td>Post BD FEV1 (%) predicted</td>
<td>50.30</td>
<td>49.00</td>
<td>62.63</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Comparison of clinical characteristics between the P:F derived microbiome clusters: "N" denotes the total number of samples in a cluster. Numbers in brackets represent in percentage the number of samples for each parameter.

### 5.3.2 Validating P:F microbiome cluster by qPCR

P:F ratios derived from 454 sequencing data yielded some interesting results that showed differences between subgroups during different clinical COPD states. However, analyses were limited by the relatively small sample size of 19 subjects. Quantitative PCR is a much cheaper and quicker method compared to sequencing that can provide information about the P:F ratio in sputum. It was hypothesised that qPCR would give data that strongly correlated with 454 sequence analysis and this would enable examination of a larger proportion of the available SEFR set.

#### 5.3.2.1 Developing 16S rDNA based Proteobacteria and Firmicutes qPCR

A Proteobacteria and Firmicutes quantitative PCR for deriving (P:F) in COPD sample was developed. Although there are several 16S primer pairs that are specific for the Firmicutes phylum, no 16S primer pairs exist that are sufficiently specific and sensitive to target the entire phylum Proteobacteria. Earlier analyses of the sample set identified bacteria belonging to *Haemophilus*, *Moraxella* and *Pseudomonas* genera to be the most
dominant Proteobacteria. All belong to Gammaproteobacteria (Y’proteo). Therefore, 16S primers targeting the Y’proteo class were selected for qPCR (Table 5.3).

To choose primer pairs with optimum sensitivity and specificity for both bacterial groups in-silico check against the RDP database and in vitro qPCR assay on pure culture was performed. All but one of the Y’proteo primer pairs listed (Table 5.3) generated very long amplicons, not suitable for qPCR. Therefore, oligo-sequence Y’proteo871F, Y’proteo946F and Y’proteo1080F were used as forward primers with Y’proteo1202R, producing products of size 331bp, 256bp and 122bp respectively. For detection of Firmicutes, the three listed primer pairs (Table 5.3) generating <200bp were analysed.

**Table 5.3 16S primer pairs targeting Y’proteo and Firmicutes:**

<table>
<thead>
<tr>
<th>Primer</th>
<th>Target group</th>
<th>Primer sequence(5’-3’)</th>
<th>Amplicon size(’bp)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y’proteo 395F</td>
<td>Gammaproteobacteria (16S</td>
<td>OMATGCGCGGTGTGGTGA</td>
<td>476</td>
<td>(Mühling et al., 2008)</td>
</tr>
<tr>
<td>Y’proteo 871R</td>
<td>gene)</td>
<td>ACCCCCAGGGCGGTCDATCTA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y’proteo 1080F</td>
<td></td>
<td>TGCTAGCCTGCTGTGTTGGA</td>
<td>122</td>
<td>(Bacchetti De Gregoris et al., 2011)</td>
</tr>
<tr>
<td>Y’proteo 1202R</td>
<td></td>
<td>CGTATGGGCCCATGAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y’proteo 382F</td>
<td></td>
<td>AGGCAGCGAGGGGAAATA</td>
<td>981,417</td>
<td>(Kleijn et al., 2007)</td>
</tr>
<tr>
<td>Y’proteo 946F</td>
<td></td>
<td>AAATGAATGACGAGGGGCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y’proteo 1363R</td>
<td></td>
<td>ACTAGGCAAGTCCGACTTCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firm 350F</td>
<td>Firmicutes (16S</td>
<td>GGCAGCAGTRGGGAARCTTCA</td>
<td>464</td>
<td>(Mühling et al., 2008)</td>
</tr>
<tr>
<td>Firm 814R</td>
<td>gene)</td>
<td>AACTGATYAAAGGAATGGACG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firm 929F</td>
<td></td>
<td>TGAACCTAAAGGAATGGACG</td>
<td>112</td>
<td>(Bacchetti De Gregoris et al., 2011)</td>
</tr>
<tr>
<td>Firm 1040R</td>
<td></td>
<td>ACCATGACACATGTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firm 934F</td>
<td></td>
<td>GGAGYATGTTGTTTAAATTCGAGCA</td>
<td>126</td>
<td>(Guo et al., 2008)</td>
</tr>
<tr>
<td>Firm 1060R</td>
<td></td>
<td>AGCTAGCAGACACATGAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firm probe</td>
<td></td>
<td>GTCGACCTGCGTTGGCTGTA</td>
<td>70</td>
<td>(Armougom et al., 2009)</td>
</tr>
<tr>
<td>Firm 1224R</td>
<td></td>
<td>CAATTTCAACAGGTCTGTT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firm probe</td>
<td></td>
<td>GTCACACTCAGATGCC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**16S primer pairs targeting Y’proteo and Firmicutes:** Primers beginning with Firm are Firmicutes phylum specific oligos and the numbers represent their position in E. coli 16s gene while, Y’proteo are Y’proteo group specific primers.

In-silico sensitivity of Firmicutes 16S oligo sequences was determined as the proportion of matches with the Firmicutes, specifically *Streptococcus* in the RDP database and for Y’proteo oligo sequences the proportion of matches produced with Y’proteo class
Comparison of SEFR samples based on the relative abundance of the 2 dominant phyla bacteria *Haemophilus*, *Moraxella* and *Pseudomonas*. Specificity of Firmicutes and Yproteo oligos was measured as the proportion of matches with Firmicutes and Yproteo sequences respectively, against the total of 16S bacterial matches. The proportion of non-specific cross-matches of Firmicutes primers with the three dominant Yproteobacteria and Yproteo oligos with *Streptococcus* was also examined (Table 5.4).

Among the Firmicutes targeting oligos, Firm934F had fewer matches and poorer sensitivity for the target group. Firm1063F, Firm1060R and Firm1224R all had good sensitivity but exhibited poor specificity, producing matches with Yproteo group as well. For the Yproteo oligos, Yproteo1080F had very low specificity and produced a high percentage matches with *Streptococcus*. On this basis, Firmicutes primer pair (928F, 1040R) and Yproteo (871F, 1202R) were chosen for SYBR green qPCR assay (Table 5.4).

**Table 5.4: Insilico sensitivity and specificity check of Firmicutes and Yproteo oligos**

<table>
<thead>
<tr>
<th>Primer</th>
<th>Target Phylum</th>
<th>Sensitivity [hits in the cart (%)]</th>
<th>Specificity [Proportion hits inside cart against total (16S) bacterial hits (%)]</th>
<th>% GammaProteobacteria sequences (RDP)</th>
<th>% Streptococcus sequences (RDP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Firm 928F</td>
<td>Firmicutes (RDP-Firmicutes 440938 sequences downloaded in the cart)</td>
<td>64.78</td>
<td>41.23</td>
<td>5.3 [Moraxella 0.04%, Pseudomonas (0.05%)]</td>
<td>63</td>
</tr>
<tr>
<td>Firm 1040R</td>
<td>Firmicutes (RDP-Firmicutes 440938 sequences downloaded in the cart)</td>
<td>57.91</td>
<td>72.46</td>
<td>0.08</td>
<td>47</td>
</tr>
<tr>
<td>Firm 934F</td>
<td>Firmicutes (RDP-Firmicutes 440938 sequences downloaded in the cart)</td>
<td>55.03</td>
<td>54.63</td>
<td>14 [Pseudomonas 65%]</td>
<td>53</td>
</tr>
<tr>
<td>Firm 1060R</td>
<td>Firmicutes (RDP-Firmicutes 440938 sequences downloaded in the cart)</td>
<td>61.67</td>
<td>58.69</td>
<td>0.09</td>
<td>47</td>
</tr>
<tr>
<td>Firm 1063F</td>
<td>Firmicutes (RDP-Firmicutes 440938 sequences downloaded in the cart)</td>
<td>65.9</td>
<td>27.7</td>
<td>40 [Haemophilus 1.5%, Moraxella 80%, Pseudomonas 67%]</td>
<td>47</td>
</tr>
<tr>
<td>Firm 1224R</td>
<td>Firmicutes (RDP-Firmicutes 440938 sequences downloaded in the cart)</td>
<td>64.25</td>
<td>33.28</td>
<td>58 [Haemophilus 54%, Moraxella 86%, Pseudomonas 52%]</td>
<td>46</td>
</tr>
<tr>
<td>Firm probe</td>
<td></td>
<td>63.88</td>
<td>48.5</td>
<td>0.18</td>
<td>46</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Primer</th>
<th>Target Phylum</th>
<th>Sensitivity [hits in the cart (%)]</th>
<th>Specificity [Proportion hits inside cart against total (16S) bacterial hits (%)]</th>
<th>% Firmicutes seq (RDP)</th>
<th>% Haemophilus, Moraxella, Pseudomonas (RDP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yproteo 871F</td>
<td>Yproteo bacteria (RDP-Yproteo bacteria 215820 sequences downloaded in the cart)</td>
<td>42.35</td>
<td>80.88</td>
<td>0.01</td>
<td>Haemophilus (90), Moraxella (93), Pseudomonas (57)</td>
</tr>
<tr>
<td>Yproteo 1060F</td>
<td>Yproteo bacteria (RDP-Yproteo bacteria 215820 sequences downloaded in the cart)</td>
<td>60.5</td>
<td>11.77</td>
<td>66 [Streptococcus 46%]</td>
<td>Haemophilus (55), Moraxella (86), Pseudomonas (87)</td>
</tr>
<tr>
<td>Yproteo 946F</td>
<td>Yproteo bacteria (RDP-Yproteo bacteria 215820 sequences downloaded in the cart)</td>
<td>46.38</td>
<td>77.12</td>
<td>0.04</td>
<td>Haemophilus (67), Moraxella (89), Pseudomonas (55)</td>
</tr>
<tr>
<td>Yproteo 1202R</td>
<td>Yproteo bacteria (RDP-Yproteo bacteria 215820 sequences downloaded in the cart)</td>
<td>51.32</td>
<td>54.24</td>
<td>0.04</td>
<td>Haemophilus (54), Moraxella (86), Pseudomonas (55)</td>
</tr>
</tbody>
</table>
As discussed in previous chapters, *Haemophilus* and *Streptococcus* are two of the most abundant groups frequently detected in sputum samples from COPD subjects. Therefore, sensitivity for the target group was further examined for the chosen primer pairs by performing qPCR assays in varying proportions of pure culture mixtures of *Streptococcus* and *Haemophilus*. Serial dilutions of pure cultures till $10^{-4}$ of the target group, along with the neat non-target group were used to make culture mixes (Figure 5.6).

Both primer sets retained specificity when the non-target group was present at a $\log_{10}$ threefold higher concentration than the target group in mixture, with little non-specific amplification (Figure 5.6). Overall, $\gamma$proteo primers showed one $\log_{10}$ fold lower readings for *Haemophilus* than expected.

**Figure 5.6**

![Performance of Firmicutes and $\gamma$proteo primer pairs in qPCR of mixed cultures](image)

Performance of Firmicutes and $\gamma$proteo primer pairs in qPCR of mixed cultures: Pure culture DNA, standardized at $10^7$ copies/$\mu$l, of *S. pneumoniae* (SP) and *H. influenzae* (HI) were used for dilutions and control for both qPCR assay. For the Firmicutes assay SP was the target bacterium and HI the non-target and vice versa for the $\gamma$proteo assay. In both assays the target bacterium was diluted in equal proportion
Comparison of SEFR samples based on the relative abundance of the 2 dominant phyla and in 10 fold dilution series up to $10^{-4}$ with the neat non-target bacterial DNA. Diluted target DNA was also used as control in the assays. X axis shows the name of the culture and dilution. The bars in red shows the calculated readings of target DNA in undiluted and in dilution mixes with non-target DNA and, undiluted non-target DNA Blue bars represent the observed readings of the target and non-target DNA in these assays.

5.3.2.2 Correlation and validation of qPCR derived P:F ratio

The optimised qPCR was performed in the 30 COPD subjects of the SEFR set to determine whether P:F ratios and subsequent clustering of subjects in this way was comparable to outcomes with 454 sequencing.

The P:F ratio of all 120 samples of the SEFR set were calculated with qPCR, using values for copies/μl obtained from the ϒproteo and Firmicutes assays. There was a strong and significant correlation between these qPCR derived P:F ratios and the ratios obtained in the 110 samples with available 454 sequence data ($R^2 = 0.728$, p <0.0001) (Figure 5.7).

Hierarchical clustering of qPCR derived data for the P:F ratio in the 19 subjects clustered with 454 sequencing data yielded a 3-cluster model that showed very good agreement with the model obtained using 454-sequencing data (Rand index 0.79) (Figure 5.8). Further analysis to characterise discrepant samples was performed with PCA analysis of qPCR derived P:F ratios. There were 3 subjects (13, 17 and 15) that differed in their cluster membership between the methods (Figure 5.9). All three subjects were assigned to the PF cluster derived from 454 sequencing and HP cluster with qPCR.
Comparison of SEFR samples based on the relative abundance of the 2 dominant phyla

**Figure 5.7**

![Scatter plot graph showing correlation between 454 and qPCR derived P:F ratio](image)

Correlation between 454 and qPCR for P:F ratio: The scatter plot graph shows correlation between the 454 and qPCR derived P:F ratio with the best fit line (black line) and 95% CI (dotted lines). X axis represents the 454 P:F ratio and Y axis qPCR derived P:F ratio. Each sample is represented as squares.

**Figure 5.8**

![Hierarchical clustering](image)

Comparison of cluster models based on P:F ratio obtained by qPCR and microbiomics: Hierarchical clustering was performed using Euclidean measure and average distance method.
Comparison of SEFR samples based on the relative abundance of the 2 dominant phyla

Figure 5.9

Grouping of qPCR derived P:F ratio subjects labelled according to 454 P:F clusters: PCA plot based on qPCR ratio and subjects labelled according to the cluster they belonged to in the 454 P:F cluster

5.3.2.3 Clustering of all 30 subjects based on qPCR P:F ratio

With qPCR, a complete dataset of the P:F ratio was obtained for all four visit times in each of the 30 subjects of the SEFR set. Hierarchical clustering of these samples yielded a 3-cluster model (Figure 5.10). Bootstrapping confirmed significant membership of samples to individual clusters validating the cluster model except for one sample that didn't fall in any of the three clusters and was excluded from further analysis. The clusters formed described the same patterns identified previously with 454 sequencing data and were accordingly named

HF cluster - characterised by a low a P:F ratio for samples at all visits with significantly lower Proteobacteria abundance than the PF cluster at exacerbation (p < 0.05), follow-up (p<0.01) and recovery (p<0.001) visits and the HP cluster (p<0.001) for exacerbation samples.

HP cluster - characterised by significantly higher P:F ratio at exacerbation visit compared to PF cluster (p<0.05) and HF cluster (p<0.001)
**PF cluster** - characterised by significantly higher P:F ratio at both Follow-up and recovery visit compared to the other two cluster.

The HF cluster was the largest group (n=18) with 10 of the 11 subjects excluded from the 454-sequencing analysis being assigned to this group. Both HF and HP clusters retained membership of all the subjects from the earlier 454 sequence based cluster model. In addition the qPCR HP cluster included two further subjects - one was initially excluded from the 454-sequencing output due to low read numbers and the other (Patient 15) had been assigned to the PF cluster with 454 sequencing. A further two patients of the 454-PF cluster (13 and 17) were assigned to the HF cluster in the qPCR P:F cluster model.

Figure 5.10
Clustering of 30 subjects based on qPCR P:F ratio: A- the heatmap representing the clustering of 30 subjects based on qPCR derived ratio. B- Bootstrapping analysis to test the significance of the cluster model. Red boxes represent the three significant clusters with >95% value.

5.3.2.4 Changes in the qPCR based P:F ratio across visits in 30 subjects

Similar to the findings with 454-sequence based data, no significant change in the P:F ratio was observed across visits at a population level. However significant between-visit changes were observed within each cluster that was comparable with those observed for the corresponding 454-sequencing clusters (Figure 5.11). In the HP cluster, exacerbation samples had a significantly higher P:F ratio than samples at the other three visits (p<0.05 (S visit); p<0.001 (F and R visit)). In the PF cluster, recovery samples had a significantly higher P:F ratio than samples at stable and exacerbation visits (p<0.05). For the HF cluster there was again no significant change in the ratio across visits.

Figure 5.11

Change in the qPCR derived P:F ratio across 4 visits within each cluster: Chart represents mean and ±SD in each visit group within P:F clusters

5.3.2.5 Clinical characteristics of qPCR derived P:F clusters

The association between these microbiome clusters and biological markers of bacterial infection was examined in two ways. Firstly, comparisons of baseline clinical characteristics of the qPCR based clusters in all 30 subjects were performed. These were broadly similar to those observed for the 454 based clusters in 19 subjects and
showed no significant difference between the subjects of three clusters based on smoking status, GOLD stage, treatment at exacerbation and BMI (Table 5.5A). The sputum neutrophil count was higher in members of the HP cluster at stable state (Table 5.5B) and exacerbation frequency in the previous 12 months was highest in the HF cluster. HF cluster had significantly higher exacerbation frequency (P=0.05) compared to the PF cluster (Table 5.5). However, in contrast with 454-sequencing, exacerbations treated with steroids alone were not confined to the PF cluster and members of the HP cluster did not have a lower BMI, suggesting that these observations previously are likely to have arisen by chance with a smaller sample set. Additionally, the qPCR based P:F ratio clusters were compared with our previously reported cytokine based biological clusters to define exacerbation phenotypes (Bafadhel et al., 2011). A significant (p=0.039) relationship was found between the PF ratio cluster and the exacerbation phenotypes characterised by the cytokine clusters (Table 5.5). Exacerbations of the HP and PF clusters appeared to be quite specific in their type, mapping to bacteria and Th1 predominant biological clusters for the HP group and to the eosinophil / Th2 predominant biological cluster for the PF group. In contrast, exacerbations of the HF cluster exhibited considerable heterogeneity in their biological characterisation, with representation of all four biological phenotypes (Table 5.5A).

Secondly, P:F ratio clusters were also mapped to biological markers of infection collected at two or more of the visits. These included the % sputum neutrophil count, serum C-reactive protein (CRP) and sputum IL-1β that were compared at stable and exacerbation state between clusters and within clusters across visits to determine if changes in these markers were cluster specific and associated with changes in the P:F ratio (Table 5.5B). In the exacerbation visit there was a significant difference in CRP (p=0.0008) and IL-1β (p=0.02) between the three clusters with highest increase in HP cluster. A significantly greater increase in CRP (p=0.0005) and IL-1β (p=0.0021) from stable to exacerbation visit was also observed in the HP cluster compared to other 2 clusters. Neutrophil counts also were highest in the HP cluster for both stable and exacerbation visit. Within the HP cluster, the exacerbation visit showed highest mean for all three biomarker tested compared to rest of the visit group while, recovery visit within PF cluster showed higher values for CRP and neutrophil. HF cluster showed lowest mean values for neutrophil and IL-1β across visit groups compared to other two
Comparison of SEFR samples based on the relative abundance of the 2 dominant phyla cluster. These trends in clinical biomarkers were coherent with changes observed in ratios in these clusters.

Table 5.5

<table>
<thead>
<tr>
<th>Metadata Groups</th>
<th>HP (N=5)</th>
<th>PF (N=6)</th>
<th>HF (N=18)</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOLD Stage</td>
<td>2 (60)</td>
<td>3 (50)</td>
<td>8 (44.44)</td>
<td>0.467</td>
</tr>
<tr>
<td></td>
<td>3 (40)</td>
<td>1 (6.66)</td>
<td>8 (44.44)</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>4 (100)</td>
<td>2 (33.33)</td>
<td>2 (11.11)</td>
<td></td>
</tr>
<tr>
<td>Male / N (%)</td>
<td>5 (100)</td>
<td>5 (83.33)</td>
<td>13 (72.22)</td>
<td>0.383</td>
</tr>
<tr>
<td>Smoking status</td>
<td>Ex-smoker / N (%)</td>
<td>5 (100)</td>
<td>4 (66.67)</td>
<td>12 (66.67)</td>
</tr>
<tr>
<td>Type of treatment</td>
<td>Antibiotic &amp; Steroid</td>
<td>3 (60)</td>
<td>3 (50)</td>
<td>12 (66.67)</td>
</tr>
<tr>
<td>Antibiotic</td>
<td>2 (40)</td>
<td>0</td>
<td>4 (22.22)</td>
<td></td>
</tr>
<tr>
<td>Steroid</td>
<td>0</td>
<td>3 (50)</td>
<td>2 (11.11)</td>
<td></td>
</tr>
<tr>
<td>BMI kg/m²</td>
<td>27.46</td>
<td>25.52</td>
<td>25.57</td>
<td>0.737</td>
</tr>
<tr>
<td>Exacerbation in past 12 months</td>
<td>3.4</td>
<td>1.7</td>
<td>5.0</td>
<td>0.028</td>
</tr>
<tr>
<td>Cytokine cluster</td>
<td>TNF (bacterial predominant)</td>
<td>3 (60)</td>
<td>1 (16.66)</td>
<td>2 (11.11)</td>
</tr>
<tr>
<td></td>
<td>Th1-IP10 (Viral predominant)</td>
<td>2 (40)</td>
<td>0</td>
<td>6 (33.33)</td>
</tr>
<tr>
<td></td>
<td>Th2, CCL17/M5 (eosinophil dominant)</td>
<td>0</td>
<td>4 (66.67)</td>
<td>5 (27.78)</td>
</tr>
<tr>
<td></td>
<td>Pauci inflammatory</td>
<td>0</td>
<td>0</td>
<td>4 (22.22)</td>
</tr>
</tbody>
</table>

Comparison of clinical parameters between and within the P:F cluster: Table A - Compares the demographic data between the three clusters. "N" denotes the number of subjects in each cluster. Numbers in brackets represent in percentage the number of samples for each parameter. Table B - S, E, F and R visit column gives the mean value of the clinical biomarkers within each cluster. Paired t test column and RM-Anova shows the p value for within subject difference between stable and exacerbation and across all 4 visits respectively. "n" denotes subject number, NP=not performed, ND= no data.

205
The linear association of change in the PF ratio between stable, exacerbation and recovery visit with change in the three clinical biomarkers were also explored. A significant linear correlation was found for between change in P:F ratio and change in CRP and IL-1 beta but not sputum neutrophils (Table 5.6A). For CRP differences was observed in this correlation across clusters, with the highest correlation observed for the HP cluster (Table 5.6B). A similar analysis could not be performed for IL-1 beta due to a lack of data-points.

**Table 5.6**

<table>
<thead>
<tr>
<th>Dependant Variable Independent variable</th>
<th>Pearson Correlation co-efficient</th>
<th>p value</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta$ P:F ratio (E-S &amp; R-E)</td>
<td>$\Delta$ CRP (E-S &amp; R-E)</td>
<td>.445</td>
<td>.001</td>
</tr>
<tr>
<td>$\Delta$ neutrophil (E-S &amp; R-E)</td>
<td></td>
<td>.029</td>
<td>.842</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dependant Variable Independent variable</th>
<th>Pearson Correlation co-efficient</th>
<th>p value</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta$ P:F ratio (E-S)</td>
<td>$\Delta$ CRP (E-S)</td>
<td>.476</td>
<td>.011</td>
</tr>
<tr>
<td>$\Delta$ neutrophil (E-S)</td>
<td>.53</td>
<td>.793</td>
<td>27</td>
</tr>
<tr>
<td>$\Delta$ IL-1beta (E-S)</td>
<td>.613</td>
<td>.012</td>
<td>16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dependant Variable Independent variable</th>
<th>Pearson Correlation co-efficient</th>
<th>p value</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta$ P:F ratio (R-E)</td>
<td>$\Delta$ CRP (R-E)</td>
<td>.465</td>
<td>.014</td>
</tr>
<tr>
<td>$\Delta$ neutrophil (R-E)</td>
<td>.130</td>
<td>.546</td>
<td>24</td>
</tr>
</tbody>
</table>

**Correlation between change in P:F ratio and clinical biomarkers:** A-Rows represent linear (Pearson) correlations for the between visit change ($\Delta$) in P:F ratio and corresponding change in stated clinical biomarkers at exacerbation from stable (E-S) and at recovery from exacerbation (R-E) considered together and individually. (N) - total number of samples analysed. B- Rows represent correlations as described in (A), stratified by P:F cluster.
5.3.3 Grouping of COPD subjects based on PCoA pattern across visit in subjects

In section (4.3.3.2) PCoA plots at OTU level demonstrated considerable heterogeneity in patterns of change within individual subjects across visits during an exacerbation event. Moreover the patterns showed some differences between subject clusters defined by the P:F ratio (Figure 5.5), although this was not consistent.

Through subjective interrogation, three broad patterns of change across sequential visits were characterised and grouping performed (Figure 5.12):

Group A: Subjects showing little change in their microbiome across visits.

Group B: Subjects with a progressive change in their microbiome from stable to exacerbation to recovery visit.

Group C: Subjects with a big change in their microbiome from stable to exacerbation visit, followed by a return to the stable microbiome at the recovery visit.

Data for the follow-up visit was excluded from this analysis as these samples were collected only 2 weeks after the exacerbation visit. It is probable that samples at this time-point would comprise a microbiome in transition following treatment and likely add noise and further complexity to the analysis.
Comparison of SEFR samples based on the relative abundance of the 2 dominant phyla

**Figure 5.12**

*Grouping of subjects based on change in microbiome at OTU level:* each of the scatterplot represents the PC1 co-ordinates along the X axis and PC2 cordinates along the Y axis of samples based PCoA analysis. Sample colours are subject specific and a sequential line starting from Stable visit sample and ending at recovery visit shows the change in the microbiome within subjects across the 3 visit group.

The patterns of longitudinal change in the microbiome described with OTU based grouping broadly resembled the patterns observed in the P:F clusters. In OTU Group A, 5 of 9 subjects were members of the HF cluster; for Group B 4 of 5 subjects belonged to the PF cluster.

To explore the hypothesis of whether substantial changes in microbial composition during exacerbation events are likely to be representative of bacterial infection, clinical biomarkers of infection were compared between the microbiome groups characterised above. Change in percent sputum neutrophils, serum CRP and sputum IL-1β were compared at stable and exacerbation state between and within the OTU groups (Table 5.7).
There was no significant difference between or within the OTU groups for any of the biomarkers. Overall, exacerbation samples in all OTU groups had higher levels of neutrophil, CRP and IL-1β compared to other visit types. Interestingly, changes from stable to exacerbation visit in IL-1β and CRP were most marked in group B but the sample size was too small to make any statistical determination.

Table 5.7

<table>
<thead>
<tr>
<th>Metadata Groups</th>
<th>GroupA</th>
<th>GroupB</th>
<th>GroupC</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>P/F cluster</td>
<td>n=9</td>
<td>n=5</td>
<td>n=5</td>
<td>0.103</td>
</tr>
<tr>
<td>PF</td>
<td>3 (11.1)</td>
<td>4 (80)</td>
<td>3 (60)</td>
<td></td>
</tr>
<tr>
<td>PM</td>
<td>5 (55.5)</td>
<td>0</td>
<td>1 (20)</td>
<td></td>
</tr>
<tr>
<td>GOLD Stage</td>
<td>n=2</td>
<td>n=4</td>
<td>n=2</td>
<td>0.292</td>
</tr>
<tr>
<td>GOLD Stage</td>
<td>6 (66.6)</td>
<td>1 (20)</td>
<td>2 (40)</td>
<td></td>
</tr>
<tr>
<td>Smoking status</td>
<td>n=3</td>
<td>n=2</td>
<td>n=2</td>
<td>0.783</td>
</tr>
<tr>
<td>Ex-smoker / N (%)</td>
<td>6 (60.6)</td>
<td>4 (80)</td>
<td>3 (60)</td>
<td></td>
</tr>
<tr>
<td>Type of treatment</td>
<td>n=3</td>
<td>n=2</td>
<td>n=2</td>
<td>0.039</td>
</tr>
<tr>
<td>Antibiotic &amp; Steroid</td>
<td>6 (66.6)</td>
<td>1 (20)</td>
<td>4 (80)</td>
<td></td>
</tr>
<tr>
<td>Antibiotic</td>
<td>3 (33.3)</td>
<td>1 (20)</td>
<td>1 (20)</td>
<td></td>
</tr>
<tr>
<td>Steroid</td>
<td>0</td>
<td>3 (60)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Cytokine cluster</td>
<td>n=2</td>
<td>n=3</td>
<td>n=1</td>
<td>0.503</td>
</tr>
<tr>
<td>TNF (bacterial predominant)</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>TNF_1 (T10) (T10 predominant)</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>TNF_2 (T10/C17/K1) (eosinophil predominant)</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Pseudomonas inflammatory</td>
<td>n=2</td>
<td>n=3</td>
<td>n=1</td>
<td>0.669</td>
</tr>
<tr>
<td>Exacerbation in previous 12 months</td>
<td>3.9</td>
<td>3</td>
<td>4.8</td>
<td></td>
</tr>
<tr>
<td>Post BD FEV1 (%) predicted</td>
<td>66.32</td>
<td>40.50</td>
<td>50.4</td>
<td>0.109</td>
</tr>
<tr>
<td>BMI kg/m²</td>
<td>25.35</td>
<td>24.85</td>
<td>24.47</td>
<td>0.942</td>
</tr>
</tbody>
</table>

Comparison of clinical parameters between and within the OTU groups: Table A- Compares the demographic data between the three groups. "N" denotes the number of subjects in each group. Numbers in brackets represent in percentage the number of samples available for each parameter. Two of the group A samples were outliers in P:F cluster and therefore not included in PF cluster comparison. Table B - S, E, F and R visit column gives the mean value of the clinical biomarkers within each group. Paired t test column and RM-ANOVA shows the p value for within subject difference between stable and
exacerbation and across all 4 visits respectively. "n" denotes subject number, NP=not performed, ND= no data.

5.4 Discussion

5.4.1 Stratification of samples based on two dominant phyla

The aim of this study was to investigate the bacterial changes occurring in subjects with COPD over the course of an exacerbation event. It was proposed that during exacerbation there would be a significant disturbance in the microbiome structure at the time of exacerbation and following treatment, with restoration of the microbiome in most post exacerbation samples. However, contrary to this idea, there was no detectable difference observed in the exacerbation microbiome structure at population level. One reason for this is the heterogeneity of factors that can precipitate COPD exacerbations (Bafadhel et al., 2011; Gao et al., 2013). Both infectious and non-infectious triggers are recognised, suggesting that only a subgroup of the exacerbations that were studied are likely to have been caused by bacterial infection. Moreover changes in the microbiome structure during a bacterial infection may be too subtle to be detected with a population based approach. Although sequencing techniques provide detailed information about the microbiome, the volume of data that is generated may be associated with too much ‘noise’ to identify statistically significant differences.

In this study, the microbiome pattern in samples collected at all time-points was dominated by members of the phyla Proteobacteria and Firmicutes. This led to the idea that the balance between Proteobacteria and Firmicutes might provide a basis for stratifying the microbiome in COPD sputum. This view is supported by several studies of the gut and oral microbiome, which have indicated that the balance between different components in the microbiome may have biological and clinical significance (Ley et al., 2006; Saini et al., 1999). Also an aggravation episode in a chronic disease state may be polyfactorial and hence subgrouping of samples based on certain biological criteria might help identify bacterial associated subgroups as demonstrated in Cystic Fibrosis study (Carmody et al., 2013).

To address these issues, unbiased cluster analysis was performed to test if exacerbation phenotypes, on the basis of the pattern of P:F ratio expression across the four visits of each exacerbation episode, could be characterized. This resulted in formation of three
distinct clusters. Analysis of individual visit samples between the three clusters showed significant variation in exacerbation and recovery P:F ratio while, stable and follow-up visit exhibited similar P:F ratio across the three P:F clusters. This suggests that exacerbation and recovery P:F ratio were probably the key determinants of cluster structure, implying that the clusters may be useful for characterising different phenotypes of COPD exacerbation.

The P:F ratio pattern of individual samples were concordant with OTU (0.03) patterns, indicating that *Haemophilus*, *Moraxella* from Proteobacteria and *Streptococcus* from Firmicutes are the main drivers of the P:F ratio pattern. Within individual clusters, analysis of change in P:F ratio across the 4 visit points for the first time showed evidence of a subgroup (HP cluster) in which a microbiomic disturbance at exacerbation was associated with an increase in Proteobacteria abundance that was restored towards the stable profile following treatment. Comparison of baseline clinical metadata showed higher neutrophil count and decrease in lung function in HP cluster, although not significant, compared to other two clusters. Increasing neutrophil counts are considered to be indicators of bacterial infection (Banerjee et al., 2004; Soler et al., 1999). Lung function decline has also been associated with bacterial detection in COPD samples (Wilkinson et al., 2003). But due to small sample number it is difficult to evaluate the importance of this cluster. Previously *H. influenzae* and *M. catarrhalis* have been associated with increase in neutrophil airway inflammation markers (Hill et al., 2000; Marin et al., 2010) and sputum purulence (Patel et al., 2002) suggesting a strong role of these bacterial groups in pathogenesis of COPD.

The PF cluster was characterised by a significant increase in Proteobacteria proportion at recovery visit, with *Haemophilus* dominance in 6 out of 8 recovery samples. Interestingly, only the PF group included patients treated with steroids alone. PF cluster recovery samples also had a significantly uneven microbiome distribution pattern compared to the remaining three visits. It is therefore possible that the changes observed at the recovery visit could indicate inappropriate treatment of developing bacterial infection or a secondary bacterial infection.

The HF cluster showed a relatively stable P:F across the 4 visits and it is possible that bacterial infection was less likely to have been a significant contributor to the exacerbation event in subjects of this cluster. However, *S. pneumoniae* is associated
Comparison of SEFR samples based on the relative abundance of the 2 dominant phyla

with COPD exacerbations and there is close homology between this and other frequently occurring mitis group *Streptococci* which could not be reliably differentiated with a sequencing profile resolution to genus level (Větrovský and Baldrian, 2013; Whatmore et al., 2000). Hence, it is difficult to evaluate if there was pneumococcus switch occurring at exacerbation in subjects of the HF group.

### 5.4.2 Validating P:F microbiome cluster by qPCR

As the P:F ratio based stratification showed promising trends, a qPCR based assay to derive P:F ratio was set up to examine if similar clusters and trends can be observed including all 30 subjects. It was not possible to have 16S rDNA based primers specific enough to amplify the whole Proteobacteria group hence primers targeting the Gamma Proteobacteria class were selected, as COPD pathogens mainly belong to this class. Significant correlation and good cluster agreement between sequencing and qPCR generated clusters supported the use of qPCR for extending the P:F ratio based analyses to the complete dataset of 120 samples from 30 subjects. A three cluster model was generated that was comparable with the outcome using 454 sequencing based P:F ratios. Consistent with the earlier observations, the qPCR based HF cluster showed no change in P:F ratio across the 4 visits; the HP cluster was characterised by a significant rise in Proteobacteria abundance at exacerbation; and the PF cluster exhibited an increasing P:F ratio in post-exacerbation recovery samples. Comparing membership between the qPCR and sequence based cluster model, assignment to the PF cluster was most discrepant, with 3 PF subjects from sequencing moving to HP and HF clusters with qPCR. From a modelling perspective, this suggests that the PF cluster is an intermediate group that overlap with the other two groups, implying some shared characteristics with both of them.

Identifying clusters exhibiting the same patterns of change of P:F across visits during exacerbation in a larger sample set with qPCR strengthens the validity of the observations that significant increases in the proportion of Proteobacteria in the HP and PF clusters at different time-points during exacerbation may be associated with a role for bacterial infection in these subgroups. This was examined further by mapping the microbiome clusters to markers of airway and systemic inflammation and clinical outcomes. The three qPCR P:F clusters did not map to disease severity, BMI, gender, smoking status or treatment given post-exacerbation. However important differences between the clusters were identified in the pattern of exacerbation based on phenotypes.
Comparison of SEFR samples based on the relative abundance of the 2 dominant phyla

of cytokine expression. The HP cluster comprised exacerbations belonging to bacterial (60%) and viral predominant (40%) exacerbation phenotypes, suggesting an infectious aetiology. Although a positive association between the detection of rhinovirus and an increase in Proteobacteria load, especially *H. influenzae* has previously been reported (Molyneaux et al., 2013; Wilkinson et al., 2006), no co-occurrence of virus was detected in the HP cluster.

Serum CRP, IL-1β and sputum neutrophils were markers of inflammation serially measured at different time-points during exacerbation. All of these have been shown to be good predictive markers of bacterial exacerbation (Bafadhel et al., 2011; Daniels et al., 2010; Garcha et al., 2012; Marin et al., 2010). Significantly higher increases in CRP and IL-1β from stable to exacerbation visit were observed in the HP cluster compared with the other two clusters.

Interestingly, subjects of the HF cluster reported the highest number of exacerbations in the past 12 months but did not show changes in either microbiome composition or in the clinical biomarkers at exacerbation during the study. This suggests that the trigger for exacerbations in this group is heterogeneous and might not be associated with infection. In keeping with this, cytokine profiles were heterogeneous and clinical biomarkers of infection were unchanged during exacerbation. The PF cluster was characterised by eosinophil pre-dominant cytokine cluster. 50% of PF exacerbations were treated by steroids alone and had very high Proteobacteria abundance at recovery. One possible explanation for the high Proteobacteria in recovery samples might be linked to treatment failure and airway inflammation leading to bacterial infection in post-exacerbation samples.

Overall, the change in P:F ratio at exacerbation compared with stable and recovery vists showed a good linear association with the corresponding change in CRP and IL-1β. However there was wide variation in the correlation for the three clusters with the highest correlation observed in the HP cluster. This further supports the idea that exacerbations in this subgroup were probably due to infection. Interestingly, although neutrophil counts are widely used as a biomarker for bacterial related COPD (Banerjee et al., 2004; Soler et al., 1999) there wasn't a significant correlation between the change in exacerbation P:F ratio and change in neutrophil count. This is likely to be due to the
Comparison of SEFR samples based on the relative abundance of the 2 dominant phyla

poor specificity of this marker as neutrophils have been reported to be raised during both viral and/or bacterial infection (Papi et al., 2006).

### 5.4.3 Grouping of COPD subjects based on PCoA pattern across visit in subjects

PCoA patterns based on the first two component indices derived from OTU (0.03) that accounts for most of the microbial variability between samples, were explored for within subject changes across the S, E and R visits. This was done with the objective of defining subjective groups on the basis of the magnitude and type of change across the three visits. Again three subgroups were identified: Group A with not much shift in microbial structure; Group B showing a progressive change from stable to recovery visit; and Group C with a microbiome shift at exacerbation and restoration at recovery. Although the subgroups identified by this method showed similar trends to clusters obtained using the P:F ratio, there was no significant agreement between the two approaches in subject membership. Furthermore, comparison of clinical biomarkers between these three subgroups did not identify any significant differences or correlations.

### 5.5 Conclusions

The P:F ratio clustering utilised an unsupervised method comparing the relative abundance between the representative phyla of COPD pathogens to define exacerbation phenotypes. With the P:F ratio analyses tentatively three subgroups have been identified in which there are different microbiomic responses during exacerbation. The HP cluster shows disturbance and restoration (a classic bacterial infection pattern), the PF cluster an intriguing pattern at recovery and the HF cluster no change. These clusters also associated well with the immune markers of bacterial infection. At present there are no mechanistic hypotheses to explain how these patterns may be related to the pathogenesis of exacerbations. However, PF ratio analyses are amenable to contemporaneous analyses in diagnostic laboratories and could be used to assign patients to different treatment subgroups in routine practice. Replication of these findings in a larger patient population is clearly desirable before exploring PF-based stratification in a therapeutic trial.
6 Concluding remarks and Future work

COPD is one of the leading causes of death across the world and acute exacerbation episodes are the major contributor to this phenomenon. Characterising the role of bacteria in COPD is important as processes caused by bacteria may be modifiable with antibiotic therapy. For example antibiotic treatment for acute exacerbations is currently based on the number of clinical symptoms diagnosed, with the majority of the population receiving antibiotics. In this study >80% of the exacerbations received antibiotics. However, a better understanding of the role of bacteria in exacerbation may help to target antibiotic treatment more effectively. Our current understanding of the role and type of bacteria associated with different stages of COPD is primarily founded on culture, serological and antibiotic trial based studies. Recently, molecular techniques have been utilised to characterise the COPD microbial community and differences in the microbiome have been compared in different disease severity groups to inform changes in the microbiome associated with disease progression. But these studies have been limited by small sample size, lack of longitudinal samples and differences in disease severity in cohorts from different studies. This project was part of a prospective observational study involving 145 subjects with data on various immune marker expression levels, PCR based virology data, molecular and culture based bacterial data, patient history and demographics. To date this study has the largest number of longitudinal samples analysed for bacterial profiling.

The main objective of this study was to characterize and determine the bacterial changes occurring during exacerbation from stable state using molecular techniques. Preliminary work involved developing qPCR assays to investigate if there was an increase in the total bacterial load at exacerbation along with changes in specific airway pathogen numbers. The qPCR assays performed showed good agreement with routine culture results and although qPCR was significantly more sensitive, there was concordance in the relative proportions of individual pathogens detected across the sample set. Irrespective of the visit type H. influenzae was the most frequently and abundant bacterial group detected followed by M. catarrhalis and S. pneumoniae. These results
suggested that *H. influenzae*, *S. pneumoniae* and *M. catarrhalis* are key components of the COPD microbiome. A total bacterial load of >$10^6$ genome/ml was observed in all the COPD samples with no difference in load between stable and exacerbation. No differences in detection or load of any of the specific bacteria tested were observed between these two stages. Together these observations would suggest that bacteria do not play a significant role in the pathogenesis of acute exacerbations. However, COPD is a complex disease with heterogeneity across multiple domains including severity stage of disease, exacerbation type and aetiology. For exacerbations, this is reflected in the differences in treatments that are prescribed during an event. Within the study population it is likely that the presence of high bacterial numbers with high detection rate of pathogens in stable state of some subjects as well as heterogeneity in exacerbation aetiology contributed to obscuring the changes occurring at exacerbation.

Sequencing and characterization of 16S rDNA in COPD samples further supported the observations of qPCR and culture techniques with dominance of the bacterial genera *Streptococcus, Haemophilus and Moraxella* in almost all the samples. Good correlation between the qPCR and culture results of *H. influenzae* and *M. catarrhalis* with microbiome abundance of *Haemophilus* and *Moraxella* reinforces the likely association of these two Proteobacteria with the airway microbiome of COPD, a finding that has been previously reported using culture. In addition to *Streptococcus*, sequences belonging to *Actinomyces, Prevotella, Veillonella* and *Rothia* genera were also identified in >80% of the samples. Streptococcal species along with *Prevotella* and *Veillonella* are known to be part of the normal oral and URT microbiome(Aas et al., 2005). In this study Streptococcal dominant samples showed positive association with the frequently observed low abundance bacterial constituents listed above whereas the association was negative in *Haemophilus* and *Moraxella* dominated samples.

The reduced diversity of *Haemophilus* dominant samples and their negative association with bacterial groups belonging to the healthy oral and respiratory flora suggests a relationship between Proteobacteria dominance and COPD pathogenesis. *Haemophilus* species and other Gram negative bacteria are known to produce membrane bound lipopolysaccharide (LPS) antigenic molecules as virulence factors (Moran et al., 1996). These molecules are capable of acting as immunostimulators and immunomodulators thus, playing an important role in the pathogenesis of certain bacterial infections. Furthermore, LPS molecules in many bacterial strains aid in the evasion of host immune
defenses through antigenic diversity and host molecule mimicking (Moran et al., 1996).

It is possible that LPS plays a significant role in pathogenesis of COPD. The positive correlation observed in this study between increase in Proteobacteria ratio and change in bacterial related immune markers at exacerbation supports this hypothesis.

Overall, sputum samples at exacerbation visits were not associated with either a change in diversity or bacterial composition compared to non-exacerbation state. These findings are consistent with the absence of difference in total or specific bacterial load that was observed using qPCR. As mentioned before, this is likely to be due to the heterogeneous aetiology of COPD exacerbations, with a significant proportion not caused by bacterial infection. Exacerbation samples among the four visit types also showed a wide range in diversity further supporting the heterogeneity theory in COPD. Understanding the role of bacteria is made more complex by their frequent persistence through mechanisms to evade host immune clearance and antibiotics. The suggestion that COPD pathogens become chronic colonisers was supported by the consistency of microbial community composition observed in multiple stable samples of exacerbating subjects. Microbiome analysis of data stratified based on COPD severity (GOLD stage), smoking status or treatment did not show any conclusive differences in microbial diversity within or between groups, suggesting that although these variables may be important for describing clinical phenotypes of COPD, they are less likely to have a cause or effect association with the sputum microbiome.

In contrast, characterising the heterogeneity of COPD exacerbation episodes on the basis of the sputum microbiome identified a potential subset of exacerbations in which bacteria were likely to play a role. Based on the negative association observed between Firmicutes and Proteobacteria as dominant bacterial groups, the relative abundance of these two phyla provided a useful summary variable of microbial constituents for clustering exacerbation episodes. The formation of three clusters, with distinct changes in microbiome, by unsupervised classification, highlights the need for phenotyping the COPD population and pitfalls of deriving conclusions based on agglomerating the entire COPD population.

In addition to the multifactorial nature of COPD, study of the disease is complicated by methodological factors such as sample types used for representing the airway microbiome and difficulty with collecting longitudinal samples from large and
Concluding remarks and Future work

representative populations. Furthermore, despite being highly sensitive with a wider dynamic range than routine culture, the molecular techniques have their own limitations. With high-throughput sequencing technology becoming more widely available, the downstream sequence curation pipeline and the biostatistical analysis utilised to explore the bacterial role in COPD is a new and fast growing area. This presents new challenges with analysis and interpretation of large and high-dimension datasets, particularly as sequencing and biostatistical algorithms are continuously updated. A recommended or universally accepted approach does not exist and therefore a cautious approach is required before drawing any conclusions from such analysis. Some of these limitations that could influence the outcomes of the study are discussed below in more detail.

6.1 Limitations of the study

6.1.1 Sample and clinical study design

6.1.1.1 Sample type
Use of sputum as representative of lower respiratory tract microbiome has been controversial and criticised for a number of reasons. Sputum samples are considered as not being true representative of the lung microbiota. Although various precautions such as saline mouthwash prior to sputum collection and separation of sputum plugs from saliva before sample processing are performed to reduce the probability of contamination, oral surfaces and URT harbour a plethora of organisms. As a consequence, it is impossible to achieve complete elimination of bacterial communities from these surfaces. Recently Cabrera-Rubio and co-workers utilised 454 sequencing to characterize the respiratory tract microbiome in five patients. They reported sputum samples underrepresented the microbial diversity of lung compared to tissue or BAL samples. There was considerable variation in the microbiome structure between all of the different sample types within individuals although similar bacterial groups were identified across these sample types (Cabrera-Rubio et al., 2012). In contrast to this, other studies in healthy and COPD subjects suggest the lung microbiome to be a continuation of the URT and oral microbiome (Charlson et al., 2011; Pragman et al., 2012).
Despite these limitations of sputum samples, in routine clinical practice, it is the easiest, cheapest and most practical sample type to obtain in longitudinal studies of large populations. Performing invasive procedures, especially during exacerbation in moderate and severe COPD subjects can be difficult (Bhowmik et al., 2000). There can also be difficulties with additional ethics requirement and volunteer recruitment for trials with invasive procedures. Finally, with reports of variation in microbial structure from tissue sampling from discrete multiple sites of lungs within an individual (Erb-Downward et al., 2011) it is difficult to state what would be the ideal sample type for COPD microbiome analysis.

One limitation of this study was that healthy subjects were not included. The LRT and lungs were previously believed to be bacteriologically sterile, however sequencing studies refute this, showing the existence of bacterial communities in health (Charlson et al., 2011). Furthermore, differentiating the COPD microbiome from healthy microbiome has been complicated due to the considerable overlap between the two groups (Erb-Downward et al., 2011) and considerable variability of the healthy lung microbiome between individuals (Erb-Downward et al., 2011).

6.1.1.2 Sample collection and storage

Processing of fresh sputum samples would have been the preferred choice to reduce the biases that might be introduced due to loss of cell integrity of fragile bacterial cells due to freeze thawing. But due to the multi-centred nature of this study, multiple-processing from the same sample and lag between the clinical sample collection date and molecular study start date meant that samples had to be frozen. A study performed in our laboratory using 10 of the COPD sputum samples from our project, extracted using the same DNA protocol before and after freezing showed no significant differences in TBL between frozen and non-frozen samples, although difference in the relative abundance of microbiome constituents between frozen and non-frozen samples has been reported in gut and other studies (Bahl et al., 2012; Cardona et al., 2012; Sekar et al., 2009). 

Lauber and co-workers examined the effect of storage time and temperature on the microbiome from soil and human fecal and skin samples. They showed no significant effect of storage time or temperature on microbiome structure or diversity and concluded that environmental factors and biases in molecular techniques are likely to be a greater source of variation between samples (Lauber et al., 2010). Whether significant difference in microbiome composition exists between frozen and non-frozen samples is
not clear and requires further investigation. Another limitation of frozen samples is the inability to use pre-treatment methods to eliminate interfering qPCR signals from host DNA. It would have been preferable to use pre-extraction treatment (MolYsis) for enriching the bacterial cells and eliminating host DNA. But use of such treatment showed a significant loss of bacterial load and in some cases this involved complete loss of signal from *H. influenzae* and *S. pneumoniae*.

For this study mostly spontaneous sputum samples were collected but in some cases sputum expectoration was induced. A study performed by Bhowmik et al. comparing induced and spontaneous sputum reported a higher proportion of viable cells recovered from induced sputum then spontaneous (Bhowmik et al., 1998). Whether variation in the sputum collection method (spontaneous vs induced) and the time between sputum collection and processing effects the molecular outcomes are not known. Our study was not designed to address these questions and further studies are needed.

### 6.1.2 Factors contributing to bias and variability in molecular studies:

#### 6.1.2.1 DNA extraction

The two main factors influencing the choice of DNA extraction method is its cell lysis efficiency and the quality and yield of DNA extract from complex microbial communities. Recent studies on gut, oral and respiratory microbiome have shown differences in microbiome composition and structure between different extraction and pre-treatment methods used to lyse all bacterial cells in mixed microbial studies (Abusleme et al., 2014; Kennedy et al., 2014; Lazarevic et al., 2013; Willner et al., 2012). Various techniques such as freeze thawing cycles, enzymatic lysis involving peptidoglycan hydrolysing enzymes such as lysozyme and mutanolysin, mechanical bead beating and high temperatures are utilised for achieving complete lysis of bacterial cells. It has been suggested that bead beating techniques have higher cell lysis efficiency then enzymatic lysis methods especially with tougher cell walled bacteria such as members of Actinobacteria within the microbiome (Kauffmann et al., 2004; Lazarevic et al., 2013; Salonen et al., 2010). In the COPD study performed by Pragman et al., Actinobacteria formed a dominant constituent of the microbiome in most samples (Pragman et al., 2012). This might be due to the bead beating technique used in this study. However, Erb-Downward and co-workers who had also utilised the same DNA extraction technique didn't show abundance of this group in COPD samples (Erb-
Concluding remarks and Future work

Downward et al., 2011). In our study, members of Actinobacteria group were represented as part of the core microbiome but present in low proportions. One risk with use of harsher mechanical conditions for cell disruption is the possible loss of DNA from fragile cells like S. pneumoniae and H. influenzae or downstream chimera formation during amplification due to excessive DNA fragmentation (Yuan et al., 2012). Furthermore, the bacterial cell integrity in sputum samples of this study could have been compromised due to use of frozen samples for analysis. The Qiagen (QIAamp DNAmini kit) commercial DNA extraction kit based on enzymatic lysis used in this study was previously tested in a gut microbiota study showing 95% DNA extraction efficiency, good repeatability with > 99% microbial similarity index between replicates and produced a near identical microbiome structure when compared with the bead beating lysis method (Li et al., 2003).

6.1.2.2 qPCR reaction variability:

In absolute quantification, amplification efficiency is based on the standard curve and doesn't take into consideration variations in PCR efficiency between individual samples. The presence of inhibitory substances in samples can reduce their amplification efficiency and may lead to complete inhibition of amplification producing false negative results (Wilson, 1997). A synthetic nucleotide sequence or non-specific DNA target at a known standard concentration may be spiked into samples to detect amplification inhibition and this is referred to as an internal control. One of the limitations of the qPCR assays in this study was that internal control were not used, although comparisons of culture positives with qPCR positives provided a validation check for qPCR with >98% of culture positive samples showing detection with qPCR. A set of samples that were culture positive but qPCR negative, high qPCR baseline signal producing samples and qPCR positives were checked for inhibition by performing a MTB spiked assay (3.3.2.1). Except for some high baseline producing samples, none of the samples tested showed amplification inhibition. The high baseline samples were diluted to an optimum concentration that reduced the inhibitory substance effect in PCR but at the same time allowed amplification of the target product. Although these results and optimisations suggest that it is unlikely that lack of use of internal control could have contributed to any significant bias in this study, we intend to incorporate an internal control in further qPCR assays.
Concluding remarks and Future work

Other potential sources of bias in qPCR are the factors involved in the construction of the standard curve including the initial quantification of the standard curve template, serial dilution of the template and the algorithmic determination of the Ct value (Love et al., 2006; Smith and Osborn, 2009). To reduce the variability between assays criteria for passed qPCR assay based on $R^2$, amplification efficiency and Ct threshold were set as recommended by previous literature (Smith et al., 2006). Despite these criteria, minor variations in standard curve and Ct values between assays can be a source of variation between assays due to the logarithmic properties of the standard curve (Smith and Osborn, 2009). It is therefore advisable if possible to perform quantification assay on longitudinal samples within an assay to avoid inter assay variations.

One of the limitations of the DNA qPCR is that it will not discriminate between live and dead bacteria. Propidium monazide dye (PMA) binds preferentially to dsDNA with high affinity in a photolysis reaction thus, permanently modifying the DNA and making it unavailable for amplification. Performing this pre-treatment on samples would lead to crosslinking of the dye with free and compromised cell DNA to generate readings from only live cells (Nocker et al., 2009). But in this study, as frozen samples were used, performing PMA assay could have produced biased results.

6.1.2.3 16S rDNA based microbial community PCR reaction

Inherent biases with PCR reactions become more pronounced while working with 16S rDNA primers for capturing the complex microbial community (von Wintzingerode et al., 1997). Some of the common problems with amplification are production of primer dimers, spurious non-specific products, chimeras and sequence read errors. In PCR assays targeting a single bacterial population it is possible to eliminate the production of non-specific products and primer dimers by increasing the annealing temperature and target specific primers. In the case of 16S rDNA based community PCR, use of reaction conditions that are too stringent might lead to loss of amplification products from some bacterial population while the universal nature of primers are prone to annealing non-specifically. For the amplification product to be a close representation of bacterial diversity in the sample, lower primer annealing temperatures are utilised in PCR while lowering the number of PCR cycles, post PCR gel purification and AMPure magnetic bead based PCR product clean-up reduce the formation and removal of primer-dimers and non-specific products. In a mixed microbial PCR, differential amplification of PCR
products can occur due to GC content variability, differences in binding energies resulting from primer degeneracy (mixed nucleotide primers) and preferential template reannealing (longer homologs compared to primers) with increasing template concentration inhibiting formation of template primer hybrids (Polz and Cavanaugh, 1998). The template reannealment inhibition of PCR can cause a significant bias with the final concentrations not reflecting the initial ratio of the microbial community in a sample and is a function of the PCR cycle number (Suzuki and Giovannoni, 1996). To minimise the potential for such a bias the PCR total cycle number is generally maintained below the normally used 35 PCR cycles in analyses of microbial community structure (Wintzingerode et al., 1997). High fidelity DNA polymerase was used to reduce PCR based sequencing errors.

6.1.2.4 16S rDNA primers and 16rDNA hypervariable region for bacterial community study:

Based on previous comparative bacterial studies utilising the 16S rDNA gene, a catalogue of so called universal primers that are complementary to conserved stretches of the eubacterial 16S rDNA gene have been generated (Baker et al., 2003; Lane et al., 1985; Nossa et al., 2010; Wang and Qian, 2009). However the ‘universality’ of these primers is questionable as absolutely conserved regions of the rDNA gene normally extend only to consecutive tetranucleotides and the primers are developed based on the evaluation of complementarity across the bacterial domain of identified species (Baker et al., 2003; Forney et al., 2004). Despite these limitations of 16S rDNA primer universality, it has been suggested that approximately 70% complementarity between the primers and the template is sufficient for successful annealment and amplicon production provided the PCR stringency conditions are sub-optimal (Baker et al., 2003). Traditionally, the whole length amplifying 16S gene primers have been used for Sanger sequencing but with the advent of high throughput and qPCR sequence technologies 16S primers creating shorter amplicon products are generated. The primers used to target the nine hypervariable 16S rDNA regions have been shown to differ in their taxonomic coverage of bacterial groups while the taxonomic classification accuracy and resolution is biased by the choice of the 16S rDNA hypervariable region for community analysis (Chakravorty et al., 2007; Liu et al., 2008; Wang and Qian, 2009; Youssef et al., 2009). Insilico comparison of taxonomic assignments between three combined hypervariable region amplicons: V1-V3, V3-V5 and V6-V9 generated from the full
length 16S rDNA sequence database are reported to be comparable in their accuracy of microbial composition (Vilo, 2012). Although study of individual hypervariable regions has suggested the V2 and V4 hypervariable regions to offer the best classification accuracy (Vilo, 2012; Wang et al., 2007), the widely used primer pair (27F-338R) targeting the V2 region has been reported to lack sensitivity, with poor amplification of members belonging to *Bifidobacterium* genus and Verrucomicrobia phylum. These groups form the core microbiome constituents in gut and soil microbiota respectively (Bergmann et al., 2011; Hayashi et al., 2004). Overall, primers targeting the mid 16S rDNA region from V3 to V6 variable regions are considered to have the best predicted coverage (Liu et al., 2008; Wang and Qian, 2009). In this study V6-V8 16S rDNA was targeted as this was the only region with optimal 454 sequencing length to allow microbiome analysis based on information from three hypervariable regions. The V6 region utilised in this study is suggested to have comparable discriminatory power to whole 16S rDNA gene even at genus level (Chakravorty et al., 2007; Huse et al., 2008).

### 6.1.2.5 Analysis pipelines

Accurately quantifying relative abundance of microbial constituents in 16S rDNA based microbial studies is affected by the existence of variable and multiple copy numbers of this gene amongst bacteria (Crosby and Criddle, 2003; Farrelly et al., 1995). At the time we analysed our dataset it wasn't possible to normalise the bacterial groups detected based on their 16S copy number. However, this is unlikely to have had any significant influence on the observed microbial patterns due to occurrence of similar 16SrDNA copy numbers among the observed dominant COPD bacterial constituents. The recently developed PICRUSt (phylogenetic investigation of communities by reconstruction of unobserved states) pipeline allows normalization of the gene copy number to predict functional composition and abundance based on genome content of microbes identified by the phylogenetic marker (Langille et al., 2013). Due to time constraints, this analysis could not be utilised in the current study but will be incorporated in further analysis pipelines.

### 6.2 Conclusions

This study is one of the first to begin characterising the role of the bacteria in COPD exacerbations using molecular methods. The different approaches used to investigate the COPD microbiome implicate a role for increasing Proteobacteria dominance, mainly...
driven by *H. influenzae* in COPD. Although *S. pneumoniae* is also considered to play an important role in COPD it was difficult to explore its role in detail due to considerable sequence homology with the mitis group. Although the study did not provide information sufficient to predict an exacerbation based the microbiome pattern at stable state or a clear microbiome signature at exacerbation, the P:F ratio based clustering has identified trends that merit further study in larger cohorts.

### 6.3 On-going and Future work

#### 6.3.1 COPD microbiome and P:F ratio based analysis with larger sample cohort:

Currently a larger cohort of samples (~ 500) from 140 exacerbation episodes comprising three or four sequential visits has been sequenced utilising the V4 and V5 16S rDNA region. The objective is to investigate if the trends identified using P:F ratio based clustering in this study are consistent and reach statistical significance. The new sample cohort also includes several exacerbation episodes from an individual, providing an opportunity to study whether the microbiome structure varies between exacerbations within an individual.

The 120 samples from this study have also been sequenced using the new primers to explore if significant differences in the COPD microbiome structure and composition are detected between the two primer sets used.

#### 6.3.2 Deep sequencing for greater sampling depth to understand the role of rarer species in COPD microbiome:

Species present in low abundance could be a core microbiome constituent but often cannot be detected consistently across samples due to inadequate sampling depth. For example, 16s rDNA microbiome analysis of the human distal oesophagus yielded 6800 sequences but revealed only one of the 166 species, *S. mitis*, shared by all 34 subjects sampled (Nossa et al., 2010). The recently developed ultra-deep illumina sequencing technology (Miseq) can achieve a 10 to 100 fold higher sequencing depth compared with 454 at a sequence length of up to 400 bases at a substantially lower cost (Caporaso et al., 2012). We have sequenced 24 samples on the Miseq illumina sequencer, utilising the new V4 and V5 16S rDNA targeting primers. Direct comparison with 454 generated sequences is possible and will be explored.
6.3.3 PMA based microbial community analysis:
Results on bacterial load in COPD samples obtained from qPCR assays are biased due to poor discrimination of DNA based assays between live and dead bacteria, particularly in antibiotic treated samples. As mentioned earlier the PMA based qPCR assay couldn't be undertaken on frozen samples. Recently, PMA treated DNA was utilised for comparing microbial diversity and composition between PMA treated and non-treated Cystic Fibrosis sputum samples revealing differences in microbial membership between PMA and non-PMA treated samples (Rogers et al., 2013). It would be desirable to set up a pilot study for developing the PMA based DNA assays for quantification assays and microbial community analysis.

6.3.4 Whole genome sequencing (WGS) and Metatranscriptomics:
In this study we were able to characterise in some detail the resident bacterial community in COPD sputum. However, bacteria do not exist in isolation but in a complex and dynamic ecosystem that includes viral and fungal communities as well as host tissue. Furthermore the techniques we have used are unable to inform function and activity. WGS, also known as shotgun sequencing, is based on sequencing the DNA fragment library generated from shearing the total genome of a sample and building genome assemblies of incomplete genomes or incomplete chromosomes of prokaryotes or eukaryotes. In COPD it is believed that airway damage from viral infection allows colonisation of pathogenic bacteria and may therefore play an active role in triggering exacerbations (Green, 1968; Smith et al., 1976). This is supported by a recent study that reported a significant increase in bacterial load, especially \textit{H. influenzae}, in COPD subjects 15 days after being infected with Rhinovirus (Molyneaux et al., 2013). The WGS approach can help to build a comprehensive picture of eukaryotic, prokaryotic and viral communities present within a sample. This will begin to reveal important relationships that are likely to exist between these groups. WGS can also determine if specific genes rather than certain microbial groups are associated with exacerbation, based on the functional (biochemical and metabolic) properties of the microbial community (Gill et al., 2006; Petrosino et al., 2009). Metatranscriptomics enables investigating the actively transcribed ribosomal and messenger RNA from a community (Gilbert and Hughes, 2011). Together these novel and rapidly developing technologies are likely to help transform our understanding of COPD.
References


Belozersky, A. N., and A. S. Spirin, 1958, A Correlation between the Compositions of

Ben Salah, I., T. Adekambi, D. Raoult, and M. Drancourt, 2008, rpoB sequence-based


Bergmann, G. T., S. T. Bates, K. G. Eilers, C. L. Lauber, J. G. Caporaso, W. A. Walters, R. Knight,
and N. Fierer, 2011, The under recognized dominance of Verrucomicrobia in soil

Bhowmik, A., T. A. Seemungal, R. J. Sapsford, J. L. Devalia, and J. A. Wedzicha, 1998,
Comparison of spontaneous and induced sputum for investigation of airway

Gill, C. M. Fraser-Liggett, and D. A. Relman, 2010, Bacterial diversity in the oral cavity

Björkstén, B., E. Sepp, K. Julge, T. Voor, and M. Mikelsaar, 2001, Allergy development and the

Gill, C. M. Fraser-Liggett, and D. A. Relman, 2010, Bacterial diversity in the oral cavity

Björkstén, B., E. Sepp, K. Julge, T. Voor, and M. Mikelsaar, 2001, Allergy development and the

393-4.

Bogdan, J. R., and J. T. Curtis, 1957, An Ordination of the Upland Forest Communities of Southern
Wisconsin: Ecological Monographs, v. 27, p. 325-349.

British Lung Foundation, 2007, Invisible lives: Chronic Obstructive Pulmonary Disease (COPD):
Finding the missing millions.


Chao, A., 1984, Nonparametric estimation of the number of classes in a population, p. 265-270.


Elizabeth van, P.V., Alex van Belkum, John P. Hays, 2008, Principles and technical aspects of PCR amplification, Springer.


Free, R. C., 2005, Molecular analysis of the lower respiratory tract microbiota, University of Leicester, United Kingdom.


Greisen, K., M. Loeffelholz, A. Purohit, and D. Leong, 1994, PCR primers and probes for the 16S rRNA gene of most species of pathogenic bacteria, including bacteria found in cerebrospinal fluid: J Clin Microbiol, v. 32, p. 335-51.


Roche, 2008, 250th Peer-reviewed Publication Documents the Success of 454 Sequencing.


Sethi, S., 2000a, Bacterial infection and the pathogenesis of COPD: Chest, v. 117, p. 286S-91S.


Appendices

Additional tables and figures

Figure A1.1: Variability between fresh, frozen and automated DNA extraction methods

QIAcube is the automated DNA extraction method using the Qiagen Gram positive protocol. There was no significant difference observed in TBL between fresh, frozen and automated extraction methods.

Figure A1.2: QPCR interassay variability

Three Proteobacteria qPCR runs were performed on 11 samples to test the effect of interassay variability on the number of Proteobacteria copies obtained from each run. There was no significant difference (p=0.978) between the three runs.
Table A1.1: Chronic colonisation at stable state

<table>
<thead>
<tr>
<th>Pathogen Tested (%)</th>
<th>Always positive (%)</th>
<th>Mixed (%)</th>
<th>Never positive (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>qPCR for any 4 of the pathogens tested (%)</td>
<td>70.37</td>
<td>29.63</td>
<td>0.00</td>
</tr>
<tr>
<td>qPCR for <em>H. influenzae</em> (%)</td>
<td>59.26</td>
<td>40.74</td>
<td>0.00</td>
</tr>
<tr>
<td>qPCR for <em>S. pneumoniae</em> (%)</td>
<td>25.93</td>
<td>48.15</td>
<td>25.93</td>
</tr>
<tr>
<td>qPCR for <em>M. catarrhalis</em> (%)</td>
<td>14.81</td>
<td>77.78</td>
<td>7.41</td>
</tr>
</tbody>
</table>

27 subjects with samples available at following three (baseline, 3months and 6 months) stable state were selected to determine the detection pattern of the four pathogens tested by qPCR. Columns "Always positive", "Mixed" and "Never positive" represents the % of subjects with all three stable visit, 1 to 2 stable visit and none of the stable visit positive for pathogens tested, respectively.

Table A1.2: Patient table and their 454 numbering in SEFR group

<table>
<thead>
<tr>
<th>Subject_number</th>
<th>Hospital Numbers</th>
<th>454 numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3  3.1  3.11  3.12</td>
<td>S01  E01  F01  R01</td>
</tr>
<tr>
<td>3</td>
<td>2  2.1  2.11  2.12</td>
<td>S02  E02  F02  R02</td>
</tr>
<tr>
<td>7</td>
<td>4  4.1  4.11  4.12</td>
<td>S03  E03  F03  R03</td>
</tr>
<tr>
<td>9</td>
<td>5  5.1  5.11  5.12</td>
<td>S04  E04  F04  R04</td>
</tr>
<tr>
<td>21</td>
<td>5  5.1  5.11  5.12</td>
<td>S05  E05  F05  R05</td>
</tr>
<tr>
<td>23</td>
<td>3  3.1  3.11  3.12</td>
<td>S07  E07  F07  R07</td>
</tr>
<tr>
<td>27</td>
<td>3  3.1  3.11  3.12</td>
<td>S08  E08  F08  R08</td>
</tr>
<tr>
<td>76</td>
<td>2  2.2  2.21  2.22</td>
<td>S09  E09  F09  R09</td>
</tr>
<tr>
<td>39</td>
<td>2  5.1  5.11  5.12</td>
<td>S10  E10  F10  R10</td>
</tr>
<tr>
<td>40</td>
<td>5  5.1  5.11  5.12</td>
<td>S11  E11  F11  R11</td>
</tr>
<tr>
<td>44</td>
<td>5  5.1  5.11  5.12</td>
<td>S12  E12  F12  R12</td>
</tr>
<tr>
<td>47</td>
<td>2  2.1  2.11  2.12</td>
<td>S13  E13  F13  R13</td>
</tr>
<tr>
<td>52</td>
<td>2  2.2  2.21  2.22</td>
<td>S14  E14  F14  R14</td>
</tr>
<tr>
<td>56</td>
<td>2  2.1  2.11  2.12</td>
<td>S15  E15  F15  R15</td>
</tr>
<tr>
<td>57</td>
<td>2  2.1  2.11  2.12</td>
<td>S16  E16  F16  R16</td>
</tr>
<tr>
<td>58</td>
<td>2  2.1  2.11  2.12</td>
<td>S17  E17  F17  R17</td>
</tr>
<tr>
<td>60</td>
<td>2  2.1  2.11  2.12</td>
<td>S18  E18  F18  R18</td>
</tr>
<tr>
<td>63</td>
<td>4  2.1  2.11  2.12</td>
<td>S19  E19  F19  R19</td>
</tr>
<tr>
<td>71</td>
<td>2  3.1  3.11  3.12</td>
<td>S20  E20  F20  R20</td>
</tr>
<tr>
<td>74</td>
<td>3  3.1  3.11  3.12</td>
<td>S21  E21  F21  R21</td>
</tr>
<tr>
<td>78</td>
<td>2  2.1  2.11  2.12</td>
<td>S22  E22  F22  R22</td>
</tr>
<tr>
<td>79</td>
<td>2  2.1  2.11  2.12</td>
<td>S23  E23  F23  R23</td>
</tr>
<tr>
<td>83</td>
<td>2  2.1  2.11  2.12</td>
<td>S24  E24  F24  R24</td>
</tr>
<tr>
<td>93</td>
<td>2  2.1  2.11  2.12</td>
<td>S25  E25  F25  R25</td>
</tr>
<tr>
<td>96</td>
<td>4  2.1  2.11  2.12</td>
<td>S26  E26  F26  R26</td>
</tr>
<tr>
<td>102</td>
<td>3  2.1  2.11  2.12</td>
<td>S27  E27  F27  R27</td>
</tr>
<tr>
<td>115</td>
<td>2  3.1  3.11  3.12</td>
<td>S28  E28  F28  R28</td>
</tr>
<tr>
<td>122</td>
<td>3  3.2  3.21  3.22</td>
<td>S29  E29  F29  R29</td>
</tr>
<tr>
<td>123</td>
<td>2  2.1  2.11  2.12</td>
<td>S30  E30  F30  R30</td>
</tr>
</tbody>
</table>
Table A1.3: Patient table and their 454 numbering in SSE group

<table>
<thead>
<tr>
<th>Hospital Number</th>
<th>Subject_number</th>
<th>Sequential stable visit number</th>
<th>454 number</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>3.00</td>
<td>A01</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>4.00</td>
<td>B01</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>5.00</td>
<td>C01</td>
<td></td>
</tr>
<tr>
<td>76</td>
<td>2.00</td>
<td>A02</td>
<td></td>
</tr>
<tr>
<td>76</td>
<td>3.00</td>
<td>B02</td>
<td></td>
</tr>
<tr>
<td>76</td>
<td>4.00</td>
<td>C02</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>2.00</td>
<td>A03</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>4.00</td>
<td>B03</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>5.00</td>
<td>C03</td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>2.00</td>
<td>A04</td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>3.00</td>
<td>B04</td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>4.00</td>
<td>C04</td>
<td></td>
</tr>
<tr>
<td>57</td>
<td>2.00</td>
<td>A05</td>
<td></td>
</tr>
<tr>
<td>57</td>
<td>3.00</td>
<td>B05</td>
<td></td>
</tr>
<tr>
<td>57</td>
<td>4.00</td>
<td>C05</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>2.00</td>
<td>A06</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>3.00</td>
<td>B06</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>4.00</td>
<td>C06</td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>2.00</td>
<td>A07</td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>3.00</td>
<td>B07</td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>4.00</td>
<td>C07</td>
<td></td>
</tr>
<tr>
<td>78</td>
<td>2.00</td>
<td>A08</td>
<td></td>
</tr>
<tr>
<td>78</td>
<td>4.00</td>
<td>B08</td>
<td></td>
</tr>
<tr>
<td>78</td>
<td>5.00</td>
<td>C08</td>
<td></td>
</tr>
<tr>
<td>83</td>
<td>2.00</td>
<td>A09</td>
<td></td>
</tr>
<tr>
<td>83</td>
<td>3.00</td>
<td>B09</td>
<td></td>
</tr>
<tr>
<td>83</td>
<td>4.00</td>
<td>C09</td>
<td></td>
</tr>
<tr>
<td>96</td>
<td>2.00</td>
<td>A10</td>
<td></td>
</tr>
<tr>
<td>96</td>
<td>3.00</td>
<td>B10</td>
<td></td>
</tr>
<tr>
<td>96</td>
<td>4.00</td>
<td>C10</td>
<td></td>
</tr>
<tr>
<td>123</td>
<td>2.00</td>
<td>A11</td>
<td></td>
</tr>
<tr>
<td>123</td>
<td>4.00</td>
<td>B11</td>
<td></td>
</tr>
<tr>
<td>123</td>
<td>6.00</td>
<td>C11</td>
<td></td>
</tr>
</tbody>
</table>
Table A1.4: Sample numbers of stable samples sequenced both in SEFR and SSE set

<table>
<thead>
<tr>
<th>454 NO_SEFR</th>
<th>454 NO_SSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>S7</td>
<td>1A</td>
</tr>
<tr>
<td>S8</td>
<td>2C</td>
</tr>
<tr>
<td>S12</td>
<td>3A</td>
</tr>
<tr>
<td>S14</td>
<td>4A</td>
</tr>
<tr>
<td>S15</td>
<td>5A</td>
</tr>
<tr>
<td>S16</td>
<td>6A</td>
</tr>
<tr>
<td>S20</td>
<td>7A</td>
</tr>
<tr>
<td>S22</td>
<td>8A</td>
</tr>
<tr>
<td>S24</td>
<td>9A</td>
</tr>
<tr>
<td>S26</td>
<td>10C</td>
</tr>
<tr>
<td>S30</td>
<td>11A</td>
</tr>
</tbody>
</table>

Table A1.5: SEFR set samples for bacterial community analysis after sequence filtration steps

<table>
<thead>
<tr>
<th>Stable</th>
<th>Exacerbation</th>
<th>Follow-up</th>
<th>Recovery</th>
</tr>
</thead>
<tbody>
<tr>
<td>S01</td>
<td>&lt;500 reads</td>
<td>F01</td>
<td>R01</td>
</tr>
<tr>
<td>S02</td>
<td>E02</td>
<td>F02</td>
<td>R02</td>
</tr>
<tr>
<td>S03</td>
<td>E03</td>
<td>F03</td>
<td>R03</td>
</tr>
<tr>
<td>&lt;500 reads</td>
<td>E04</td>
<td>&lt;500 reads</td>
<td>R04</td>
</tr>
<tr>
<td>S05</td>
<td>&lt;500 reads</td>
<td>F05</td>
<td>R05</td>
</tr>
<tr>
<td>S06</td>
<td>E06</td>
<td>F06</td>
<td>R06</td>
</tr>
<tr>
<td>S07</td>
<td>E07</td>
<td>F07</td>
<td>R07</td>
</tr>
<tr>
<td>S08</td>
<td>E08</td>
<td>F08</td>
<td>R08</td>
</tr>
<tr>
<td>S09</td>
<td>E09</td>
<td>F09</td>
<td>R10</td>
</tr>
<tr>
<td>S10</td>
<td>E10</td>
<td>F10</td>
<td>&lt;500 reads</td>
</tr>
<tr>
<td>S11</td>
<td>E11</td>
<td>F11</td>
<td>R11</td>
</tr>
<tr>
<td>S12</td>
<td>E12</td>
<td>F12</td>
<td>R12</td>
</tr>
<tr>
<td>S13</td>
<td>E13</td>
<td>F13</td>
<td>R13</td>
</tr>
<tr>
<td>S14</td>
<td>E14</td>
<td>F14</td>
<td>R14</td>
</tr>
<tr>
<td>S15</td>
<td>E15</td>
<td>F15</td>
<td>R15</td>
</tr>
<tr>
<td>S16</td>
<td>E16</td>
<td>F16</td>
<td>R16</td>
</tr>
<tr>
<td>S17</td>
<td>E17</td>
<td>F17</td>
<td>R17</td>
</tr>
<tr>
<td>S18</td>
<td>E18</td>
<td>F18</td>
<td>R18</td>
</tr>
<tr>
<td>S19</td>
<td>E19</td>
<td>F19</td>
<td>R19</td>
</tr>
<tr>
<td>S20</td>
<td>E20</td>
<td>F20</td>
<td>R20</td>
</tr>
<tr>
<td>S22</td>
<td>&lt;500 reads</td>
<td>F22</td>
<td>R22</td>
</tr>
<tr>
<td>S24</td>
<td>E24</td>
<td>F24</td>
<td>R24</td>
</tr>
<tr>
<td>&lt;500 reads</td>
<td>E25</td>
<td>F25</td>
<td>R25</td>
</tr>
<tr>
<td>S26</td>
<td>E26</td>
<td>F26</td>
<td>R26</td>
</tr>
<tr>
<td>&lt;500 reads</td>
<td>E27</td>
<td>&lt;500 reads</td>
<td>R27</td>
</tr>
<tr>
<td>S28</td>
<td>&lt;500 reads</td>
<td>F28</td>
<td>R28</td>
</tr>
<tr>
<td>S29</td>
<td>&lt;500 reads</td>
<td>F29</td>
<td>R29</td>
</tr>
<tr>
<td>S30</td>
<td>E30</td>
<td>F30</td>
<td>R30</td>
</tr>
<tr>
<td>25</td>
<td>23</td>
<td>26</td>
<td>27</td>
</tr>
</tbody>
</table>
Table A1.6: SSE and NES set samples remaining for bacterial community analysis after filtration steps

<table>
<thead>
<tr>
<th>SSE 1</th>
<th>SSE 2</th>
<th>SSE 3</th>
<th>NES</th>
</tr>
</thead>
<tbody>
<tr>
<td>A01</td>
<td>B01</td>
<td>C01</td>
<td>NES01</td>
</tr>
<tr>
<td>A02</td>
<td>&lt; 500 reads</td>
<td>&lt; 500 reads</td>
<td>NES02</td>
</tr>
<tr>
<td>A03</td>
<td>B03</td>
<td>C03</td>
<td>NES03</td>
</tr>
<tr>
<td>A04</td>
<td>B04</td>
<td>C04</td>
<td>NES04</td>
</tr>
<tr>
<td>A05</td>
<td>B05</td>
<td>C05</td>
<td>NES05</td>
</tr>
<tr>
<td>A07</td>
<td>B07</td>
<td>C07</td>
<td>NES06</td>
</tr>
<tr>
<td>A08</td>
<td>B08</td>
<td>C08</td>
<td>NES07</td>
</tr>
<tr>
<td>A09</td>
<td>B09</td>
<td>&lt; 500 reads</td>
<td>NES08</td>
</tr>
<tr>
<td>A10</td>
<td>&lt; 500 reads</td>
<td>C10</td>
<td>NES09</td>
</tr>
</tbody>
</table>
| A11   |    B11 |  C11 |     10  8  8  9

Table showing samples that were subset from earlier 120 samples

Definitions

Logit Transformation: A logit is defined as the logarithm of the odds. If p is the probability of an event, then \((1−p)\) is the probability of not observing the event, and the odds of the event are \(p/(1−p)\). The logit of a number \(p\) between 0 and 1 is given by the formula:
\[
\text{logit}(p) = \log\left(\frac{p}{1−p}\right)
\]

Principal component analysis (PCA): This method reduces the dimensions of a complex data set allowing visualisation of complex data. It converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components using orthogonal transformation. The first principal component has the largest possible variance and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to (i.e., uncorrelated with) the preceding components.

Principal co-ordinate analysis (PCoA): Principal Coordinates Analysis (PCoA, = Multidimensional scaling, MDS) is a method to explore and to visualize similarities or dissimilarities of data. It starts with a similarity matrix or dissimilarity matrix (= distance matrix) and assigns for each item a location in a low-dimensional space, e.g. as a 3D graphics.

Rand index: This function calculates the Rand Index for two different clustering outcomes. The Rand Index gives a value between 0 and 1, where 1 means the two clustering outcomes match identically.

Sequence read (Phred) quality score: are assigned to each nucleotide base call in automated sequencer traces and used for determination of accurate, quality-based consensus sequence. Phred quality scores \(Q\) are defined as a property which is logarithmically related to the base-calling error probabilities \(P\).
\[
Q = -10 \log_{10} P
\]

Chao(S<sub>es</sub>): This diversity measure estimates the total species richness present within a community.
\[
S_{\text{est}} = S_{\text{obs}} + \frac{n_1^2}{2n_2}
\]
where,
$S_{\text{est}}$ = the estimated richness
$S_{\text{obs}}$ = the observed number of species
$n_1$ = the number of OTUs with only one sequence (i.e. "singletons")
$n_2$ = the number of OTUs with only two sequences (i.e. "doubletons")

**Shannon-Weiner diversity index (H):** The diversity index takes into account both abundance and evenness of species present in the community. It is defined as the proportion of species $i$ relative to the total number of species ($p_i$) is calculated, and then multiplied by the natural logarithm of this proportion ($\ln p_i$). The resulting product is summed across species ($S$), and multiplied by $-1$.

$$H = -\sum_{i=1}^{S} p_i \ln p_i$$

A high value of $H$ would be a representative of a diverse and equally distributed community and lower values represent less diverse community.

**Shannon equitability ($E_H$):** measures the evenness of a community and is calculated by diving the value of $H$ with $H_{\text{max}}$, which equals to $\ln S$($S$=number of species encountered). Its value ranges between 0 and 1, with being complete evenness.

$$E_H = H/H_{\text{max}} = H/\ln S$$

**Simpson diversity index (1-D):** This index also takes into account the relative abundance of the individual species present along with the number of species present and is represented by the equation:

$$\text{Simpson} = 1 - \sum (p_i)^2$$

The index has 0 to 1 measure with increasing number representing increasing diversity of a community.

**UniFrac distance measure:** UniFrac method performs a pairwise comparison between two organismal communities utilizing their phylogenetic information. A phylogenetic tree based on all the taxa present is drawn and branch leading to taxa from both samples is marked as "shared" and branches leading to taxa which are present only in one sample are marked as "unshared". The distance between the two samples is then calculated as (the sum of "unshared" branch lengths)/(the sum of all tree branch lengths (= shared+unshared)).

**Euclidean distance measure:** Euclidean distance $d(x,y)$ is given as the square root of the sum of squared differences between corresponding elements of the two vectors.

$$d(x,y) = \sqrt{\sum_{i=1}^{S} (x_i - y_i)^2}$$

**Bray curtis distance measure:** is a statistic used to quantify the compositional dissimilarity between two different samples, based on counts at each sample and given by the formula:
\[ d^{BCD}(i, j) = \frac{\sum_{k=0}^{n-1} |y_{i,k} - y_{j,k}|}{\sum_{k=0}^{n-1} (y_{i,k} + y_{j,k})} \]

\( d^{BCD} \) is the Bray-Curtis dissimilarity between the objects \( i \) and \( j \), \( k \) is the index of a variable and \( n \) is the total number of variables.

**Morisita Horn distance measure:** The dissimilarity between the structures of two communities is given by the following formula.

\[
D_{Morisita-Horn} = 1 - 2 \frac{\sum S_{A,i} S_{B,i}}{\sum (\frac{S_{A,i}}{n})^2 + \sum (\frac{S_{B,i}}{m})^2}
\]

where,

- \( S_{A,i} \): the number of individuals from community A in the \( i \)th OTU
- \( S_{B,i} \): the number of individuals from community B in the \( i \)th OTU
- \( n \): the number of individuals in community A
- \( m \): the number of individuals in community B

**Permanova:** This method allows to test the simultaneous response of multiple variables (e.g., bacterial genera) to one or more factors (e.g., various time points of COPD stages i.e. visits) in an ANOVA experimental design on the basis of any distance measure, using permutation methods. The permutation method repeatedly performs random shuffling of factor labels and each time, a new value of \( F(\pi) \), is calculated. If the null hypothesis were true, then the \( F \)-statistic actually obtained with the real ordering of the data relative to the factor labels will be similar to the values obtained under permutation. If, however, there is a significant effect of treatments, then the value of \( F \) obtained with the real ordering will appear large relative to the distribution of values obtained under permutation.

**Scripts for Qiime**

Qiime scripts for analysing the SEFR set

```
sudo mount -t vboxsf Qiime Shared_Folder/
# checking my mapped files are fine
check_id_map.py -m stable_mapping2.txt -o mapping_output/
check_id_map.py -m exacerbation_mapping2.txt -o mapping_output/
check_id_map.py -m followup_mapping2.txt -o mapping_output/
check_id_map.py -m recovery_mapping2.txt -o mapping_output/

# assign samples to multiplex reads
```
# Remove unwanted taxa

denoise_nhCOPDfil_chimera.fasta

usearch4.2.66_win32.exe
database.start>run>cmd

## to remove the bacterial chimeras using uchime..downloaded usearch and put it into the directory of input file and ref

# Remove sequences that did not align

chimeric sequences

####### Now go through and remove the bacterial chimeras using uchime..downloaded usearch and put it into the directory of input file and ref

denoised/S/centroids.fasta, denoised/E/centroids.fasta, denoised/F/centroids.fasta, denoised/R/centroids.fasta

e 1e

denoised/S/denoiser_mapping.txt, denoised/E/denoiser_mapping.txt, denoised/F/denoiser_mapping.txt, denoised/R/denoiser_mapping.txt

# COPD

denoise_wrapper.py

denoise_COPDfil2.fasta

# COPD

sffinfo GXCYKM104.sff > GXCYKM104.sff.txt

sffinfo GXCYKM103.sff > GXCYKM103.sff.txt

sffinfo GXCYKM104.sff > GXCYKM104.sff.txt

#SEQUENCE FILTERING AND CURATION #

########### SEQUENCE FILTERING AND CURATION ###########

#1) Blast against a chicken filter file containing chicken 18S rRNA and mitochondrial sequences (obtained from NCBI)

# The Filter file used will of course vary depending on host and will need to be compiled for each host.

# For samples from humans, download the human genomic and transcripts blast database from NCBI ftp://ftp.ncbi.nih.gov/

# This consists of 3 databases, so do the blast hits iteratively

exclude seqs by blast.py -i COPD_denoised.fasta -d ~/Desktop/Shared_Folder/Denoising/denoise_nonhumanCOPD_failures.fasta -o humanblastCOPD_e 1e-10 -p -m 100 -w 28 -n

exclude seqs by blast.py -i humanblastCOPD.non-matching -d ~/Desktop/Shared_Folder/Denoising/denoise_nonhumanCOPD_failures.fasta -o humanblastCOPD_e 1e-10 -p -m 100 -w 28 -n

exclude seqs by blast.py -i humanblastCOPD.non-matching -d ~/Desktop/Shared_Folder/Denoising/denoise_nonhumanCOPD_failures.fasta -o humanblastCOPD_e 1e-10 -p -m 100 -w 28 -n

# aligning the bacterial sequence

parallel_align_seqs_pynast.py -i denoise_nonhumanCOPD.fna -o aligned/ -a uclust -e 150 -p 0.75 -O 32 -t

Greengenes/Align/core_set_aligned.fasta.imputed

#### Now go through and remove non-bacterial 16S rRNA sequences, which failed to align to the reference 16S database, and chimeric sequences

# Remove sequences that did not align- using the fasta file containing sequences that did not align

filter_fasta.py -i denoise_nonhumanCOPD.fna -a ~/Desktop/Shared_Folder/Denoising/denoise_nonhumanCOPD_failures.fasta -o denoise_nonhumanCOPDfiltered.fasta -n

## to remove the bacterial chimeras using uchime..downloaded usearch and put it into the directory of input file and ref database.start=run='cmd'=go to the

usearch4.2.66_win32.exe --uchime denoise_nonhumanCOPDfiltered.fasta --db gold.fda --uchimeout denoise_nhcpCOPDf11.nonchimeric.uchime --uchimealns denoise_nhcpCOPDf11.chimera.aln --chimeras denoise_nhcpCOPDf11.chimeras.fasta --nonchimeras denoise_nhcpCOPDf11_good.fasta

# 4) Remove unwanted taxa

assign_taxonomy.py -i denoise_COPDf12.fasta -m rd-p -c 0.80 -o rdp_assignmentdenoiseCOPD/

# Write the sequence ID of sequences classified as being Cyanobacteria

261
grep "Cyanobacteria" rdp_assignmentdenoiseCOPD/denoise_COPDfil12_tax_assignments.txt >
rdp_assignmentdenoiseCOPD/denoiseCOPD_Cyanobacteria.txt
#Remove sequences with these ID from our fasta file
cd ...
filter_fasta.py -d denoise_COPDfil12.fasta -s rdp_assignmentdenoiseCOPD/denoiseCOPD_Cyanobacteria.txt -o denoiseCOPD_filtered3.fasta -n
#To count how many Cyanobacteria sequences we removed
grep -c "Cyanobacteria" rdp_assignmentdenoiseCOPD/denoiseCOPDfil12_tax_assignments.txt

#5) Removal by Sample Name
#There are some cases when we might want to remove certain samples altogether with >500 reads
extract_seq_by_sample_id.py -i denoiseCOPD_filtered3.fasta -s S4,S21,S23,S25,S27,E1,E5,E21,E22,E23,E28,E29,F4,F21,F23,F27,R9,R21,R23 -o denoiseCOPD_filtered4.fasta -n

# filter distance matrix from longitudinal samples from stable and exacerbation visit
data through plots.py
wfs_bdiv_even620/lonwei_unifrac_dm.txt
for compari
# stable to exacerbation change in smoking status, copd severity, bactreial associated exacerbation change

## to filter out longitudinal samples distance matrix only
filter_distance_matrix.py -i wf_bdiv_even620/weighted_unifrac_dm.txt -o lonwei_unifrac_dm.txt --sample_id_fp longisample_id_list.txt
filter_distance_matrix.py -i wf_bdiv_even620/unweighted_unifrac_dm.txt -o lonunwei_unifrac_dm.txt --sample_id_fp longisample_id_list.txt

## to filter out longitudinal samples from stable and exacerbation visit distance matrix from lonwei_unifrac_dm.txt for comparing stable to exacerbation change in smoking status, copd severity, bactreial associated exacerbation

#QIIME also has a workflow script for beta diversity calculations
#This difference in sequencing depth may affect the apparent microbial community composition (eg new OTUs may be detected as an artefact of deeper sequencing)
#It is likely that there is a different number of sequences associated with each sample.
#It will be useful to look at how many sequences we are getting for each sample
per_library_stats.py -i otu_table.txt
summarize_taxa_through_plots.py -i otu_table.txt -o wf_taxa_summary -m COPD454Mapping.txt

cat stable_mapping2.txt excursion_mapping2.txt followup_mapping2.txt > COPD454Mapping.txt

## 1. Taxonomy based classification

#Classify the sequences
assign_taxonomy.py -i denoiseCOPD_filtered4.fasta -m rdp -c 0.80 -o rdp_assignmentdenoiseCOPDfil/

## 2. OTU- and Phylogenetic-based approach using QIIME

pick_ottus_through_ottu_table.py -i denoiseCOPD_filtered4.fasta -p ~/Desktop/Shared_Folder/custom_parameters5.txt -o otus

#It will be useful to look at how many sequences we are getting for each sample
per_library_stats.py -i otus/otu_table.txt
summarize_taxa_through_plots.py -i otu_table.txt -o wf_taxa_summary -m COPD454Mapping.txt

cat stable_mapping2.txt excursion_mapping2.txt followup_mapping2.txt > COPD454Mapping.txt

## ALPHA DIVERSITY USING QIIME

#Create a custom parameters file (eg alpha_params.txt) containing the metrics that you would like to be calculated
alpha_rarefaction.py -i otu_table.txt -m COPD454Mapping.txt -o wf_arare/ -p ~/Desktop/Shared_Folder/custom_parameters1.txt -t rep_set.tre

## BETA DIVERSITY

## Determining an appropriate sequencing depth to apply to all samples

#BETTY CURTISS and moristisa-horn diversity measure

beta_diversity_through_plots.py -i otu_table.txt -m COPD454Mapping.txt -o wf_bdiv_even620 BCMH/ -t rep_set.tre -e 620 -p ~/Desktop/Shared_Folder/custom_parameters5.txt

# for visits
beta_diversity_through_plots.py -i otu_table.txt -m COPD454Mapping.txt -o wf_bdiv_even620/ -t rep_set.tre -e 620

# for only stable, exacerbation, recovery and longitudinal data
beta_diversity_through_plots.py -i otu_table.txt -m PCllusterMapping.txt -o wf_bdivPF_even620_1/ -t rep_set.tre -e 620
beta_diversity_through_plots.py -i otu_table.txt -m stableMapping.txt -o wf_bdivS_even620_1/ -t rep_set.tre -e 620
beta_diversity_through_plots.py -i otu_table.txt -m excorrMapping.txt -o wf_bdivE_even620_1/ -t rep_set.tre -e 620
beta_diversity_through_plots.py -i otu_table.txt -m recMapping.txt -o wf_bdivR_even620_1/ -t rep_set.tre -e 620

# bray curtis and moristisa-horn diversity measure
beta_diversity_through_plots.py -i otu_table.txt -m COPD454Mapping.txt -o wf_bdiv_even620 BCMH/ -t rep_set.tre -e 620 -p ~/Desktop/Shared_Folder/custom_parameters5.txt

# to filter out longitudinal samples distance matrix only
filter_distance_matrix.py -i wf_bdiv_even620/weighted_unifrac_dm.txt -o lonwei_unifrac_dm.txt --sample_id_fp longisample_id_list.txt
filter_distance_matrix.py -i wf_bdiv_even620/unweighted_unifrac_dm.txt -o lonunwei_unifrac_dm.txt --sample_id_fp longisample_id_list.txt

# to filter out longitudinal samples from stable and exacerbation visit distance matrix from lonwei_unifrac_dm.txt for comparing stable to exacerbation change in smoking status, copd severity, bactreial associated exacerbation

filter_distance_matrix.py -i wf_bdiv_even620/lonwei_unifrac_dm.txt -o SE_lonwei_unifrac_dm.txt --sample_id_fp longisampleSE_id_list.txt
filter_distance_matrix.py -i wf_bdiv_even620/lonwei_unifrac_dm.txt -o ER_lonwei_unifrac_dm.txt --sample_id_fp longisampleER_id_list.txt

262
principal_coordinates.py -i wf_bdiv_even620/SE_lonwei_unifrac_dm.txt -o SE_lonwei_unifrac_coords.txt
principal_coordinates.py -i wf_bdiv_even620/ER_lonwei_unifrac_dm.txt -o ER_lonwei_unifrac_coords.txt

## making 3d biplot
make_3d_plots.py -i wf_bdiv_even620/weighted_unifrac_pc.txt -m COPD454Mapping.txt -t wf_taxa_summary1/otu_table_L6.txt -n taxa_kep 10 -o 3d_biplotweight1
make_3d_plots.py -i wf_bdiv_even620/unweighted_unifrac_pc.txt -m COPD454Mapping.txt -t wf_taxa_summary1/otu_table_L6.txt -n taxa_kep 10 -o 3d_biplotunweight2

## make distance boxplots
make_distance_boxplots.py -d wf_bdiv_even620/unweighted_unifrac_dm.txt -m COPD454Mapping.txt -f "Visit" -o unweightedboxplot1 --suppress_all_between --suppress_all_within -n 999
make_distance_boxplots.py -d wf_bdiv_even620/weighted_unifrac_dm.txt -m COPD454Mapping.txt -f "Visit" -o weightedboxplot --suppress_all_between --suppress_all_within -n 999
make_distance_boxplots.py -d wf_bdiv_even620/lonunwei_unifrac_dm.txt -m PFclusterMapping.txt -f "Individual" -o lonweightedboxplot1 --suppress_all_between --suppress_all_within --save_raw_data
make_distance_boxplots.py -d wf_bdiv_even620/lonunwei_unifrac_dm.txt -m PFclusterMapping.txt -f "Individual" -o lonunweightedboxplot --suppress_all_between --suppress_all_within --save_raw_data -n 999

## make distance boxplots for smoking, COPD severity at stable and exacerbation state
make_distance_boxplots.py -d wf_bdivS_even620/weighted_unifrac_dm.txt -m stableMapping.txt -f "Smoking" -o weightedsmokingboxplot1 --suppress_all_between --suppress_all_within --save_raw_data -n 999
make_distance_boxplots.py -d wf_bdivS_even620/weighted_unifrac_dm.txt -m exaMapping.txt -f "Smoking" -o weightedexasmokingboxplot1 --suppress_all_between --suppress_all_within --save_raw_data -n 999
make_distance_boxplots.py -d wf_bdivS_even620/weighted_unifrac_dm.txt -m COPD454Mapping.txt -f "COPDSeverity" -o weighteddasmokingboxplot1 --suppress_all_between --suppress_all_within --save_raw_data -n 999
make_distance_boxplots.py -d wf_bdivS_even620/weighted_unifrac_dm.txt -m exaMapping.txt -f "COPDSeverity" -o weighteddasmokingboxplot1 --suppress_all_between --suppress_all_within --save_raw_data -n 999
make_distance_boxplots.py -d wf_bdivS_even620/weighted_unifrac_dm.txt -m COPD454Mapping.txt -f "COPDSeverity" -o weighteddasmokingboxplot1 --suppress_all_between --suppress_all_within --save_raw_data -n 999
make_distance_boxplots.py -d wf_bdivS_even620/weighted_unifrac_dm.txt -m COPD454Mapping.txt -f "COPDSeverity" -o weighteddasmokingboxplot1 --suppress_all_between --suppress_all_within --save_raw_data -n 999

## make distance boxplots for Visit
make_distance_boxplots.py -d wf_bdiv_even620/weighted_unifrac_dm.txt -m COPD454Mapping.txt -f "Visit" -o weightedboxplot --suppress_all_between --suppress_all_within -n 999
make_distance_boxplots.py -d wf_bdiv_even620/weighted_unifrac_dm.txt -m COPD454Mapping.txt -f "Visit" -o weightedboxplot --suppress_all_between --suppress_all_within -n 999
make_distance_boxplots.py -d wf_bdiv_even620/weighted_unifrac_dm.txt -m COPD454Mapping.txt -f "Visit" -o weightedboxplot --suppress_all_between --suppress_all_within -n 999
make_distance_boxplots.py -d wf_bdiv_even620/weighted_unifrac_dm.txt -m COPD454Mapping.txt -f "Visit" -o weightedboxplot --suppress_all_between --suppress_all_within -n 999

## comparing categories.py for visit groups
compare_categories.py --method permanova -i wf_bdiv_even620/weighted_unifrac_dm.txt -m COPD454Mapping2.txt -c Visit -o permanovaout -n 999
compare_categories.py --method permanova -i wf_bdiv_even620/unweighted_unifrac_dm.txt -m COPD454Mapping2.txt -c Visit -o permanovaunweii_out -n 999

Qime scripts to analyse SSE and NES set

# checking my mapped files are fine
check_id_map.py -m stableseq_mapping.txt -o mapping_output/
check_id_map.py -m stableseq1_mapping.txt -o mapping_output/

# assign samples to multiplex reads
split_libraries.py -m stableseq_mapping.txt -f C_H2EEAF01.fna -q C_H2EEAF01.qual -o split_library_output/SS1 -l 200 -L 1000 -s 25 -a 0 -H 8 -M 2 -h 10 -e 1.5
split_libraries.py -m stableseq1_mapping.txt -f sample_D/D_HJK0XVJ03.fna -q sample_D/D_HJK0XVJ03.qual -o split_library_output/SS2 -l 200 -L 1000 -s 25 -a 0 -H 8 -M 2 -h 10 -e 1.5

#Load the Roche sequencer module (version number may differ)
module load gs454/2.5.3

# Use the sffinfo command - process each region/sff file separately
sff files in the scratch/micro/kh132/stable_454
sffinfo C_H2EEAF01.sff > C_H2EEAF01.sff.txt
sffinfo D_HJK0XVJ03.sff > D_HJK0XVJ03.sff.txt
denoise_wrapper.py -v -i C_H2EEAF01.sff.txt -f split_library_output/SS1/seqs.fna -o denoised/SS1/ -m stableseq_mapping.txt --titanium
denoise_wrapper.py -v -i D_HJK0XVJ03.sff.txt -f split_library_output/SS2/seqs.fna -o denoised/SS2/ -m stableseq1_mapping.txt --titanium

#remove unmapped reads
filter_fasta.py -f denoisestable_nonhuman.fna -o denoisestable/denoised/denoised_unmatched.fasta -n

to remove bacterial chimeras using uchime.
C:/Qiime/Denoiser/NonHuman

assign_taxonomy.py -i denoised/denoised/denoised_unmatched.fasta -m rdp -e 1e-10 -p 0.97 -m 100 -w 28 -n

#to remove samples with low reads <500 reads. sample names are B2,C2,A6,B6,C6,C9,B10
filter_fasta.py -f denoisestable_nonhuman.fna -o denoisestable/denoised/denoised_unmatched.fasta -n
denoisedstable_nonhumanfiltered.fasta

#4) Remove unwanted taxa
assign_taxonomy.py -i denoisedstable_filtered2.fasta -m rdp -e 0.80 -o rdp_assignmentdenoisedstable/
grep "Cyanobacteria" rdp_assignmentdenoisedstable/denoisedstable_filtered2_tax_assignments.txt >
rdp_assignmentdenoisedstable/denoisedstable_Cyanobacteria.txt
filter_fasta.py -f denoisedstable_filtered2.fasta -s rdp_assignmentdenoisedstable/denoisedstable_Cyanobacteria.txt -o denoisedstable_filtered3.fasta -n

#now need to remove samples with low reads <500 reads. sample names are B2,C2,A6,B6,C6,C9,B10
extract_seqs_by_sample_id.py -i denoisedstable_filtered3.fasta -o denoisedstable_filtered4.fasta -s B2,C2,A6,B6,C6,C9,B10

#to give assignment only till genus level in 1.4.0 try doing it in qiime 1.3.0

to filter out longitudinal samples distance matrix only
filter_distance_matrix.py -i wf_bdivsub_even1710/unweighted_unifrac_dm.txt -m subjectMapping.txt -t rep_set.tre

beta_diversity_through_plots.py -i otu_table.txt -m stable454Mapping.txt -o wf_taxa_summary -m stable454Mapping.txt

alpha_rarefaction.py -i otu_table.txt -m stable454Mapping.txt -o wf_taxa/alpha_rarefaction.txt

#QIIME also has a workflow script for beta diversity calculations
beta_diversity_through_plots.py -t rep_set.tre -e 500

#to obtain the loadings co-ordinates
make_3dplots.py -i wf_bdivsub_even1700/unweighted_unifrac_pc.txt -m subjectMapping.txt -t

#to filter out longitudinal samples distance matrix only
filter_distance_matrix.py -i wf_bdivsub_even1700/unweighted_unifrac_dm.txt -m subjectMapping.txt -t wf_taxa_summary/otu_table_L6.txt -o "taxa_keep 10 -o 3d_biplotunweight
make_3dplots.py -i wf_bdivsub_even1700/weighted_unifrac_dm.txt -m subjectMapping.txt -t wf_taxa_summary/otu_table_L6.txt -o "taxa_keep 10 -o 3d_biplotweight

#to compare 3d plots of the repeated stable samples in SEFR and SSE samples
cc stable_454/taxonomy.otu_qiime/otufilter_distance_matrix.py -i wf_bdivsub_even500/weighted_unifrac_dm.txt -o repeatedstable1_weuinfrac_dm.txt --sample_id_fp repeatedstable1_samplelist.txt

264
R scripts for heatmap, hierarchical clustering and PCA plot

R scripts for SEFR set

ClassifierTable1 <- read.table("D:/Koirobi/Qiime/COPDdenoise_454/Denoising/NonhumanCOPD/rdp_assignmentdenoiseCOPDfil/rdpgenuscount_R1.txt", sep="\t", header=T)

#Remove last row containing total counts
ClassifierTable1 <- ClassifierTable1[1:(length(ClassifierTable1[,1])-2),]

#Order according to Phylum name, then Class etc
ClassifierTable1 <- ClassifierTable1[order(ClassifierTable1[, "Phylum"], ClassifierTable1[, "Class"], ClassifierTable1[, "Order"], ClassifierTable1[, "Family"], ClassifierTable1[, "Genus"]),]

#For plotting rdp genus
ClassifierTable1.Filter <- ClassifierTable1[, "TotalCounts"] >= 50 & ClassifierTable1[, "Max"] >= 0.01 ,
write.table(ClassifierTable1.Filter, "D:/Koirobi/Qiime/COPDdenoise_454/Denoising/NonhumanCOPD/rdp_assignmentdenoiseCOPDfil/GenusCount49.txt", sep="\t")

ClassifierTable1.Datafilt <- ClassifierTable1.Filter[6:(length(ClassifierTable1.Filter[1:]-2))]

#Logit transformation
ClassifierTable1.lg <- log(ClassifierTable1.Datafilt/(1-ClassifierTable1.Datafilt)+0.00001)

GenusTableNamesfilt1 <- paste(ClassifierTable1.Filter[,1], ClassifierTable1.Filter[,5], sep=" | ")

#Set column colours
ColColours <- NULL
ColColours[grep("S", colnames(ClassifierTable1.Datafilt))]<- "green"
ColColours[grep("E", colnames(ClassifierTable1.Datafilt))] <- "red"
ColColours[grep("F", colnames(ClassifierTable1.Datafilt ))]<- "yellow"
ColColours[grep("R", colnames(ClassifierTable1.Datafilt ))]<- "blue"

#[Optional] Set up column separators for the heatmap- this will vary depending on your samples and how you want your heatmap to look
colsepx <- 1:101
colsexcude <- c(3,7,11,13,16,20,24,28,31,35,39,43,47,51,55,59,63,67,71,75,78,82,85,89,91,94,97)
colsepx <- setdiff(colsepx, colsexcude)

#Draw heatmap
matrix.min<- min(ClassifierTable1.Datafilt[ClassifierTable1.Datafilt!=0])
matrix.max<- max(ClassifierTable1.Datafilt)
pairs.breaks<- c(0, seq( matrix.min, matrix.max, length.out=11) )
color.palette=colorRampPalette(c("green","yellow","pink","purple"))(10)

heatmap.2(as.matrix(ClassifierTable1.Datafilt), Rowv=F, Colv=F, ColSideColors=ColColours ,col=c("white",color.palette),breaks=pairs.breaks,labRow=GenusTableNamesfilt1, scale="none", key=T, na.rm=T, xlab="Sample", ylab="Genus", trace="none", margin=c(3,15), colsep=colsepx, rowsep=c(0:49), sepcolor="black", sepwidth=c(0.0005,0.0005), add.expres=cf abline(v=(colexclude + 0.5), col="red", lwd=2) )

heatmap.2(as.matrix(ClassifierTable1.Datafilt), Rowv=F, Colv=T, ColSideColors=ColColours ,col=c("white",color.palette),breaks=pairs.breaks,labRow=GenusTableNamesfilt1, scale="none", key=T, na.rm=T, xlab="Sample", ylab="Genus", trace="none", margin=c(3,15), colsep=colsepx, rowsep=c(0:49), sepcolor="black", sepwidth=c(0.0005,0.0005 ) )

##heatmap clustering logit transformed
matrix.min<- min(ClassifierTable1.lg[ClassifierTable1.lg!=<11.00])
matrix.max<- max(ClassifierTable1.lg)
pairs.breaks<- c(-11.00, seq( matrix.min, matrix.max, length.out=17) )
color.palette=colorRampPalette(c("green","yellow","pink","purple"))
heatmap.2(as.matrix(ClassifierTable1.lg), Rowv=F, Colv=T, ColSideColors=ColColours ,col=c("white",color.palette),breaks=pairs.breaks,labRow=GenusTableNamesfilt1, scale="none", key=T, na.rm=T, xlab="Sample", ylab="Genus", trace="none", margin=c(3,15), colsep=colsepx, rowsep=c(0:49), sepcolor="black", sepwidth=c(0.0005,0.0005 ) )

##drawing PCA plots
ClassifierTable1 <- read.table("D:/Koirobi/Qiime/COPDdenoise_454/Denoising/NonhumanCOPD/rdp_assignmentdenoiseCOPDfil/rdpgenuscount_R1.txt", sep="\t", header=T)
#Remove last row containing total counts
ClassifierTable1 <- ClassifierTable1[ 1: (length(ClassifierTable1[,1])-2),]
library(MASS)
ClassifierTable.Data <- ClassifierTable1[, 6:(length(ClassifierTable1[,1])-2)]
ClassifierTable.pca <- prcomp(t(ClassifierTable.Data), retx=T)
eqscplot(ClassifierTable.pca$x[,1:2], main="PCA plot of Classification at Genus level using relative abundances", pch=".", xlab="PC1(60.25% variance)", ylab= "PC2(17.93% variance)"
#Label points with site names and colour accordingly
S <- grep("S", colnames(ClassifierTable.Data))
E<- grep("E", colnames(ClassifierTable.Data))
F<- grep("F", colnames(ClassifierTable.Data))
R<- grep("R", colnames(ClassifierTable.Data))
text(ClassifierTable.pca$x[S,1:2], colnames(ClassifierTable.Data)[S], cex=0.9, col="green")
text(ClassifierTable.pca$x[E,1:2], colnames(ClassifierTable.Data)[E], cex=0.9, col="red")
text(ClassifierTable.pca$x[F,1:2], colnames(ClassifierTable.Data)[F], cex=0.9, col="yellow")
text(ClassifierTable.pca$x[R,1:2], colnames(ClassifierTable.Data)[R], cex=0.9, col="blue")

Generadata <- read.table("D:/Koirobi/Qiime/COPDdenoise_454/Denoising/NonhumanCOPD/rdp_assignmentdenoiseCOPDfil/genera.txt", header=T)
#Remove last row containing total counts and sum & last column containing Totalcounts and Max
Generadata <- Generadata[ 1: (length(Generadata[,1])-2),1:(length(Generadata[,1])-2)]
Generadata.pca <- prcomp(t(Generadata), retx=T)
eqscplot(Generadata.pca$rotation[,1:2], main="Plot of Genera level loadings", pch=".", xlab="PC1", ylab= "PC2")
text(Generadata.pca$rotation[,1:2], rownames(Generadata),cex=0.8,col="Purple")

###Plotting Heatmap for Phylum level(RDP) in R ( all phyla)
PhylumTable <- read.table("D:/Koirobi/Qiime/COPDdenoise_454/Denoising/NonhumanCOPD/rdp_assignmentdenoiseCOPDfil/Phylumcount_for_R.txt", sep="\t", header=T)
#Remove last row containing total counts and sum & last column containing Totalcounts and Max
PhylumTable <- PhylumTable[ 1: (length(PhylumTable[,1])-2),1:(length(PhylumTable[,1])-2)]

#Make the first column the row name and then remove the column
PhylumTableNames <- PhylumTable[,1]
PhylumTable <- PhylumTable[2:length(PhylumTable[,1])]

# Set up column separators for the heatmap- this will vary depending on your samples and how you want your heatmap to look
colsepx <- c(3,7,11,13,16,20,24,28,31,35,39,43,47,51,55,59,63,67,71,75,78,82,85,89,91,94,97)
colsepx <- setdiff(colsepx, colexclude)

#Set column colours
ColColours <- NULL
ColColours[grep("S", colnames(PhylumTable))]<- "green"
ColColours[grep("E", colnames(PhylumTable ))]<- "red"
# Draw heatmap
matrix.min <- min(PhylumTable[PhylumTable != 0])
matrix.max <- max(PhylumTable)
pairs.breaks <- c(0, seq( matrix.min, matrix.max, length.out = 11 ) )
color.palette = colorRampPalette(c("green","yellow","pink","purple"))(10)

heatmap.2(as.matrix(PhylumTable), Rowv = F, Colv = F, ColSideColors = ColColours, col = c("white",color.palette), breaks = pairs.breaks, labRow = PhylumTableNames, scale = "none", ylab = "Phylum", trace = "none", margin = c(10, 10), colsep = colsep, rowsep = c(0:16), sepal = "black", sepwidth = c(0.0005, 0.0005), add.expr = c( abline(v = (colexclude + 0.5), col = "red", lwd = 2 )) )

heatmap.2(as.matrix(PhylumTable), Rowv = F, Colv = T, ColSideColors = ColColours, col = c("white",color.palette), breaks = pairs.breaks, labRow = PhylumTableNames, scale = "none", ylab = "Phylum", trace = "none", margin = c(10, 10), colsep = colsep, rowsep = c(0:16), sepal = "black", sepwidth = c(0.0005, 0.0005))

## log change in proteo to firmratio
RatioChange <- read.table("C:/Users/bashi/Qiime/Denoising/NonhumanCOPD/rdp_assignmentdenoiseCOPDfil/ProteoFirmRatio.txt", header=T)

heatmap.2(t(RatioChange), Rowv = F, Colv = T, col = maPalette(8, high = "yellow", mid = "green", low = "blue"), scale = "none", breaks = seq(from = -3, to = 3, length = 9), na.rm = T, xlab = "Subject", ylab = "Visits", trace = "none", cexRow = 0.8)
heatmap.2(t(RatioChange), Rowv = F, Colv = T, distfun = function(RatioChange) dist(RatioChange, method = 'euclidean'), hclustfun = function(RatioChange) hclust(RatioChange, method = 'average'), col = maPalette(8, high = "yellow", mid = "green", low = "blue"), scale = "none", breaks = seq(from = -3, to = 3, length = 9), na.rm = T, xlab = "Subject", ylab = "Visits", trace = "none", cexRow = 0.8)

# bootstrapping the cluster analysis
cluster.visits <- pvclust(RatioChange, method.hclust = "average", method.dist = "euclidean", nboot = 1000)
plot(cluster.visits)
class.samples <- pvclust(t(RatioChange), method.hclust = "average", method.dist = "euclidean", nboot = 1000)
plot(cluster.samples)
pvrect(cluster.samples, alpha = 0.95)

R script for SSE and NES set

ClassifierTable <- read.table("D:/Koirobi/Qiime/stable_454/taxonomy_otu_qiime/rdp_analysis/Allgenuscount_R1.txt", sep = "\t", header = T)
# Remove last row containing total counts
ClassifierTable <- ClassifierTable[1:(length(ClassifierTable[,1])-2),]

# Filter our rows that don't meet criteria eg total count of >29 or >1% in at least one sample
ClassifierTable.Filter1 <- ClassifierTable[ ClassifierTable[, "TotalCounts"] >= 29 & ClassifierTable[, "Max"] >= 0.01 , ]
# Out of 143 at genus level remaining 45 genera for 0.02% total abundance.
# Order according to Phylum name, then Class etc
ClassifierTable.Filter <- ClassifierTable.Filter[order(ClassifierTable.Filter[, "Phylum"], ClassifierTable.Filter[, "Classs"], ClassifierTable.Filter[, "Order"], ClassifierTable.Filter[, "Family"], ClassifierTable.Filter[, "Genus"]), ]

# To make separate heatmaps for the sequential stable samples and baseline samples of exacerbeted and nonexacerbated stable sample
write.table(ClassifierTable.Filter, "D:/Koirobi/Qiime/stable_454/taxonomy_otu_qiime/rdp_analysis/dominant_genuscount1.txt", sep = "\t")

# To plot the filtered genus groups in the exacerbated and non-exacerbated group stable groups
ClassifierTable_S1 <- read.table("D:/Koirobi/Qiime/stable_454/taxonomy_otu_qiime/rdp_analysis/dge_A_NES.txt", sep = "\t", header = T)
# Order according to Phylum name, then Class etc
ClassifierTable_S1 <- ClassifierTable_S1[order(ClassifierTable_S1[, "Phylum"], ClassifierTable_S1[, "Classs"], ClassifierTable_S1[, "Order"], ClassifierTable_S1[, "Family"], ClassifierTable_S1[, "Genus"]), ]
ClassifierTable_S1filt <- ClassifierTable_S1[6:(length(ClassifierTable_S1[,1]))-0]

GenusTableNames_S1 <- paste(ClassifierTable_S1[,1], ClassifierTable_S1[,5], sep = " | ")
# Set column colours
ColColours <- NULL
ColColours[grep("A", colnames(ClassifierTable_S1filt))] <- "red"
ColColours[grep("NES", colnames(ClassifierTable_S1filt))] <- "green"
library(gplots)
library(marray)
matrix.min <- min(ClassifierTable_S1filt[ClassifierTable_S1filt!=0])
matrix.max <- max(ClassifierTable_S1filt)
pairs.breaks <- c(0, seq(matrix.min, matrix.max, length.out=11))
color.palette <- colorRampPalette(c("green","yellow","pink","purple"))(10)

# Draw heatmap
heatmap.2(as.matrix(ClassifierTable_S1filt), Rowv=NULL, Colv=TRUE, dendrogram= "column", ColSideColors=ColColours, col=c("white",color.palette), labRow=GenusTableNames_S1, breaks=pairs.breaks, scale="none", key=T, na.rm=T, xlab="Sample", ylab="Genus", trace="none", margin=c(5,30), rowsep=c(0:43), colsep=c(1:19), sepcolor="black", sepwidth=c(0.0005,0.0005))

### To plot the filtered genus groups in the sequential stable groups
ClassifierTable_S <- read.table("D:/Koirobi/Qiime/stable_454/taxonomy_otu_qiime/rdp_analysis/dgc_withoutNES.txt", sep="\t", header=T)

# Order according to Phylum name, then Class etc
ClassifierTable_S <- ClassifierTable_S[order(ClassifierTable_S[, "Phylum"], ClassifierTable_S[, "Classs"], ClassifierTable_S[, "Order"], ClassifierTable_S[, "Family"], ClassifierTable_S[, "Genus"]), ]
ClassifierTable_Sfilt <- ClassifierTable_S[, 6:(length(ClassifierTable_S[1,])-0)]
GenusTableNames_S <- paste(ClassifierTable_S[,1], ClassifierTable_S[,5], sep=" |")

# Set column colours
ColColours <- NULL
ColColours[grep("01", colnames(ClassifierTable_Sfilt))] <- "green"
ColColours[grep("02", colnames(ClassifierTable_Sfilt))] <- "white"
ColColours[grep("03", colnames(ClassifierTable_Sfilt))] <- "yellow"
ColColours[grep("04", colnames(ClassifierTable_Sfilt))] <- "violet"
ColColours[grep("05", colnames(ClassifierTable_Sfilt))] <- "orange"
ColColours[grep("07", colnames(ClassifierTable_Sfilt))] <- "red"
ColColours[grep("08", colnames(ClassifierTable_Sfilt))] <- "blue"
ColColours[grep("09", colnames(ClassifierTable_Sfilt))] <- "black"
ColColours[grep("10", colnames(ClassifierTable_Sfilt))] <- "grey"
ColColours[grep("11", colnames(ClassifierTable_Sfilt))] <- "pink"

library(gplots)
library(marray)

matrix.min <- min(ClassifierTable_Sfilt[ClassifierTable_Sfilt!=0])
matrix.max <- max(ClassifierTable_Sfilt)
pairs.breaks <- c(0, seq(matrix.min, matrix.max, length.out=11))
color.palette <- colorRampPalette(c("green","yellow","pink","purple"))(10)

# Draw heatmap
heatmap.2(as.matrix(ClassifierTable_Sfilt), Rowv=NULL, Colv=TRUE, dendrogram= "column", ColSideColors=ColColours, col=c("white",color.palette), labRow=GenusTableNames_S, breaks=pairs.breaks, scale="none", key=T, na.rm=T, xlab="Sample", ylab="Genus", trace="none", margin=c(3,25), rowsep=c(0:43), colsep=c(1:26), sepcolor="black", sepwidth=c(0.0005,0.0005))

### Plotting Heatmap for Phylum level (RDP) in R (all phyla)
PhylumTable <- read.table("D:/Koirobi/Qiime/stable_454/taxonomy_otu_qiime/rdp_analysis/R_stablephylumcount.txt", sep="\t", header=T)

# Remove last row containing total counts and sum & last column containing Totalcounts and Max
PhylumTable <- PhylumTable[1: (length(PhylumTable[,1])-2),1:(length(PhylumTable[,1])-2)]

# Make the first column the row name and then remove the column
PhylumTableNames <- PhylumTable[,1]
PhylumTable <- PhylumTable[,2:length(PhylumTable[,1])]

# [Optional] Set up column separators for the heatmap- this will vary depending on your samples and how you want your heatmap to look
colsepx <- 1:35
colexclude <- c(3,4,7,10,13,16,19,21,23,26)
colsepx <- setdiff(colsepx, colexclude)

# Set column colours
ColColours <- NULL
ColColours[grep("A", colnames(PhylumTable))] <- "green"
ColColours[grep("B", colnames(PhylumTable))] <- "red"
ColColours[grep("C", colnames(PhylumTable))] <- "yellow"
ColColours[grep("NES", colnames(PhylumTable))] <- "blue"

sepcolor="black", sepwidth=0.0005,0.0005), add.expr= c( abline(v=(colexclude + 0.5), col="white", lwd=2) )

268
R script for P:F ratio qPCR analysis

## Heatmap, clustering based on proteo to firm ratio in qPCR
RatioChange <- read.table("D:/Koirobi/Proteo_Firmanalysis/454sample_Rcorr.txt", header=T)
heatmap.2(t(RatioChange), Rowv=F, Colv=T, distfun=function(RatioChange)
dist(RatioChange,method="euclidean"), hclustfun=function(RatioChange) hclust(RatioChange,method="average"),
        col=maPalette(20, high="yellow", mid="green", low="blue"), scale="none",
        breaks=seq(from=-6, to=15, length=21), na.rm=T,
        xlab="Subject", ylab="Visits", trace="none", cexRow=0.8)

# bootstrapping the cluster analysis
cluster.visits <- pvclust(RatioChange, method.hclust="average", method.dist="euclidean", nboot=1000)
plot(cluster.visits)
cluster.samples <- pvclust(t(RatioChange), method.hclust="average", method.dist="euclidean", nboot=1000)
plot(cluster.samples)

## change in proteo to firmratio in qPCR of 454sample (all 120 samples)
RatioChange <- read.table("D:/Koirobi/Proteo_Firmanalysis/qPCR_454_PF/all454sample_Rcorr.txt", header=T)
heatmap.2(t(RatioChange), Rowv=F, Colv=T, distfun=function(RatioChange)
dist(RatioChange,method="euclidean"), hclustfun=function(RatioChange) hclust(RatioChange,method="average"),
        col=maPalette(20, high="yellow", mid="green", low="blue"), scale="none",
        breaks=seq(from=-6, to=15, length=21), na.rm=T,
        xlab="Subject", ylab="Visits", trace="none", cexRow=0.8)

# bootstrapping the cluster analysis
cluster.visits <- pvclust(RatioChange, method.hclust="average", method.dist="euclidean", nboot=1000)
plot(cluster.visits)
cluster.samples <- pvclust(t(RatioChange), method.hclust="average", method.dist="euclidean", nboot=1000)
plot(cluster.samples)

## Agreement of hierarchical clustering between 454 P:F ratio and qPCR P:F ratio
Cluster454 <- read.table("D:/Koirobi/Proteo_Firmanalysis/qPCR_454PF/MBiome_PFratio1.txt", header=T)
Cluster454_fit <- hclust(Cluster454_dist1, method="average")
plot(Cluster454_fit, main="Clustering 454 Proteo/Firm ratio", xlab="COPD subjects")

par(mfrow=c(2,1))
Cluster_qPCR <- read.table("D:/Koirobi/Proteo_Firmanalysis/qPCR_454PF/qPCR_PFratio.txt", header=T)
Cluster_qPCR_dist1 <- dist(Cluster_qPCR, method="euclidean")
Cluster_qPCR_fit <- hclust(Cluster_qPCR_dist1, method="average")
plot(Cluster_qPCR_fit, main="Clustering qPCR Proteo/Firm ratio", xlab="COPD subjects")

install.packages("fossil")
library(fossil)
rand.index(Cluster454_dist1, Cluster_qPCR_dist1)

## PCA plot agreement between 454 and qPCR P:F ratio
PCA_qPCR <- read.table("D:/Koirobi/Proteo_Firmanalysis/qPCR_454PF/PCA_qPCRPF.txt", header=T)

PC_qPCRFit <- prcomp(t(PCA_qPCR), retx=T)
library(MASS)
eqscplot(PC_qPCRFit[,1:2], pch=".", main="Grouping of qPCR derived P/Fratio samples based on 454Proteo-Firm clusters",
xlab="PC1",ylab="PC2",xlim=c(10,12),ylim=c(-10,10))
text(PC_qPCRFit[,1:2], colnames(PCA_qPCR), cex=0.7)
#Label points with site names and colour accordingly
HP <- grep("HP", rownames(pca_qPCRPF$s))
HF<- grep("HF", rownames(pca_qPCRPF$s))
PF<- grep("PF", rownames(pca_qPCRPF$s))
text(pca_qPCRPF$s [HP,1:2], rownames(pca_qPCRPF$s)[HP], cex=1, col="blue")
text(pca_qPCRPF$s [HF,1:2], rownames(pca_qPCRPF$s)[HF], cex=1, col="green")
text(pca_qPCRPF$s [PF,1:2], rownames(pca_qPCRPF$s)[PF], cex=1, col="red")