MICROBIAL COMMUNITIES OF THE RESPIRATORY TRACT

INVESTIGATIONS OF THE MICROBIAL COMMUNITIES OF THE RESPIRATORY TRACT IN THE ELDERLY AND IN CYSTIC FIBROSIS VIA CULTURE-DEPENDENT AND -INDEPENDENT APPROACHES

BY

FIONA J. WHELAN, M.SC., B.COMPSC.

A THESIS

SUBMITTED TO THE DEPARTMENT OF BIOCHEMISTRY & BIOMEDICAL SCIENCES

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Doctor of Philosophy (2017) McMaster University

(Biochemistry & Biomedical Sciences) Hamilton, Ontario, Canada

TITLE: Investigations of the microbial communities of the respiratory tract in the elderly and in cystic fibrosis via culture-dependent and -independent approaches

AUTHOR: Fiona J. Whelan

M.Sc., (Medical Sciences)

McMaster University, Hamilton, Canada

B.CompSc., (Bioinformatics)

University of Waterloo, Waterloo, Canada

SUPERVISOR: Dr. Michael G. Surette

NUMBER OF PAGES: 1, 254

I dedicate this work to my parents, Bill Nye, and Jurassic Park, for if it were not for these 3 things, I would have no love of science.

Lay Abstract

The microbes that live on and in us affect our health and can cause disease. Within, I investigate the association of these microbes in the airways. First, we show that the microbes in the noses and throats of the elderly differ from adults. We hypothesize that these differences may be associated with the increased incidence of respiratory infections, such as pneumonia and influenza, in this population. Second, we study the microorganisms that inhabit the lungs of individuals with cystic fibrosis. The bacteria of the lungs are the main cause of disease. In our study, we attempt to identify why these patients go through cycles of extreme sickness and hospitalization, but we were unable to find a cause of this in the microbiota. A follow up study using new techniques provided us with a better resolution of these communities which will help us better understand cystic fibrosis.

Abstract

The human microbiota is the collection of microorganisms which live on and in the human body. These organisms have been implicated in a host of diseases and disorders and nationwide initiatives have helped us understand their heterogeneity across the population in health. In this work, I investigate the respiratory tract microbiota and its correlations in age and disease. Elderly ($\geq$ 65 years of age) are at a greater risk of respiratory infection; previous studies have shown changes to the elderly gut microbiota which correlate with the health of these individuals. Thus, we investigated the upper respiratory tract in comparison to mid-aged adults to identify statistically different communities within the anterior nares and oropharynx which may be associated with increased respiratory infection risk in this population. Individuals with cystic fibrosis have a lung microbiota which contributes to the onset of pulmonary exacerbations, increased inflammation, which is the greatest cause of patient mortality. However, it is not understood what triggers these events. In this work, we used 16S rRNA gene sequencing to longitudinally identify the lung microbiota in a subset of patients but were unable to identify any consistent correlations in the lung microbiota and pulmonary exacerbation onset. In order to gain a better resolution of these communities, we combined culture-independent sequencing technology with culture-enrichment. We showed that 81.21% of OTUs representing 99.15% of the biomass of the cystic fibrosis lung is culturable and that metagenomic sequencing of these cultured communities provide better taxonomic resolution of the cystic fibrosis lung. Together, this work shows the contributions of the respiratory tract microbiota in age and disease.

Acknowledgements

In this age of science, nothing is ever done alone. I have been tremendously lucky to have been surrounded by so many influential, intelligent, creative, caring, and passionate people during my PhD studies. This document is the culmination of the efforts, support, and patience of my colleagues as much as it is of my own research.

I would like to acknowledge my supervisor, Dr. Michael Surette, for much of my academic growth and continued curiosity over the last 5 years. Nothing in this document could have been achieved without Mike's guidance and supervision. Mike's inquisitiveness and intellectual prowess are unequaled. His ability to challenge the scientific status quo has created an environment in which his trainees can ask questions important to both basic and applied science and that are simply fun to answer. Dr. Surette's excellent leadership, encouragement, and passion for science prevails in his willingness to do anything to advance the scientific careers of his trainees. I have been extremely fortunate to have had such an influential role model, and I will keep the lessons Mike has taught me for years to come.

I would also like to acknowledge the important role that my committee has played during my PhD studies. Drs. Brian Coombes and Gerry Wright have continued to challenge me academically and to encourage my career development. Both of my committee members are leading scientific investigators, and their accomplishments have continued to inspire and encourage me.

The Surette laboratory has been a wonderful scientific family to have become apart of. I am incredibly grateful for the positive and inquisitive nature of my colleagues and am so lucky to have been a part of a laboratory who is willing to have as much fun as it is to work hard. I couldn't have gotten to this point without the support of so many of you. Many of the ideas presented within this document have come about via lengthy debates and the challenging of thoughts that is so prevalent within this environment.

Further, being a member of the Department of Biochemistry & Biomedical Sciences, and the Institute for Infectious Disease Research has been a significant contribution to the development of my future career goals and objectives. These scientific families have inspired my scientific progress, helped me overcome imposter's syndrome, and taken great strides to encourage diversity in it's members, making me, as a female scientist, feel part of a much more equal playing field.

I have been amazingly fortunate to have worked alongside the Cystic Fibrosis Canada (CFC). The CFC is an amazing charitable organization whose yearly walkathons and other events and outreach programs give hope to thousands of Canadians living with CF.

I would like to acknowledge my amazing family and friends, including those who I have met during my time at McMaster. Through all the missed get-togethers, or impromptu adventures once "my code had started running", my friends have never once made me feel guilty about living the academic lifestyle. I feel amazingly fortunate to have such patient companions. Thank you, Jennifer Lau, for taking the brunt of my stress-filled rants and breaking them down into deal-able pieces; I am so grateful to have found a best friend in a colleague. Thank you to my wonderful parents who have always been extremely supportive and have never pressured me into a more conventional career path, even if it meant my computer being a common dinner guest at your table. Thank you for teaching me all I needed to accomplish all that I have. Thank you to my brother for supporting me in all that I do... and encouraging me to take a break every now and then. Your inherent questioning of all that is around us is infectious and has contributed greatly to how I approach scientific questioning. And, last but not least, thank you to the incomparable Duane Vaughan. Not only have you contributed amazing insight into my after-hours scientific banter, but you have always provided me with the stability necessary to get through the lowest points of this PhD experience. In everything that we do, you help me "pizza slice" my way through things that I never thought possible, whether it be running long-distance races, cycling up insane hills, or battling my way through the text of a Results section. Every day, I am so amazed by you.

Lastly, thank you to John K. Samson for writing a wonderfully upbeat song about writing an academic thesis.

Contents

List of Figures

List of Tables

Abbreviations and Symbols

AIA	Actinomycetes isolation agar
ANOVA	Analysis of variance
ASL	Airway surface liquid
BAL	Bronchoalveolar lavage
Bcc	*Burkholderia cepacia* complex
Beef	Cooked meat broth with 1.5% agar
BHI	Brain heart infusion agar
BLAST	Basic Local Alignment Search Tool
BSA	Bovine serum albumin
CAMI	Critical Assessment of Metagenomic Interpretation
CAZ	Ceftazidime
CBA	Columbia blood agar
CHOC	Chocolate agar
CF	Cystic fibrosis
CFTR	Cystic fibrosis transmembrane conductance receptor
CFU	Colony forming units
CIPRO	Ciprofloxacin

CNA	Columbia agar with 5% sheep's blood
COPD	Chronic obstructive pulmonary disease
DNA	Deoxyribonucleic acid
eLSA	Extended local similarity analysis
FAA	Fastidious anaerobe agar
FEV1	Forced expiratory volume in 1 second
HOMD	Human Oral Microbiome Database
KVLB	Tryptic soy agar with kanamycin, vancomyin, Vitamin K, hemin, & laked blood
HMP	Human Microbiome Project
IL	Interleukin
IQR	Interquartile range
MAC	MacConkey agar
MOXI	Moxifloxacin
MRSA	Metichillin-resistant *Staphylococcus aureus*
MSA	Mannitol salt agar
NCBI	National Center for Biotechnology Information
NHC	Nursing home cohort
NIH	National Institutes of Health
NTM	Nontuberculous mycobacteria
OFPBL	Oxidation-fermentation polymyxin bacitracin lactose agar
OTU	Operational Taxonomic Unit
LRT	Lower respiratory tract

lytA	autolysin gene
PC	Principal coordinate
PCA	Principal components analysis
PCR	Polymerase chain reaction
PCoA	Principal coordinate analysis
PE	Pulmonary exacerbation
PEA	Phenylethyl alcohol agar with 5% sheep's blood
PERMANOVA	Permutational multivariate analysis of variance
PLCA	Plate coverage algorithm
QIIME	Quantitative Insights Into Microbial Ecology
rRNA	Ribosomal ribonucleic acid
RDP	Ribosomal Database Project
SCV	Small colony variants
SMG	*Streptococcus* Milleri group
sl1p	Surette Laboratory 16S rRNA gene processing pipeline
TIP	Tobramycin inhaled powder
TOB	Tobramycin
T-RFLP	Terminal Restriction Fragment Length Polymorphism
TSY	Tryptic soy agar
UPGMA	Unweighted pair group method with arithmetic mean
URT	Upper respiratory tract
v3	variable 3 region of the 16S rRNA gene
v34	sequence including both variable regions 3 and 4 of the 16S rRNA gene

v35	sequence including both variable regions 3 and 5 of the 16S rRNA gene
v4	variable 4 region of the 16S rRNA gene

Chapter 1

Introduction

1.1 The human microbiota & its associated microbiome

The human microbiota is the plethora of microbes (bacteria, viruses, fungi) that live on or in the human host. While the terms microbiota and microbiome are sometimes used interchangeably to define this community, it is generally accepted that *microbiota* refers to the micro-organisms themselves, whereas *microbiome* refers to the genetic collective of the microbes (Cho and Blaser, 2012; Marchesi and Ravel, 2015). Estimates vary, and are often exaggerated, but the number of cells that make up the human microbiota are on par with a given human's eukaryotic cells, at a ratio of 1:1 (Sender *et al.*, 2016), encouraging some to refer to the human body and its inhabitants as a 'supraorganism' (Caporaso *et al.*, 2011; Turnbaugh *et al.*, 2007).

Because multicellular eukaryotes and bacteria have inhabited the same environmental niches for millions of years, it is likely that each has shaped the evolution of the other and that ancestors of *Homo sapiens'* have been living in symbiosis with a microbiota for millions of years (Ley *et al.*, 2008). Most members of these communities are considered *commensal* organisms; they generally do not pose harm to their human host who often benefits from the interaction (Hugon *et al.*, 2015; Littman and Pamer, 2011). For example, the bacteria that inhabit the human gastrointestinal tract help us absorb nutrients by breaking down complex carbohydrates and fibres (Littman and Pamer, 2011; Jandhyala *et al.*, 2015). In an extreme example, members of the gut microbiota of certain seaweed-consuming populations, such as the Japanese, have obtained the ability to digest carbohydrates from seaweed, allowing the human gut to absorb more nutrients from their ocean diet (Hehemann *et al.*, 2010). Further, studies in germ-free mice (mice lacking a microbiota) and human infants demonstrate the important role that these symbionts play in immune maturation (for e.g. Hapfelmeier *et al.* (2010); Sudo *et al.* (1997); Jakobsson *et al.* (2014) reviewed in Round and Mazmanian (2009); Matamoros *et al.* (2013)).

The initiative that has been the most influential in guiding the community's understanding of the human microbiota has been the Human Microbiome Project (HMP). Initiated in 2008, the HMP aims to better understand the human microbiota inhabiting various sites within and on the average healthy adult. Phase 1 of this project surveyed 18 body sites in 242 individuals, and found heterogeneity within a given individual across various sites in/on the body as well as some heterogeneity between ethnic/racial groups (The Human Microbiome Project Consortium, 2012b).

The HMP laid the ground work for future studies by providing microbiota data for a comprehensive group of healthy adult controls to which comparisons can be made.

Disruption or deregulation (i.e. dysbiosis) of the human microbiota has also been implicated in various diseases and disorders. In these situations, commensal organisms can undertake pathogenic behaviours due to changes in their environment, or interactions with the host or other members of the microbiota (Hugon *et al.*, 2015). Irritable bowel syndrome, inflammatory bowel disease, obesity, and various cancers (e.g. colon and breast) have all been associated with an altered gut microbiota relative to healthy controls (Gilbert *et al.*, 2016; Collins, 2014; Frank *et al.*, 2007; Ley *et al.*, 2006; Vogtmann *et al.*, 2016; Hieken *et al.*, 2016). Further, some evidence has been shown towards a microbial component of disorders along the gut-brain axis including autism spectrum disorder, depression, and Parkinson's disease (Kang *et al.*, 2013; Naseribafrouei *et al.*, 2014; Sampson *et al.*, 2016). The association between the microbiota and these conditions is complex. These diseases and disorders could be associated with a *functional change* in the community, perhaps due to a change (loss or gain) in relative abundance of a set of species which perform a given function, or due to a change in nutrient availability in the environment. In this case, changes to the microbiome may be an important factor in driving disease, whereas individual changes in the microbiota are less important. Conversely, a particular strain(s) may be associated with disease, indicating a *taxonomic change*. The community might be more specifically perturbed by the gain or loss of a particular strain(s), emphasizing the importance of individual members of the microbiota (Gilbert *et al.*, 2016). The *cause and effect* relationship between microbial shifts and disease are also important

to consider in this context. In some cases, there is data to support a causal role; for example, *in vivo* transplant of disease-associated gut microbiota have been able to recapitulate disease phenotypes in naive mice in models of inflammatory bowel disease (Schaubeck *et al.*, 2016; De Palma *et al.*, 2017). However, there are also data showing that changes to the microbiota can be driven by inflammation and/or disease. For example, chemical or genetic materials which induce intestinal inflammation have been shown to cause dysbiosis of the gut microbiota in mice (Lupp *et al.*, 2007).

1.2 Studying the human microbiome

1.2.1 Culture-independent approaches

Human microbiome studies have expanded in both number and scale in the last 15 years. It has long been believed that the human microbiota is an unculturable majority, leaving researchers eager for high-resolution, culture-independent techniques to best study these populations. Recently, advances in next generation sequencing technologies provide new approaches, including 16S rRNA gene sequencing and shotgun metagenomics, to characterize complex microbial communities.

16S ribosomal RNA (rRNA) sequencing

One of the fundamental aims of human microbiome research seeks to uncover which microbes are present in a given polymicrobial community. 16S rRNA gene sequence comparisons have been used to estimate the phylogeny of prokaryotes since the 1970s (Fox *et al.*, 1977, 1980). The DNA sequence of the 16S rRNA gene is used to assign

taxonomy to microbial species because it contains areas of high conservation and variability across the bacterial kingdom, allowing comparisons of both highly diverse and closely related bacterial species (Fox *et al.*, 1977). Improvements in next generation sequencing technologies, most notable of which being Illumina, have provided the ability to sequence regions of this gene in high-throughput. Universal primers have been designed to amplify variable regions of interest from mixed microbial communities by taking advantage of flanking areas of conservation. Some of the first efforts to adopt this approach were applied to 454 sequencing, but were later adopted to Illumina technologies with the design of primers amplifying the variable 3 and 5 regions (Caporaso *et al.*, 2010a). The addition of a short (6-12 base) unique barcode to each primer creates indexed primers such that multiple samples can be uniquely labeled and sequenced using a single Illumina sequencing run (Caporaso *et al.*, 2010a). Many additional primer sets have been designed for a variety of variable regions within the 16S rRNA gene sequence, including indexed primers for the variable 3 (Bartram *et al.*, 2011) and variable 4 (Caporaso *et al.*, 2010a; Walters *et al.*, 2016; Parada *et al.*, 2016; Apprill *et al.*, 2015) regions.

With improvements in sequencing technology, the bottleneck is now in the data analysis. A single multiplexed run of the Illumina MiSeq can produce 15 Gb of information that must then be split into individual samples based on indexed barcodes, and checked for sequencing quality base-by-base. Then, sequences can be grouped into Operational Taxonomic Units (OTUs) based on sequence similarity; OTUs are typically designed as a cluster of sequences which share 97% sequence identity, a threshold previously shown to differentiate bacterial species (Konstantinidis and Tiedje, 2005).

OTUs are then given a specific taxonomic assignment before analyses can be computed. Various tools, workflows, and algorithms have been published to assist at each of these steps. The most widely used is Quantitative Insights Into Microbial Ecology (QIIME, pronounced 'chime'), which consists of a series of commmand-line Python scripts for taking raw sequencing data through to community analyses (Caporaso *et al.*, 2010c). QIIME, and other tools such as mothur (Schloss *et al.*, 2009), use pre-existing algorithms in addition to creating their own. Because of the quantities of data, algorithms for OTU clustering and taxonomic assignment are generally heuristic approaches, meaning that they find approximate solutions since finding exact solutions would be too slow to be reasonably computed (Eddy, 2004). As a result, the use of various heuristic algorithms can significantly alter the outcome of one's analyses depending on how close each heuristic is to the actual data, thereby affecting the biological validity of the results. Unfortunately, there is no consensus as to which workflows, algorithms, and methods should be used; as a result, each research group analyses their 16S rRNA communities differently from each other, making it difficult to compare results across human microbiota studies. As an example of this, Walters *et al.* reanalyzed 5 independent studies of the association of the gut microbiota in obesity using one common computational pipeline and found differences in biological outputs such as α- and β-diversity across studies (Walters *et al.*, 2014).

Shotgun metagenomic sequencing

Metagenomics is used to profile the genes and other genetic elements present in a sampled community, without necessarily gaining the knowledge as to which member of the community contributed any particular element or which are being transcribed

(Gilbert and Dupont, 2011). Instead of asking "who is there" as is the case with marker gene studies such as 16S rRNA gene sequencing, shotgun metagenomics instead asks "what is there", in terms of the functional potential of a microbial community. The term was coined by Handelsman *et al.* in 1998 with the sequencing of DNA from an environmental soil sample using bacterial artificial chromosome (BAC) vector clones, which revealed a large amount of genetic variability within the community (Handelsman *et al.*, 1998). Since 2006, the decreased costs and improvements in sequencing technology have made large-scale shotgun metagenomic sequencing efforts more feasible (Temperton and Giovannoni, 2012). Metagenomics has allowed comparisons of the metabolism of our own genomes to that of the microbiome of our gut (Gill *et al.*, 2006) and has identified functional stability in these communities despite taxonomic variation across individuals (The Human Microbiome Project Consortium, 2012b). It was not until low cost, massively parallel sequencing technology became accessible that these methods were applied to human-associated communities.

Shotgun metagenomic sequencing analysis is computationally intensive. Short reads must undergo quality control before they can be assembled into contigs, given a taxonomic assignment (if possible), be searched for predicted genes, and compared to samples from other locales, patients, or timepoints (Gilbert and Dupont, 2011; Temperton and Giovannoni, 2012; Roumpeka *et al.*, 2017). Because of the extensive use of metagenomics in fields such as environmental biology, there are certain established analysis software tools and methods to help answer common questions such as the taxonomic and gene content of a community (Roumpeka *et al.*, 2017). In particular, efforts have been made to conduct comprehensive benchmarking analyses of these

tools independent of the often biased benchmarking results contained within software announcement manuscripts, making this benchmarking process more reliable than most approaches used in the field (Sczyrba *et al.*, 2017). The Critical Assessment of Metagenomic Interpretation (CAMI) Challenge represents a collaboration between at least 9 independent research groups in the field of metagenomics in order to create a standard assessment of metagenomic processing and analysis tools (Sczyrba *et al.*, 2017). The CAMI Challenge includes sequence assembly approaches, binning strategies, and taxonomic assignment, all important elements in the processing of shotgun metagenomic data (Sczyrba *et al.*, 2017). This initiative has allowed the field to focus its efforts on producing biologically accurate results using the best available tools and software.

1.2.2 Culture-dependent approaches

Before culture-independent methods were popularized, studies of the bacterial communities associated with the human host relied on microbial culture. Traditionally, only a very small percentage of the human microbiota was believed to be culturable (Rappé and Giovannoni, 2003; Stewart, 2012). This belief is repeated in the first sentence of a large portion of human microbiome related manuscripts published in the last 10 years, a sentiment that has been parroted from study to study. However, evidence exists in the literature of successful culture-dependent studies; in 1974, 25 years before the term 'microbiome' was coined, Finegold *et al.* compared the fecal microbiota of individuals from two nationalities to test the hypothesis that differences in their microbial communities may contribute to different rates of colon and bowel cancers between these groups (Finegold *et al.*, 1974). Using 10 aerobic and 19

anaerobic media, the authors recovered over 300 unique species from 40 specimens (Finegold *et al.*, 1974). Studies such as these indicate the readiness of the gut microbiota to culture. Since then, studies have begun to combine culture-dependent and -independent methods (see Section 1.2.3).

Although the advantages of culture-independent methods are often thought to outweigh those of culture, culture has a number of important advantages. Unlike DNA-based methods which cannot differentiate between live or dead organisms, culture establishes the viable members of a community. Selective media also allow for the growth of low abundant organisms often missed by insufficient 16S rRNA gene sequencing depth (Lau *et al.*, 2016; Sibley *et al.*, 2011). Finally, in order for the field of microbiome research to move beyond descriptive studies, the isolation of organisms is necessary to determine the role they play in health and disease; with culture isolates, a variety of phenotypic assays can be performed and whole genome analysis can provide imperative functional annotation of these organisms.

1.2.3 Combination of culture-dependent & -independent methods

Recent studies, including efforts from the Surette laboratory, are beginning to challenge the anti-culture antics of the field of human microbiome research by showing that a large portion of the human microbiota is culturable. In 2011, Sibley *et al.* demonstrated that culture-enrichment of the cystic fibrosis lung microbiota allows for the identification of more species of bacteria when compared to culture-independent

approaches (Sibley *et al.*, 2011). Specifically, the authors identified a 3-fold increase in organism diversity by applying 21 culture conditions, 10 aerobic and 11 anaerobic, to lung sputum samples compared to Terminal Restriction Fragment Length Polymorphism (T-RFLP) identification and 454 sequencing of the 16S rRNA gene of the sample directly (Sibley *et al.*, 2011). More recently, Lau *et al.* applied similar techniques to the human fecal microbiota and found that an average of 95% of the gut microbiota present at $> 0.1\%$ is culturable using 33 standard culture conditions in aerobic and anaerobic environments (Lau *et al.*, 2016). Other studies have also identified culturable communities in the gut microbiome (Goodman *et al.*, 2011; Lagier *et al.*, 2012; Rettedal *et al.*, 2014; Lagier *et al.*, 2016; Browne *et al.*, 2016) using culture-enriched techniques. Other communities across the human body are also culturable including the vagina (Pandya *et al.*, 2016), oral cavity (Thompson *et al.*, 2015), urinary tract (Hilt *et al.*, 2014), and healthy airways (Venkataraman *et al.*, 2015).

The understanding that the majority of the human microbiota is unculturable originated from the increase in diversity that early culture-independent studies observed (for e.g. Eckburg *et al.* (2005)). Recent studies combining culture-independent and -dependent approaches produce 2 overlapping but distinct microbial communities, suggesting that these approaches complement each other, but that neither encompasses the full extent of microbial diversity (Lagier *et al.*, 2012; Sibley *et al.*, 2011; Lau *et al.*, 2016). Interestingly, there is often more diversity observed via culture-enrichment then by direct profiling of a given sample (Lau *et al.*, 2016; Sibley *et al.*, 2011; Browne *et al.*, 2016), perhaps indicating the ability of culture-enrichment as a

means for facilitating the observance of low abundant organisms, thus avoiding issues associated with sequencing depth (Lau *et al.*, 2016).

1.3 The respiratory tract microbiota

Before we can understand the possible implications of the human microbiota in respiratory disease, we first have to understand its role and composition in health. The human respiratory tract consists of the airways between the nasal cavity and lungs; conceptually, this expanse is often split into the upper and lower respiratory tracts (URT and LRT, respectively) just below the larynx. Studies of the healthy URT microbiota from the HMP, and other subsequent studies, indicate the existence of distinct microbial communities in the nasal cavity (anterior nares and nasopharynx), throat (oropharynx), and mouth (The Human Microbiome Project Consortium, 2012b; Stearns *et al.*, 2015; Bassis *et al.*, 2014; Charlson *et al.*, 2011).

Until recently, the LRT was perceived to be a sterile environment owing to studies from the 1960's which where unable to culture LRT microbial isolates (Laurenzi *et al.*, 1961). However, improved culture-independent techniques have continuously found evidence of a LRT community in healthy individuals. Comparisons between oral washes and bronchoalveolar lavage (BAL) using the neutral model of community ecology identified a number of species in the lung microbiota which were not present in the mouth, indicating the lung microbiota as a distinct community (Morris *et al.*, 2013). Further, the lung microbiota are altered in individuals with HIV compared to healthy controls (Lozupone *et al.*, 2013) and is distinct in individuals at increased risk of pulmonary inflammation (Segal *et al.*, 2013). Charlson *et al.* used swabs and

oral washes of the URT to compare to BAL and protected brush samples of the LRT (Charlson *et al.*, 2011); the authors identified bacterial DNA in the LRT samples at a lower biomass than URT microbiota, but found these communities to be quite similar to the oropharyngeal microbiota, suggesting that the LRT microbiota is derived from its URT neighbours (Charlson *et al.*, 2011). Sampling the LRT of a healthy individual means first maneuvering an instrument through the URT which has caused a debate in the field as to what constitutes contamination and what is a true LRT microbiota (Beck *et al.*, 2012). Dickson *et al.* conducted BALs via the nasal cavity or oropharynx and found no significant differences between the obtained communities, suggesting that contamination is not a driving factor in the communities recovered from BALs (Dickson *et al.*, 2015). However, both of these methods must traverse the supraglottic space just above the trachea; Segal *et al.* compared the supraglottic community to BALs and found a significant overlap in taxa, suggesting that contamination during BAL procedure is still at issue (Segal *et al.*, 2013). In a more recent follow-up study, Dickson *et al.* tested two hypotheses as to the source of LRT communities: (a) these communities originate via dispersion along the mucosal layer of the bronchia, meaning that the microbial community would decrease in similarity to the URT as one moved further down the respiratory tract (Dickson *et al.*, 2017), and (b) LRT microbiota could be a result of microaspiration of microbes from the URT. In the latter case, one would expect the bottom of the bronchia to be most similar to the URT composition due to the effect of gravity and the general habit of humans to sit/stand upright (Dickson *et al.*, 2017). Via serial sampling, the authors showed that the latter hypothesis fit best with the data in a set of healthy adult volunteers (Dickson *et al.*, 2017). These studies, and others before them, have allowed us to confidently

assess the existence of a biomass LRT microbiota in healthy adult individuals whose composition is a consequence of one's URT community.

1.4 Respiratory infections in the elderly

Infants and young children (<2 years of age) are at a higher risk of respiratory infections (Simoes *et al.*, 2006; Williams *et al.*, 2002). This increased risk is known to have a microbial component since early life events, including method of delivery (vaginal/caesarean) and breastfeeding, affect the colonization and succession of respiratory microbiota, and respiratory health throughout life (Schenck *et al.*, 2016; van Nimwegen *et al.*, 2011; Arrieta *et al.*, 2015; Dogaru *et al.*, 2014). This increased risk in respiratory infection diminishes as the respiratory microbiota becomes more adult-like (Bogaert *et al.*, 2011; Stearns *et al.*, 2015). However, as individuals age (≥65 years old), this increased risk of respiratory infection again surfaces. Infectious diseases, such as influenza and pneumonia, increase in prevalence among individuals over the age of 65 years (Kaplan *et al.*, 2002; Crighton *et al.*, 2007; Centers for Disease Control and Prevention (CDC), 1995). Interestingly, although occurrence of pneumonia caused by *Streptococcus pneumoniae* is increased in this age group, carriage of this species in the URT is decreased. For example, in some studies 50% of surveyed children are found to be *S. pneumoniae* carriers compared to <1% of elderly individuals (Kwambana *et al.*, 2011; Ridda *et al.*, 2010). This suggests that while the elderly are less likely to be nasally colonized by this microbe, exposure to it often leads to invasive LRT infection. An interesting phenomena in support of this is the spike in elderly respiratory pneumococcal disease observed around the Christmas holidays, presumably when elderly come into contact with grandchildren and other

young children with high rates of *S. pneumoniae* nasal carriage (Walter *et al.*, 2009).

Although causes of this increased risk of respiratory infection with age are not fully understood, it has been suggested that immunosenescence plays a role. *Immunosenescence* is the deregulation of both the innate and adaptive immune response which often accompanies aging (Franceschi *et al.*, 2000; Krone *et al.*, 2014). Immunosenescence in elderly mice leads to a decrease in anti-pneumococcal antibodies (Nicoletti *et al.*, 1993); this, along with the general deterioration of the immune system, may partially explain the increased ability for pathogenic bacteria, such as those responsible for pneumococcal disease, to evade the immune defenses of the human host (Krone *et al.*, 2014).

1.4.1 Changes in the respiratory tract microbiota with age

While it is well established that the gastrointestinal microbiota is altered with age (Jeffery *et al.*, 2015; Lynch *et al.*, 2015; Zapata and Quagliarello, 2015; Park *et al.*, 2015), less is known about how the respiratory tract is affected. Hints in how this community may change as we age can be seen from studies in related fields. For example, mouse studies comparing the URT microbiota in adult and elderly colonies have shown a marked change in these communities with age (Thevaranjan *et al.*, 2016). As humans age, immunosenescence and inflammaging lead to a pro-inflammatory deregulation of the immune system, often associated with an increase in comorbidities and infectious disease (Johnstone *et al.*, 2014; Krone *et al.*, 2014; Franceschi *et al.*, 2000). Because of links between the immune system and microbiota established in human infants and mouse models, a corresponding change in the elderly URT microbiota

may occur. Some evidence towards this change is observed by de Steenhuijsen Piters *et al.* whose research established changes in relative abundance of a subset of species (*Prevotella, Veillonella, Leptotrichia, Rothia*, and *Lactobacillus*) between the healthy adult and healthy elderly oropharyngeal microbiome (de Steenhuijsen Piters *et al.*, 2015).

1.5 Chronic infection in cystic fibrosis

Cystic fibrosis (CF) is an autosomal recessive disorder characterized by abnormalities in chloride transport and absorption by the cell (Andersen, 1938; Quinton, 1989), a defect that affects a variety of organs, including the pancreas, liver, and intestine (Andersen, 1938; Riordan *et al.*, 1989). In the 1930s, when CF was first identified, persons with this disease often died of malnutrition during infancy (Andersen, 1938; O'Sullivan and Freedman, 2009; Cutting, 2014) due to the inability of the pancreas to properly move digestive enzymes from the pancreatic ducts into the intestine because of mucus buildup (Cutting, 2014). Once pancreatic enzyme replacement therapy was introduced, the mean survival rate of individuals with CF increased to early adulthood and the consequences of this disorder on the lung began to supersede malnutrition as the major cause of morbidity and mortality in the patient population (Heijerman, 2005).

The hunt to identify the genetic locus responsible for this disease was outlined in a series of landmark papers published in 1989, including the sequencing and characterization of a gene on chromosome 7 encoding the cystic fibrosis transmembrane conductance regulator (CFTR), the gene mutated in CF patients (Kerem *et al.*, 1989;

Riordan *et al.*, 1989; Rommens *et al.*, 1989). While 86.4% of CF disease is caused by a 3 bp deletion that encodes for amino acid 508, encoding a phenylalanine (F) in CFTR (ΔF508) (Kerem *et al.*, 1989; Boyle *et al.*, 2014; Cystic Fibrosis Foundation, 2013), there are over 2,000 mutations to CFTR currently identified as causative of CF disease (Elborn, 2016; Castellani *et al.*, 2008). Subsequently, it has been elucidated that various mutations within this gene can contribute to the mis-folding and rapid degradation of CFTR and/or a lack of functional CFTR on the cell outer membrane (Kerem *et al.*, 1989; Elborn, 2016). There are 6 classes of genetic mutations which affect CFTR; some of these mutations mitigate the amount of functional CFTR on the cell surface by (Class 1) decreasing CFTR production, (Class 2) causing defects in protein trafficking, leading to CFTR degradation in the endoplasmic reticulum, and (Class 3) causing defective protein regulation. The last 3 classes of mutations affect the functionality of the protein via (Class 4) a reduced ability to transport chloride through the CFTR channel, (Class 5) splicing defects that reduce the amount of functional CFTR produced by the cell, or (Class 6) a decreased stability of the protein on the cell membrane (O'Sullivan and Freedman, 2009; Elborn, 2016). Genetic complementation of the mutant CFTR gene with the wild-type version corrects the defects in chloride channel function typical of CF (Rich *et al.*, 1990), suggesting that CFTR is critical for to the CF disease phenotype.

While mutations in CFTR are recognized as the necessary genetic factor which precludes CF, there are a variety of *modifier genes* which have been recognized to play a role in the severity of the CF phenotype. For example, a wide variation in disease phenotypes are seen among those homozygous for the ΔF508 mutation (Collaco and

Cutting, 2008; Kerem *et al.*, 1990b) as well as within twin studies of CF disease (Collaco and Cutting, 2008). Evidence for the ability of these genes to modify CF disease were either hinted at from studies of related diseases, were part of known functional pathways linked to CF, or were identified via linkage studies CFTR (Collaco and Cutting, 2008). There have been more than 30 such modifier genes identified, but the search for modifier genes is limited by the number of available samples and small patient populations (Drumm *et al.*, 2005; Collaco and Cutting, 2008).

With CFTR characterized and sequenced, the majority of CF research moved towards identifying a cure for the disease. Attempts have been made to use gene therapeutics to deliver a functional copy of CFTR to mucosal sites within the body; although success has been seen *in vitro*, this therapy has had little success *in vivo* (Moss *et al.*, 2007; Pickles, 2004; O'Sullivan and Freedman, 2009). More recently, a series of molecules have been identified that improve the functionality of the mutant CFTR protein. *Correctors* are CFTR modulating proteins which act as molecular chaperones, helping to increase the number of CFTR proteins on the cell membrane whereas *potentiators* increase the function of membrane-bound CFTR (O'Sullivan and Freedman, 2009). Potentiators such as ivacaftor have shown success in phase 2 and 3 clinical trials for individuals with class 3 and 4 CFTR mutations (particularly the G551D mutation of CFTR) and are approved therapies for patients with specific mutations (Ramsey *et al.*, 2011; Accurso *et al.*, 2010). Combining this potentiator with the corrector lumacaftor has seen some success in the more prevalent class 1 and 2 mutations, particularly in the predominant ΔF508 mutation (Boyle *et al.*, 2014; Wainwright *et al.*, 2015; Cystic Fibrosis Foundation, 2013).

Interestingly, exactly how alterations in CFTR cause the symptomatology of CF has not been fully identified, though there are a number of leading hypotheses. It is known that mutations in CFTR affect mucosal membranes; specifically in the airways, a decrease in functional CFTR alters the composition of the airway surface liquid (ASL) (Clunes and Boucher, 2007). The *high-salt hypothesis* states that an excess of sodium (Na) and chloride (Cl) in ASL caused by defective Cl transport by CFTR disrupts the innate immune system (Clunes and Boucher, 2007; Smith *et al.*, 1996). Some elements of the innate immune system, such as human beta-defensin 1, are highly sensitive to salt (Goldman *et al.*, 1997); an increase in Na and Cl in the ASL might leave the epithelia vulnerable to infection (Clunes and Boucher, 2007). Conversely, the *low volume hypothesis* describes a defective CFTR that disrupts the osmotic pressure at the epithelial interface, leading to a dehydrated airway surface, impairing the ability of the epithelial cilia to clear the ASL via mucocilliary clearance (Clunes and Boucher, 2007; Matsui *et al.*, 1998; Button *et al.*, 2012). Other hypotheses suggest an inherent deregulation of the host inflammatory response (O'Sullivan and Freedman, 2009; Tirouvanziam *et al.*, 2000), or a disposition to infection due to the ability of wild-type CFTR to bind microbes such as *Pseudomonas aeruginosa* (O'Sullivan and Freedman, 2009; Campodónico *et al.*, 2008). It is possible that the phenotype observed in individuals with CF is a combination of aspects from all of these hypotheses.

What is known, however, is that the effects of dysfunctional CFTR are widespread throughout the body. Although the respiratory tract is responsible for the majority of

mortality in CF, the effects on the gastrointestinal tract are widespread in the patient population, typically manifesting as decreased digestive function and an accumulation of mucus in the intestine which can lead to ileum obstruction (De Lisle and Borowitz, 2013). Inflammation along the gastrointestinal tract is also present in the majority of CF patients, including increased levels of fecal calprotectin (Werlin *et al.*, 2010; De Lisle and Borowitz, 2013) and other inflammatory markers in the lumen (Smyth *et al.*, 2000; De Lisle and Borowitz, 2013). Additionally, incidence of Crohn's disease, an inflammatory bowel disease, is increased in CF (De Lisle and Borowitz, 2013). Perhaps related to these phenotypes, the gut microbiota in CF differs from that of age-matched non-CF controls (Nielsen *et al.*, 2016; Miragoli *et al.*, 2016; Manor *et al.*, 2016). Mutations in CFTR are associated with increased incidence of chronic pancreatitis (Cohn *et al.*, 1998) and pancreatic insufficiency (Shwachman and Kulczycki, 1958), often leading to CF-related diabetes mellitus. Cirrhosis of the liver affects 5-10% of CF patients, though the incidence of liver disease is more common (approximately 30%) (Flass *et al.*, 2015; Kobelska-Dubiel *et al.*, 2014). Dysfunctional CFTR also affects the vas deferens, bone density, and, of course, the airways (Elborn, 2016).

1.5.1 The lower airway environment

While improvements in management of CF has led to an improved predicted median survival age of 50.9 years of age (Cystic Fibrosis Canada, 2013), in the majority of CF patients, morbidity and mortality is due to complications related to the airways. In healthy individuals, the lower airway microbiota is a low-biomass collection of microbes which are generally the result of microaspiration of organisms present in the URT (see Section 1.3); however, defects in the function of CFTR in the airways of

individuals with CF create a lung environment ideal for bacterial growth (O'Sullivan and Freedman, 2009). For example, increased ASL viscosity prevents the action of the cilia on the epithelial cell surface from clearing aspirated organisms. Further, the pH of the ASL is reduced in cells without a functional CFTR (Berkebile *et al.*, 2014). Because of this, and perhaps an inherent deregulation/dysfunction of the immune system, aspirated bacteria persist in the LRT, leading to chronically colonizing lung microbiota which can reach a density of 10^5-10^{10} colony forming units (CFUs) per millilitre of sputum (Meyer *et al.*, 1997; Stressmann *et al.*, 2011b). In conjunction with this atypical lung microbiota are chronic low levels of host inflammation, including the recruitment of neutrophils (Armstrong *et al.*, 1997; Khan *et al.*, 1995); the presence of neutrophil elastase has been observed early in the lungs of CF infants (Cantin *et al.*, 2015), and this inflammatory response may even preceed microbial colonization (Heijerman, 2005). Further, increased levels of the pro-inflammatory cytokines such as interleukin (IL)-8 and IL-17 have been measured in the CF LRT (Cantin *et al.*, 2015; Armstrong *et al.*, 1997; Khan *et al.*, 1995). These inflammatory immune responses, including chronic neutrophil recruitment, can cause harm to the host epithelia via release of proteases and oxidants resulting in tissue damage, remodelling, and bronchiectasis (Cantin *et al.*, 2015; Sly *et al.*, 2013).

While this airway environment is atypical, a general level of patient stability can be maintained via an intense treatment regimen. In order to maintain a healthy steady state, patients cycle through an assemblage of antibiotics and airway clearance therapies. While there is no standard CF therapy, as treatment varies greatly by clinician preference and the response of patients to such drugs, most commonly

used antibiotics target known primary CF pathogens (Flume *et al.* (2009); Sibley *et al.* (2009); for more detail on primary pathogens see Section 1.5.2). Tobramycin is perhaps the most common anti-pseudomonal antibiotic used during maintenance periods (Ramsey *et al.*, 1999; Kopp *et al.*, 2015), often in the form of tobramycin inhaled powder (TIP) (Geller *et al.*, 2007). Other common anti-pseudomonal drugs include the β-lactam antibiotics ceftazidime, aztreonam, and amoxicillin, most of which tend to be prescribed during pulmonary exacerbation (see below) (Flume *et al.*, 2009). Additionally, various aminoglycosides, quinolones and β-lactams are used to treat exacerbations thought to be caused by *Staphylococcus aureus* and *Burkholderia cepacia* complex (Conway *et al.*, 2003; Szaff and Hoiby, 1982). In addition to antibiotics chosen for their specific bactericidal activity, azithromycin is also commonly prescribed as a maintenance antibiotic (antibiotics prescribed outside of a pulmonary exacerbation) even though its mode of action is not well understood (Saiman *et al.*, 2010; Sibley *et al.*, 2010a).

Along with antibiotics, other aspects of clinical treatment are important for the care of individuals with CF. Recombinant human DNase is often prescribed to help break down high molecular weight DNA thought to contribute to the viscosity of airway sputum (Fuchs *et al.*, 1994). Hypertonic saline is inhaled in order to improve mucocilliary clearance in the LRT and decrease the prevalence of pulmonary exacerbations (Reeves *et al.*, 2015). Further, patients are encouraged to perform various airway clearance exercises including active cycle breathing techniques, positive expiratory pressure (often via use of a resistor) and chest physical therapy. These techniques are designed to help physically remove mucus buildup from the airways

(Szego and Canadian Physiotherapy Group, 2017).

Despite these advancements in care, CF patients experience *pulmonary exacerbations* (PEs) which are defined by increased respiratory distress (Goss and Burns, 2007; Fuchs *et al.*, 1994; Ramsey *et al.*, 1999). The precise definition of PEs differ between care centres and studies but is generally defined based on symptoms such as fever, increased sputum production, dyspnea, a decrease in pulmonary function, change in chest physical examination, increased sinus pain, and/or increased blood neutrophil counts (Goss and Burns, 2007; Fuchs *et al.*, 1994; Ramsey *et al.*, 1999). While the cause of PE onset is not well defined - and has been linked to environmental effects (Goss *et al.*, 2004), viral acquisition (Hiatt *et al.*, 1999; Goss and Burns, 2007), and the strength of the immune response (Goss and Burns, 2007) - the lung microbiota is known to play a large contributing factor (Carmody *et al.*, 2013). In a percentage of PEs, a decrease in a *primary CF pathogen* is observed at PE resolution; however, this decrease does not predict improved clinical responses to PE treatment (Lam *et al.*, 2015). Instead, in many patients, a quantitative change in levels of a primary CF pathogen is not observed (Rabin and Surette, 2012). Despite this, PE symptoms typically decrease upon antibiotic therapy, implicating a bacterial role in these events.

Understanding the underlying cause of PE onset is critical to patient care. During PEs, lung function, often measured in units of forced expiratory volume in 1 second (FEV1), drops with the progression of symptoms. Following PE treatment and resolution, the patient's FEV1 often increases but rarely does lung function return to the baseline pre-PE levels (Sanders *et al.*, 2010). Thus, throughout a patient's lifetime,

cycles of relative stability and PEs progressively decrease lung function (Sibley *et al.*, 2009), ultimately causing mortality in the majority of patients (Heijerman, 2005; Lyczak *et al.*, 2002). Typically, antibiotic therapy is directed towards the dominant principle pathogen during PE treatment; however, emerging data on the complexity of the CF lung microbiota (see Section 1.5.3), and the possibility of polymicrobial infection, should alter the treatment of these events (Sibley *et al.*, 2009). Without fully understanding why/how PEs occur, it can be difficult to treat the clinical symptoms of a PE with the appropriate clinical therapies in a time-effective manner, lengthening patient discomfort and contributing to decreasing lung function. Because of this need, and the recognition that PEs are not only caused by primary CF pathogens, the importance of analyzing the totality of the LRT microbiota in CF patients is increasingly being recognized.

1.5.2 Principle pathogens in the cystic fibrosis lung

Of the bacterial species that make up the approximately 10^5 - 10^{10} CFUs/mL of sputum in the CF lung (Meyer *et al.*, 1997; Stressmann *et al.*, 2011b), there are a select number - detailed below - which have been associated with PEs and respiratory infection in CF and are the organisms targeted by standard CF clinical microbiology protocols (**Fig 1.1**) (Surette, 2014; LiPuma, 2010; Parkins and Floto, 2015).

Pseudomonas aeruginosa

Of the primary CF pathogens, the most notable is *Pseudomonas aeruginosa*. *P. aeruginosa* is present in 50%-80% of adult CF patients (Lyczak *et al.*, 2002; LiPuma,

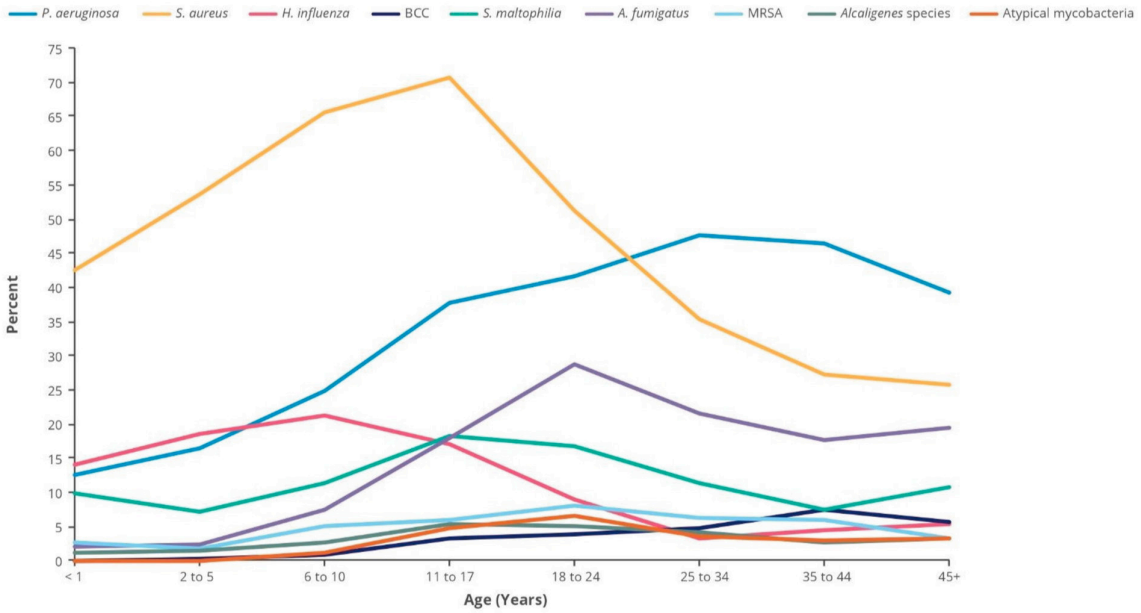


Figure 1.1: **Age-specific prevalence of respiratory infections in CF individuals, 2014.** This image, originally presented in The Canadian Cystic Fibrosis Registry's 2014 Annual Report, outlines the incidence of principle lung pathogens in CF. (This image has been reproduced with permission from CFC) (Cystic Fibrosis Canada, 2016).

2010) and 37% of all Canadians with CF (Cystic Fibrosis Canada, 2016). Initial colonization of this organism generally occurs within the first 3 years of life (Burns *et al.*, 2001) by an environmental strain of the bacterium (Lyczak *et al.*, 2002). These early stages of infection can be intermittent and involve multiple strains, but the subsequent chronic infection, when *P.aeruginosa* is not eradicated with antibiotic therapy, involves a single dominating strain (LiPuma, 2010). In most cases, mucoid variants arise which are thought to better protect the bacteria from dehydration in this unique environment (Berry *et al.*, 1989; Fegan *et al.*, 1990; Li *et al.*, 2005). Patients harbouring mucoid *P. aeruginosa* often have a poorer prognosis then those not colonized (Li *et al.*, 2005). Another adaptation that *P. aeruginosa* makes in the CF lung is the formation of small colony variants (SCVs). SCVs are small, often slower-growing isolates which differ in phenotypic and/or pathogenic diversity compared to the dominant colonizing strain (Proctor *et al.*, 2006). SCVs of *P. aeruginosa* are often antibiotic resistant and correlate with worse lung function (Malone, 2015). Further, the colonizing strain of *P. aeruginosa* becomes regionally isolated in the lung over time, allowing for increased genetic diversify and diversification of phenotypic traits (Jorth *et al.*, 2015). Although this dominant clone of *P. aeruginosa* typically colonizes CF patients long-term without super-infection of other environmental strains, in some cases epidemic strains of *P. aeruginosa* can supersede (McCallum *et al.*, 2001). The Liverpool Epidemic Strain, Prairie Epidemic Strain, and Australian epidemic strains have increased antibiotic resistance and the ability to spread patient-to-patient, making them a substantial threat to CF patient health (Cheng *et al.*, 1996; Parkins *et al.*, 2014; McCallum *et al.*, 2001; Duong *et al.*, 2015). These adaptations of *P. aeruginosa* are some examples of the evolutionary mechanisms that this organism uses within the

CF lung. Colonization with *P. aeruginosa* decreases patient prognosis and increases the onset of lung disease (Kerem *et al.*, 1990a; Kosorok *et al.*, 2001; LiPuma, 2010).

Staphylococcus aureus

Staphylococcus aureus can be cultured from approximately 48% of Canadian CF patients (Cystic Fibrosis Canada, 2013). It is thought that non-pathogenic colonization of *S. aureus* in the URT lays the foundation for LRT infections in these individuals (Ulrich *et al.*, 1998). Colonization with *S. aureus* is most common in children and adolescents but can persist into adulthood (LiPuma, 2010). Colonization, especially in tandem with other primary pathogens such as *P. aeruginosa*, leads to a poorer prognosis for CF patients (Limoli *et al.*, 2016). Methicillin-resistant *S. aureus* (MRSA) has increased in prevalence in the lungs of American CF patients; however, incidence in Canada, Europe, and Australia has remained low (**Fig 1.1**) (Parkins and Floto, 2015).

Burkholderia cepacia complex

Burkholderia cepacia complex (Bcc) currently consists of 17 species of closely related bacteria (LiPuma, 2010). While the genus *Burkholderia* contains many environmentally associated species, only a fraction have been implicated in human disease (LiPuma, 2010). The prevalence of Bcc is lower than that of other primary pathogens, infecting only 8% of adults with CF (LiPuma, 2010; Razvi *et al.*, 2009). However, infection with Bcc can have detrimental effects, including increased risk of death following lung transplant (Murray *et al.*, 2008), and the prevalence of "cepacia

syndrome" which includes a rapid decrease in lung function, pneumonia, and bacteraemia (Isles *et al.*, 1984; LiPuma, 2010). Bcc is also correlated with decreased lung function and episodic exacerbation (Govan and Deretic, 1996). Similar to those of *P. aeruginosa*, several epidemic strains of Bcc have been identified that are capable of interpatient transmission (LiPuma, 2010).

Haemophilus influenzae

Haemophilus influenzae commonly colonizes patients with CF early in childhood (**Fig 1.1**) (Bilton *et al.*, 1995; Rosenfeld *et al.*, 2001; Cystic Fibrosis Canada, 2016), and is prevalent in 10.7% of the Canadian CF population (Cystic Fibrosis Canada, 2016). This organism is a common member of the URT where it has been correlated with infections such as otitis media and community-acquired pneumonia (Sánchez *et al.*, 1999; Leibovitz *et al.*, 2003). *H. influenzae* has been implicated in biofilm formation on the surface of airway epithelia, increasing inflammatory responses and ultimately contributing to CF lung disease (Starner *et al.*, 2006).

Streptococcus Milleri/Anginosus group

Difficulties in growing and identifying these organisms have kept them below the level of detection for years; however, the development of *Streptococcus* Milleri/Anginosus group (SMG)-specific media has allowed for the elucidation of the clinical significance of this group of organisms (Bittar and Rolain, 2010; Parkins *et al.*, 2008; Sibley *et al.*, 2008, 2010b). The discovery of SMG's pathogenic role in CF exacerbations is owed in part to the realization that anti-SMG antibiotics, such as ceftriaxone and clindamycin, decreased numbers of SMG in patient sputum samples while simultaneously subsiding

symptoms of PE (Parkins *et al.*, 2008; Sibley *et al.*, 2008). It has been suggested that the SMG can act in synergy with anaerobic organisms such as *Prevotella spp.*, another common colonizer of the CF lung, to cause pulmonary infection and abscesses (Shinzato and Saito, 1994; Mendonca, 2017). Although the prevalence of SMG is not widely reported, these organisms have been identified in up to 58% of patients in some studies (Zhao *et al.*, 2012; Surette, 2014; Parkins *et al.*, 2008).

Other Principle Pathogens

Mycobacterium spp., including nontuberculous mycobacteria (NTM) is prevalent in between 2-28% of individuals (LiPuma, 2010) and are difficult to manage, given the limited clinical insight into these microbes (Parkins and Floto, 2015). *Achromobacter spp.* and *Stenotrophomonas maltophilia* are emerging CF pathogens; an increased prevalence of these species may be due to resistance to increased antibiotic usage, or simply from better culture and molecular surveillance (Parkins and Floto, 2015; Rogers *et al.*, 2003). *Alcaligenes spp.*, often associated with infections of immune compromised patients, have also been identified in the CF lung microbiota, though incidence in Canada remains low (**Fig 1.1**) (Cystic Fibrosis Canada, 2016; Tan *et al.*, 2002). Additionally, multiple research groups have demonstrated a surprisingly high number of anaerobic bacteria present in this environment (Tunney *et al.*, 2008; Sibley *et al.*, 2011), suggesting their possible importance in CF exacerbations. Fungal species, such as *Aspergillus fumigatus* also contribute to a decline in lung function and have been associated with PE (Speirs *et al.*, 2012).

1.5.3 The cystic fibrosis lung microbiota

Although it is well established that the colonization of certain bacterial species in the CF lung are associated with a poorer prognosis, the study of the CF lung environment has steadily evolved from microbiological investigations of solitary bacterial species to the study of the CF microbiota as an interconnected, ecological community (Conrad *et al.*, 2013; LiPuma, 2010). In 2003, Rogers *et al.* used culture-independent techniques to investigate CF lung microbial communities directly from BAL and sputum samples (Rogers *et al.*, 2003). Since then, molecular methods, primarily 16S rRNA gene sequencing, have been used to investigate the CF lung microbiota. The LRT of CF patients is not simply home to a handful of pathogenic organisms (Tunney *et al.*, 2008; Harrison, 2007); instead, this polymicrobial environment harbors both traditional CF pathogens and organisms generally considered to be harmless commensals.

A series of studies have outlined the CF lung microbiota across patient populations and disease states. Although a small number of primary CF pathogens are prevalent across the CF population (Fodor *et al.*, 2012), the totality of the CF lung microbiota is highly variable across individuals (Coburn *et al.*, 2015; Stressmann *et al.*, 2011a) and shown to be stable within a given individual over time (Fodor *et al.*, 2012). It has been suggested that the diversity of the CF lung microbiota declines with age and progressive lung disease (Coburn *et al.*, 2015; Bacci *et al.*, 2016). While decline in lung function can be seen in patients with progressive disease, stable patients often have similar diversity scores to the end-stage of progressive disease even through their FEV1 and other symptom scores remain high (Zhao *et al.*, 2012). Generally, a small number of OTUs whose taxonomic assignment correlate with primary CF pathogens

dominate these communities (Zhao *et al.*, 2012; Coburn *et al.*, 2015).

Interestingly, given the plethora of CF lung microbiota studies, very few correlations between these communities and the onset of PEs have been established. Zhao *et al.* did not observe any changes in lung microbial communities between baseline, PE onset, PE treatment, and recovering timepoints (Zhao *et al.*, 2012). Longitudinal studies have similarly not found statistically significant differences between PE, patient stability, and other timepoints (Cuthbertson *et al.*, 2015; Carmody *et al.*, 2015). Interestingly, Carmody *et al.* was able to establish a correlation between the genus *Gemella* and PE onset (Carmody *et al.*, 2013); however, this result has yet to be validated in other studies.

There are a number of drawbacks to using culture-independent methods to study the CF lung microbiota. Arguably the most important is the inability to differentiate between viable and non-viable cells (Surette, 2014; Whelan and Surette, 2015). While a small number of studies have used RNA sequence analysis of the 16S rRNA gene in order to capture the active subset of the CF lung community (Rogers *et al.*, 2005; Quinn *et al.*, 2014), these studies have been limited by the ability to isolate quality RNA.

A limited number of studies have moved beyond 16S rRNA gene sequencing into studies of the CF lung metagenome and metatranscriptome. Metagenomic approaches

have been used in a limited capacity as a means of demonstrating that these techniques produce similar taxonomic profiles to that observed in 16S rRNA gene sequencing studies (Hauser *et al.*, 2014). Further, metagenomics has been used to show that it is possible to identify fungal and viral components with these methods (Lim *et al.*, 2014), and that the potential metabolic functionality of the microbiota of different patients remains consistent even when their bacterial composition may not (Lim *et al.*, 2013). In the earliest study, Lim *et al.* identified antibiotic resistance mechanisms as well as potential metabolic activities of microbes within the CF lung (Lim *et al.*, 2013). Expanding these studies, the CF lung has been shown to be enriched for amino acid catabolism, folate biosynthesis, lipoic acid biosynthesis, and the fermentation product 2,3-butanedione in the CF LRT (Quinn *et al.*, 2014; Whiteson *et al.*, 2014). This technique has also been used to identify DNA viruses, including bacteriophages, (Willner *et al.*, 2009; Moran Losada *et al.*, 2016) along with dominating *P. aeruginosa* or *S. aureus* strains (Feigelman *et al.*, 2017). To-date, metagenomic studies have largely reiterated previous findings using other available technologies. A single metatranscriptomic methods paper reveals that these techniques are feasible with sputum samples (Lim *et al.*, 2013). While prior research is important and has broadened our understanding of the CF lung microbiota, all of these studies are plagued with drastically low sequencing depths, effectively burying all but the most abundant organisms thought to be interesting and important for CF disease (Whelan and Surette, 2015).

1.6 Central Paradigm

It has been well characterized that the elderly population are more susceptible to respiratory infections such as pneumonia and influenza. However, the contribution of the URT microbiota to these infections is not well understood. Sibley et al. (2011) and others have shown that the combination of both culture-independent and -dependent methods enhances the sensitivity of either method alone. The CF lung is not only inhabited by principle CF pathogens but a diverse microbial community. In order to better understand CF disease, particularly what drives the onset of PE, this community needs to be better understood in its totality.

I hypothesize that using next generation sequencing technologies to appreciate the totality of the microbial communities in the upper and lower airways will allow us a bettering understanding of health and disease. Further, using these technologies in parallel with culture-dependent techniques will increase the diversity of organisms recovered. This approach will be applied to the URT of elderly individuals and the LRT of those with cystic fibrosis in order to better understand the microbial contribution to these states. Previous to these studies, a comprehensive assessment of available 16S rRNA technologies will be performed and a reproducible pipeline will be established in order to obtain the most biologically relevant interpretation of the results.

1.6.1 Specific Hypotheses

1. I hypothesize that the microbial communities of the URT of elderly individuals is altered in comparison to adults, making these individuals more susceptible

to infection.

2. I hypothesize that using culture-dependent methods in conjunction with culture-independent advancements in sequencing technologies will improve the taxonomic and functional resolution of the CF lung microbiota in order to elucidate microbial processes within the CF lung that contribute to the onset of PE.

1.6.2 Aims

Based on the above hypotheses, my research has been broken into the following Aims as follows:

1. Create a reproducible, standardized bioinformatic pipeline for the processing of 16S rRNA gene sequencing data as a means of obtaining the most biologically accurate results in order to test our hypotheses with (**Chapter 2**).

2. Compare the anterior nares and oropharyngeal microbiota in a cohort of elderly individuals to publicly available samples of mid-aged adults from the HMP to determine whether the differences between these age groups observed in the gut microbiota are mirrored in the URT (**Chapter 3**).

3. Conduct a longitudinal assessment of the CF lung microbiota in a subset of individuals with CF in order to elucidate any potential taxonomic changes in the CF lung microbiota upon PE onset (**Chapter 4**).

4. Conduct both direct and culture-enriched molecular profiling of the 16S rRNA and metagenomic populations in CF patients during times of both exacerbation and relative stability (**Chapter 5**).

Chapter 2

sl1p: A computational pipeline for the processing and analysis of 16S rRNA microbiome sequencing data

Preface

Research presented as part of this chapter has been submitted for publication as

Whelan FJ, & Surette MG. sl1p: A computational pipeline for the processing and analysis of 16S rRNA microbiome sequencing data. *Submitted, Microbiome.*

Author Contributions: FJW is the primary, first-author of this submitted manuscript. FJW wrote all of the computer code that makes up the sl1p software, conducted all analyses, and wrote the manuscript text. FJW and MGS contributed to the intellectual design of this manuscript. All authors approved the final manuscript.

The only alterations made to this publication were for thesis continuity and formatting. Supplemental material prepared as part of this manuscript is presented in **Appendix A**.

Title page and author list

sl1p: A computational pipeline for the processing and analysis of 16S rRNA microbiome sequencing data.

Fiona J. Whelan[1], & Michael G. Surette[1,2,*]

[1]Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Canada, [2] Department of Medicine, McMaster University, Hamilton, Canada

∗ To whom correspondence should be addressed:
surette@mcmaster.ca

2.1 Abstract

Advances in next-generation sequencing technologies have allowed for detailed, molecular-based studies of microbial communities such as the human gut, soil, and ocean waters. Sequencing of the 16S rRNA gene, specific to prokaryotes, using universal PCR primers has become a common approach to studying the composition of these microbiomes. However, the bioinformatic analyses of the resulting millions of DNA sequences can be challenging, and a standardized protocol would aid in reproducible analyses. The Surette Lab 16S rRNA Pipeline (sl1p, pronounced"slip") was designed with the purpose of mitigating this lack of reproducibility by combining pre-existing tools into a computational pipeline. This pipeline automates the processing of raw 16S rRNA gene sequencing data to create human-readable tables, graphs, and figures to make the collected data more readily accessible. To choose high-performing, biologically-relevant processing options as defaults for sl1p, data generated from mock communities were compared using 8 OTU clustering algorithms, 2 taxon assignment approaches, and 3 16S rRNA gene reference databases. While all of these algorithms and options are available to sl1p users, through testing with human-associated mock communities, AbundantOTU+, the RDP Classifier, and the Greengenes 2011 reference database were chosen as sl1p's defaults. Finally, sl1p promotes reproducible research by providing a comprehensive log file, and reduces the computational knowledge needed by the user to process next-generation sequencing data. sl1p is freely available at `https://bitbucket.org/fwhelan/sl1p`.

2.2 Introduction

The recent surge of next-generation sequencing technologies have allowed the scientific community to use marker genes, most popular of which being the 16S rRNA gene, to more thoroughly understand mixed bacterial communities (i.e. microbiomes). However, the adoption of any new technology requires standards and quality control. Alongside a plethora of 16S rRNA gene amplicon studies, quality control efforts have addressed the standardization of experimental and bioinformatic methods. For example, laboratory standards have been proposed for the preparation and storage of biological samples (Sinha *et al.*, 2015; Dominiani *et al.*, 2014; Zhao *et al.*, 2011) as well as procedures for the isolation and sequencing of DNA which mitigate environmental contamination (Knudsen *et al.*, 2016; Salter *et al.*, 2014). Sequencing controls have greatly reduced variability between laboratories and datasets (Salter *et al.*, 2014). Similarly, efforts have been made to standardize the bioinformatic processing of amplicon sequencing results (Caporaso *et al.*, 2010c; Schloss *et al.*, 2009). Next-generation sequencing technologies are subject to varying levels of sequencing error; traditionally, processing of amplicon sequencing data has involved filtering based on input sequence quality, followed by clustering of sequences into Operational Taxonomic Units (OTUs) which are given a taxonomic label based on their similarity to a known database (for e.g. (The Human Microbiome Project Consortium, 2012b; Sze and Schloss, 2016; Planer *et al.*, 2016)). Choice of algorithms for quality filtering, OTU clustering, and taxonomic assignment have been shown to affect the downstream analysis of biologically meaningful results (Kopylova *et al.*, 2016).

OTU clustering, typically computed at 97% sequence similarity, can be divided by approach. Reference-based (or phylotyping) approaches, such as BLAST (Altschul

et al., 1990) and UCLUST-reference (Edgar, 2010), compare input sequences to a reference database. In contrast, de novo-based approaches are independent of a reference set. De novo approaches include hierarchical clustering methods such as Mothur's average linkage algorithm (Schloss *et al.*, 2009), and ESPIRIT (Sun *et al.*, 2009), as well as greedy algorithms such as CD-HIT (Li and Godzik, 2006; Fu *et al.*, 2012), DNACLUST (Ghodsi *et al.*, 2011), UPARSE (Edgar, 2013), and AbundantOTU+ (Ye, 2011). Similarly, choice of taxonomic assignment algorithm and reference database also vary across 16S rRNA amplicon studies.

Recent benchmark studies have helped identify some of the most accurate methods in each of these categories. For example, Kopylova *et al.* identified a series of clustering methods, including UPARSE and USEARCH, which outperformed the widely used UCLUST algorithm (Kopylova *et al.*, 2016). Schloss and colleagues have also presented numerous comparisons of OTU clustering algorithms to find that de novo methods out perform reference-based methods (Westcott and Schloss, 2015; Jackson *et al.*, 2016) and, more specifically, that the average neighbour algorithm often outperforms all others (Schloss and Westcott, 2011; Westcott and Schloss, 2015; Schloss, 2016). Some comparisons of taxonomic methods have also been performed (for e.g. (Mizrahi-Man *et al.*, 2013)).

Without a comprehensive workflow, this surplus of available methods for 16S rRNA gene data processing makes it difficult to identify the most accurate approaches. Further, because each step has been developed independently, processing often involves file and command line manipulations between steps; conducting these manipulations in high-throughput is often inaccessible to a traditionally trained microbiologist, and makes it difficult to reproduce or extend data analyses, hampering

collaboration. Widely used and important tools, such as QIIME (Caporaso *et al.*, 2010c) and Mothur (Schloss *et al.*, 2009), have aided in these issues; however, their step-by-step approach and various parameters represent a significant barrier to effective amplicon data processing and do not fully mitigate issues of reproducibility. To combat this need for ease-of-use, reproducible data processing, and want of a non-biased assessment of processing options, we developed the Surette Lab 16S rRNA gene sequencing pipeline (sl1p, pronounced "slip"), a 16S rRNA data processing software. sl1p takes Illumina-generated FASTQ files as input and automates all data processing to produce a reproducible OTU table with taxonomic assignments. This pipeline is compatible with any primer set or amplicon gene, and currently offers access to 8 OTU clustering algorithms, 2 taxonomic assignment options, 3 16S rRNA gene reference databases, and 2 phylogenetic outputs. As presented here, the default processing steps and software used in sl1p have been determined to be the most accurate available approaches based on their assessment with HMP synthetic communities, and a set of 190 individually picked isolates. All steps in data processing are recorded in a log file for future reference and reproducibility. sl1p is a tool designed to be accessible to the microbiologist without detailed bioinformatic training; as such, it is fully automated, needing one line input from the user upon startup. Further, the output of sl1p includes an R markdown file with the appropriate code to visualize read counts per sample, taxonomic assignments, α-, and β-diversity from which the user can begin their own analyses. sl1p is freely available at `https://bitbucket.org/fwhelan/sl1p`.

2.3 Methods

2.3.1 The sl1p pipeline

sl1p is a data processing pipeline developed for the automated, reproducible, and accurate processing of paired-end amplicon FASTQ data (**Fig 2.1 & Sup Fig A.1**). Input to sl1p includes (a) FASTQ reads in Illumina's standard FASTQ format, and (b) an 'file of filenames' file listing all FASTQ files and their location. Optionally, the user can also include a sequencing information file if they wish to use primer sets outside of the built in defaults (v3, (Bartram *et al.*, 2011); v34, (Caporaso *et al.*, 2010a); v4 (Caporaso *et al.*, 2010a; Walters *et al.*, 2016; Parada *et al.*, 2016; Apprill *et al.*, 2015)). Each step in sl1p's data processing approach is recorded in a log file, for future reproducibility; further, the standard error output of each step is recorded to an error file to aid in any necessary de-bugging.

During initialization, the user can use command line flags to deviate from sl1p's default functionality (**Sup Fig A.1**). By default, quality filtering consists of cutadapt (Martin, 2011) to trim the PCR primers from the FASTQ input, PANDAseq (Masella *et al.*, 2012) to align paired-end reads, sickle (https://github.com/najoshi/sickle) to quality trim the resulting pairs, and USEARCH (Edgar, 2010), as implemented in QIIME (Caporaso *et al.*, 2010c), to identify and remove chimeric sequences. Users have the choice of 8 OTU clustering approaches: 5 greedy algorithms including AbundantOTU+ (default; (Ye, 2011)), CD-HIT (Li and Godzik, 2006; Fu *et al.*, 2012), DNACLUST (Ghodsi *et al.*, 2011), UCLUST (Edgar, 2010), and UPARSE (Edgar, 2013), and 2 reference-based approaches, BLAST (Altschul *et al.*, 1990) and UCLUST (Edgar, 2010), which can either be strictly closed (UCLUST-ref-strict) or conduct

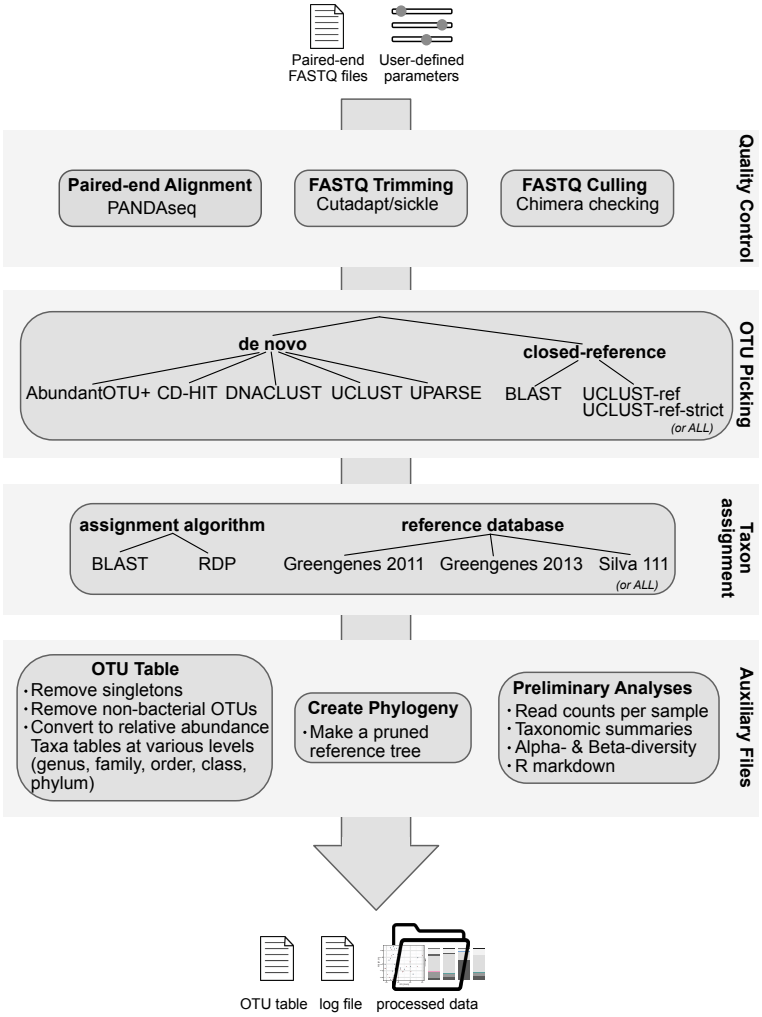


Figure 2.1: **Schematic of the sl1p pipeline.** The user input consists of FASTQ files and processing parameters. Upon input, the user can choose to deviate from the default parameters to choose from various options for OTU picking algorithms, taxonomic assignment, and reference database. Every step that sl1p utilizes is recorded in log and error files for the purposes of debugging, reference, and reproducibility. For more detail, see **Sup Fig A.1**.

closed clustering followed by de novo on any leftover sequence not matching the reference database (UCLUST-ref). Taxonomic assignment (and OTU clustering, where appropriate) can be assigned using 2 methods, BLAST or the RDP Classifier (default; (Wang *et al.*, 2007)), against 3 reference databases: Greengenes Feb 2011 (default), Greengenes Aug 2013 (DeSantis *et al.*, 2006), and Silva Release 111 (Quast *et al.*, 2013). Finally, OTU tables, phylogenies, and preliminary analyses are conducted using QIIME and R. Importantly, as part of sl1p's commandline options, the user can choose to run all possible combinations of OTU clustering algorithms, taxonomic assignment methods, and choice of reference databases automatically with one command, making comparisons of available methods reproducible and easy to approach.

The sl1p pipeline is open source and publicly available at `https://bitbucket.org/fwhelan/sl1p`. The pipeline is written in Perl and consists of one main script which calls on auxiliary scripts to aid in reformatting data between steps as necessary. Accompanying setup and install scripts are provided to download and install sl1p.

2.3.2 Generation of test datasets

The Human Microbiome Project Mock Communities (HMP-mock)

Genomic DNA of 2 unique representations of a 20 member mock community generated as part of the Human Microbiome Project (The Human Microbiome Project Consortium, 2012a) was obtained from BEI Resources (Catalog Nos. HM-782D & HM-783D; ATCC, Manassas, VA). The first sample (HMP-mock1) is an even distribution of the 20 organisms from 17 genera, whereas the second (HMP-mock2) is a staggered distribution of the same organisms (The Human Microbiome Project Consortium, 2012a). For each sample, 3 PCR replicates were generated by using 1μl of

genomic DNA PCR amplified using 1μl of dNTPs, 0.25μl of *Taq* polymerase (Life Technologies, Carlsbad, CA) and 5μl of PCR primers designed for the v3 region of 16S rRNA gene (Bartram *et al.*, 2011). These amplification products were then split across two runs of the Illumina sequencer to generate sequencing replicates. Sequencing depth ranged from 5917 to 113,084 reads with an average of 57,257. A negative PCR control was generated in parallel.

Single and Combined Isolate Controls (URTCul)

190 single colonies were picked from a collection of upper respiratory tract culture isolates (URTCul) and restreaked until pure on appropriate solid agar plates. Once pure, isolates were picked directly into 5% Chelex, boiled, and centrifuged at 13,000rpm for 5 minutes. 5μl of the supernatant was used as template for a 50μl PCR reaction of the variable regions 8F-926R (Wang *et al.*, 1999; Muyzer *et al.*, 1993) of the 16S rRNA gene and sequenced using Sanger sequencing. The resulting Sanger sequences for each isolate were taxonomically assigned using independent blastn searches against NCBI's RefSeq database. Taxonomic assignments were made to the species level; in the case of multiple species matching with percent identity within $< 1\%$ of each other, multiple species names were included in the taxonomic assignment (for e.g. g__Streptococcus;s__infantis_mitis). This dataset contained 8 unique genera and 33 unique species.

For Illumina sequencing, PCR amplification of the v34 region (341F & 806R, (Caporaso *et al.*, 2010a)) was performed and sequencing was conducted in on an Illumina MiSeq sequencer; each isolate was PCR amplified with its own unique barcoded primer (**Fig 2.2, URTCul-singles**). A contaminant was identified as any sample having

greater than 15% of reads assigned a taxonomy which differed from the taxonomy assignments of other reads at the family level or above in the OTU table produced using sl1p's defaults (9 isolates culled). The average number of sequenced reads per isolates was 12 (range 1-81).

After amplification and Illumina sequencing of each isolate individually, the raw FASTQ reads were combined *in silico* to create one sample (**Fig 2.2, URTCul-combined**). Further, the taxonomic assignments of the Sanger sequencing results were consulted to create a second *in silico* sample in which only uniquely identified taxa were combined (**Fig 2.2, URTCul-uniques**). The artificial sequencing depths of these 2 samples were 2148 and 423, respectively.

Publicly available dataset

Additionally, a publicly available dataset of human fecal microbiota samples (Bioproject Submission SUB2392090; (Moayyedi *et al.*, 2015)) was used in testing the phylogenetic outputs of sl1p displayed in Figure 2.6.

2.3.3 Data processing comparisons

All output data processing comparisons were based on OTU tables, map files, and phylogenies generated by sl1p v4.1 using the -p all -d all and -t all flags. All analyses were computed in R using phyloseq (McMurdie and Holmes, 2013), ggplot2 (Wickham, 2009), and reshape2 (Wickham, 2007) with the following exceptions. FastQC (Andrews, 2010) was used to calculate FASTQ quality scores used in Figure 2.3. Graphlan (Asnicar *et al.*, 2015) was used to visualize phylogenies as presented in Figure 2.6. All data processing was computed on a standard personal desktop computer

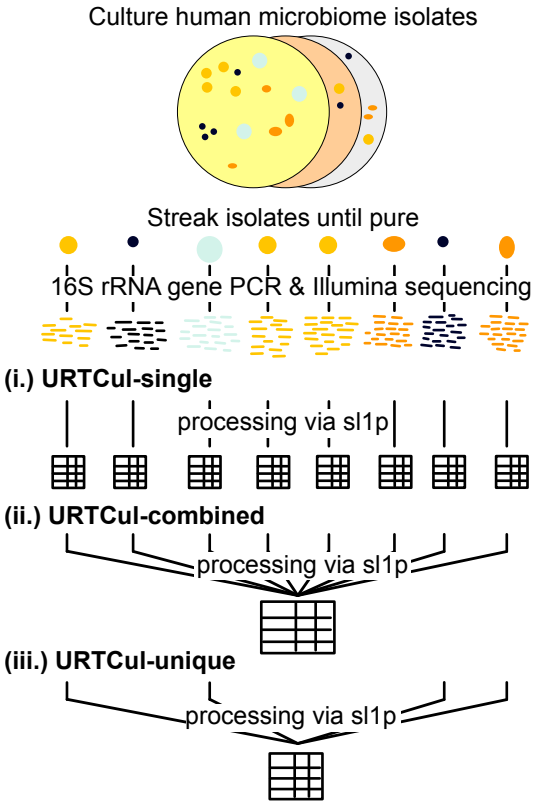


Figure 2.2: **Schematic of URTCul mock community generation.** Isolates were individually picked from solid agar plates and amplified using Sanger (amplicon length=918 bps) and Illumina (amplicon length=250bp) sequencing approaches. Following Illumina sequencing, the resulting reads from each individually sequenced isolate were analyzed individually (**URTCul-singles**), in combination (**URTCul-combined**), or as a combination of each uniquely identified taxa (**URTCul-uniques**).

running Ubuntu 14.04.

2.4 Results

The Surette laboratory 16S rRNA processing pipeline (sl1p) was developed as an automated and reproducible 16S rRNA gene sequencing processing tool. In order to determine the most accurate default settings of this tool, we systematically tested various approaches within the sl1p workflow using 2 approaches (i.) 2 HMP mock community samples (HMP-mock), and (ii.) 190 single bacterial isolates (URTCul-singles) and their combination as a totality of the 190 sequencing results (URTCul-combined) or the combination of unique taxa from this pool (URTCul-uniques).

2.4.1 sl1p removes low quality reads effectively

One of the consequences of using next-generation sequencing technologies in high-throughput is the propensity for sequencing error. For instance, Illumina technology is known to have an increased error rate towards the 3' end of the read, and that the reverse read is generally of poorer quality then the forward. Mitigating this error prior to OTU generation and taxonomic assignment is essential in order to refrain from the generation of spurious OTUs.

sl1p utilizes a multi-step approach to quality control. Immediately following removal of sequencing primers with cutadapt, forward and reverse reads are assembled using PANDAseq. While many options are available for the merging of paired-end reads, PANDAseq includes both quality filtering and read assembly. Across our PCR and sequencing replicates of HMP-mock, approximately 12.5% of raw input reads

were culled at this step (**Fig 2.3a**); the majority of culled reads were due to misalignment of forward and reverse reads. Following, cutadapt was used to remove any reads containing Illumina annealing or sequencing primers. While this step removed only 2.7% of the HMP-mock input (**Fig 2.3a**), we have found it to be an important way of removing erroneous sequencing results, and a measure of an infrequent poor Illumina sequencing run. Next, sickle was used to trim quality sequence (and to remove any reads $<$ 100 bp post-trimming). It is at this stage where the most quality-filtering is done, with an average 29.6% read loss (**Fig 2.3a**). However, it is this strict quality filtering that results in clean, high-quality paired-end reads (**Fig 2.3b**); when we compare this strict threshold with lower quality cutoffs, we begin to see a decline in the final paired-end read quality as the cutoff drops below 30 (**Sup Fig A.2**).

The last step in sl1p's quality control workflow is chimera checking. Because 16S rRNA gene amplicon data is generated via PCR amplification, chimeric sequences can be an issue, especially if the PCR amplification reaction traverses a highly conserved region as is the case for multi-variable region amplicons. As such, sl1p uses QIIME's implementation of USEARCH to conduct chimera checking on the generated paired-end reads (**Fig 2.1 & Sup Fig A.1**). This approach is database-dependent; however no significant differences were observed between sl1p's 3 options for reference database (removal of 0.36%, 0.4%, and 0.39% of reads for Greengenes 2011, 2013, and Silva Release 111, respectively on the HMP-mock data).

Following sl1p's quality control workflow, an average of 55.2% of the raw input HMP-mock reads remain. This percentage is higher than that found with the URTCul dataset (mean of 30.4%); a greater number of unassembled paired-end reads (57.9%

of raw input removed during PANDAseq alignment) were observed with the URTCul v34 sequencing, possibly due to the shorter overlap in the target sequence (**Sup Fig A.3**).

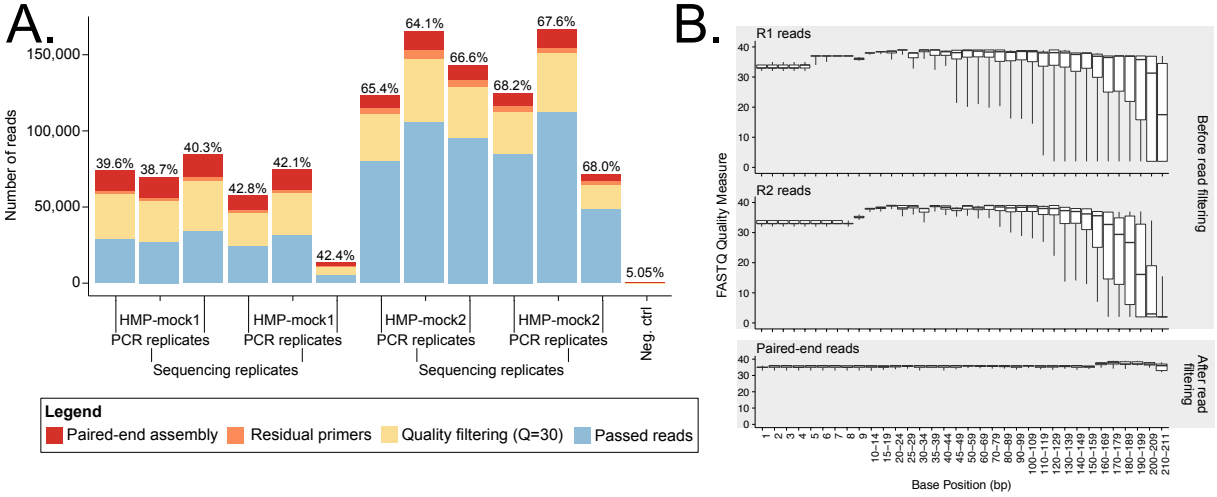


Figure 2.3: **sl1p effectively removes low quality reads. A.** sl1p's quality control workflow consists of paired-end assembly, removal of residual primers, quality trimming, and length filtering. Here, the number of reads culled at each step is presented. Inline percentages indicate the percentage of raw input reads which remain following the quality control process. **B.** This process successfully removes bases of low quality from the resulting paired-end reads as demonstrated here on the raw sequence input from 2 unique mock HMP samples sequenced using 3 PCR and 2 sequencing replicates.

2.4.2 OTU clustering algorithms produce varying numbers of OTUs compared to known input

Clustering of input reads into Operational Taxonomic Units (OTUs) has been the most well-studied effect on processed reads (Huse *et al.*, 2010; Schloss and Westcott, 2011; Barriuso *et al.*, 2011; Sun *et al.*, 2012; May *et al.*, 2014; Kopylova *et al.*, 2016). OTUs are typically clustered based on a 97% threshold based upon imperial studies identifying this as the differentiating threshold of species (Konstantinidis and Tiedje,

2005); however when sequencing is restricted to small regions within the gene, this threshold may provide differentiation between the genus and species level, depending on the organism in question (Mizrahi-Man *et al.*, 2013).

sl1p provides 8 OTU clustering approaches from which the user can choose from upon initialization of the pipeline. As expected, *de novo* clustering methods produce observed OTU numbers independent of the reference database, whereas some variability in observed OTUs is seen with reference-based approaches (**Fig 2.4**). Most of these options over-estimate the number of OTUs within the HMP-mock and URTCul datasets when compared to the known taxonomic composition (**Fig 2.4**). This is perhaps the most evident in the HMP-mock dataset where some algorithms, such as DNACLUST, over-estimated sample diversity by almost 40x (**Fig 2.4a, Sup Figs A.4-A.5**). When Swarm (Mahe *et al.*, 2014) was compared using sl1p-generated quality filtered reads, it also over-estimated sample diversity, though the removal of singletons greatly reduced the number of spurious OTUs (**Sup Fig A.6**). When OTUs with a successively small number of defined reads where culled, the number of observed OTUs quickly converged to the expected community diversity (**Sup Fig A.7**), suggesting that these spurious OTUs are often due to low abundance reads. Other algorithms, such as UPARSE, under-estimated OTU abundance (**Fig 2.4, Sup Figs A.4-A.5**). Of those tested, the approaches which most closely estimated within sample OTU diversity in the HMP-mock samples were AbundantOTU+, UCLUST closed reference picking, and UPARSE.

Within the URTCul-single dataset, in which each sample consisted of DNA from a single bacterial colony, many OTU picking algorithms over-estimated sample diversity in multiple samples (**Fig 2.4b**). UPARSE, with its own approach to sequence

Table 2.1: **CPU time for OTU clustering approaches implemented in sl1p.** All calculations were computed on a standard Desktop running Ubuntu 14.04.

OTU picking approach	**CPU time (in mins)**
AbundantOTU+	3.38
BLAST	127.17
CD-HIT	13.32
DNACLUST	0.08
UCLUST	0.21
UCLUST-ref	0.69
UCLUST-ref-strict	0.82
UPARSE	0.28

quality control (**Sup Fig A.1**), often underestimated sample diversity. However, many approaches, including AbundantOTU+, CD-HIT, DNACLUST, and UCLUST often identified the sole OTU within the sample (**Fig 2.4b**). When these individually sequenced isolates were combined, most OTU picking approaches estimated sample diversity between the known number of genera and species present within the samples(**Fig 2.4c**). Notably, UPARSE again under-estimated diversity, generating 9 and 5 OTUs in the URTCul-combined and -unique samples, which consisted of 33 species from 8 genera. As next-generation sequencing approaches become more accessible to this field, the feasibility of implementing these methods on a common laboratory desktop is increasingly more practical and should be considered (**Table 2.1**).

2.4.3 Choice of data processing algorithms affect taxonomic assignment

However, as has been previously addressed (Sun *et al.*, 2012), what is more important than simply the number of OTUs produced is how the taxonomic assignment

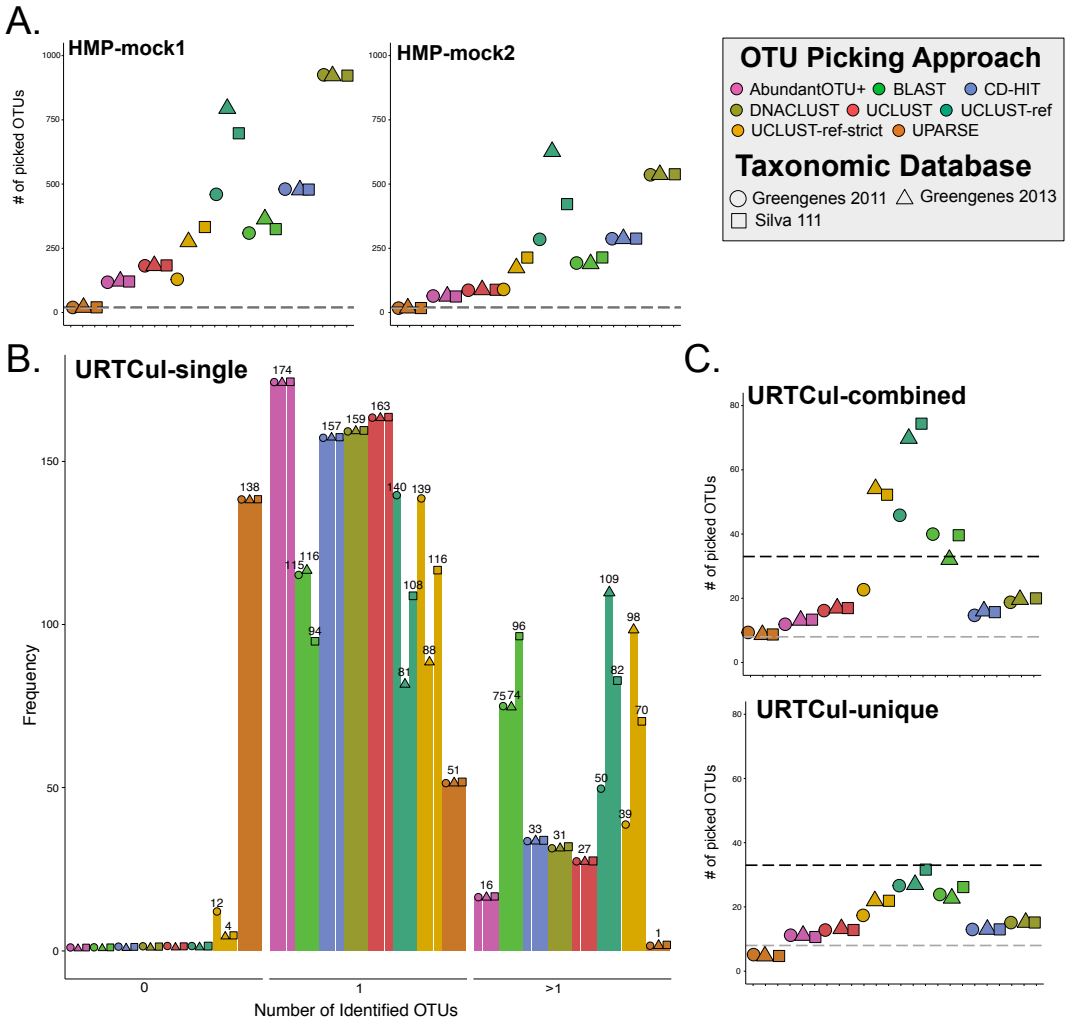


Figure 2.4: **OTU clustering methods perform variably. A.** 8 methods were used on control communities of known composition to report OTU counts compared to known sample diversity (black dotted lines = number of genus; grey dotted line = number of species). Non-bacterial sequences were removed as part of sequence processing. Similar results were obtained when singletons were also removed (**Sup Fig A.4**). **B.** A group of 190 single isolates were independently sequenced in order to test varying OTU clustering algorithm's ability to correctly identify 1 OTU within the input sample. **C.** When these individual isolates were combined, the number of OTUs generated often lies between the known number of unique genus and species within the samples.

and corresponding relative abundance of each taxa compares to the known sample composition that truly matters. To measure this, we compared the known composition of the mock datasets to the OTU composition generated via sl1p's options for OTU clustering, taxon assignment, and reference database (**Fig 2.5**). The processing options which showed the most similarity to a given mock community was highly sample-dependent; for example, a combination of UPARSE, BLAST, and reference database Greengenes 2011 showed the most similarity to the HMP-mock1 sample whereas AbundantOTU+ and the RDP Classifier replaced UPARSE and BLAST as the most accurate OTU picking algorithm and taxonomic assignment method in HMP-mock2 and URTCul-combined (**Fig 2.5a,c**). Further, the combination which produced the most similar output to the known composition of HMP-mock1 (UPARSE, BLAST, and Greengenes 2011) produced one of the least similar outputs in URTCul-combined (**Fig 2.5a,c**). In the URTCul-singles dataset, the most abundant OTU's taxonomic assignment was compared with the results of taxonomic assignment based on full-length Sanger sequencing of the 16S rRNA gene. In this dataset, the RDP Classifier produced the highest number of correctly assigned taxa accompanied with either Greengenes 2011 and the Silva database (**Fig 2.5b**). These results indicate the impact of sample composition as well as choice of OTU picking approach, taxon assignment method, and reference database on the underlying biological implications of these data.

To further quantify these differences, comparisons can be made between the known taxa and relative abundance compared to each set of OTU picking, taxonomic assignment, and reference database options (**Sup Fig A.8-A.9**). At this level of resolution,

independent of the number of OTUs assigned to each genera, we can see that the proportions of each genera output from sl1p reflect the expected proportions in each of the HMP-mock samples. However, in some sets of processing options, some mistakes are made in taxonomic assignment. The combination of the RDP Classifier and Greengenes 2013 database, for example, incorrectly identifies *Flexispira* of the family *Helicobacteraceae* in place of the *Heliobacter* genus (**Sup Fig A.8-A.9**). In other cases, the correct assignment is made, though more conservatively left at the family, order, or class level (**Sup Fig A.8-A.9**); for example, Greengenes 2013 using BLAST as the taxon assignment algorithm assigns some OTUs to the class *Bacilli*, failing to differentiate between the *Bacillus, Listeria, Staphylococcus, Enterococcus,* and *Streptococcus* species present in this mock community. Overall, across all methods and the HMP-mock samples, BLAST in combination with Greengenes 2011 was the only combination to provide no errors in taxonomic assignment. This accuracy comes with a small increase in computing time compared with the RDP Classifier (data not shown).

2.4.4 Choice of processing methods affect biologically relevant results of 16S rRNA gene sequencing

Like all bioinformatic pipelines and processing workflows, what is most important in the output is the reflection of the true underlying biology in the results. While 16S rRNA sequencing data can be analyzed in a number of ways in order to answer many unique research questions, calculations of α and β diversity are often fundamental to analyses. α diversity, or within sample diversity, is a calculation performed on each

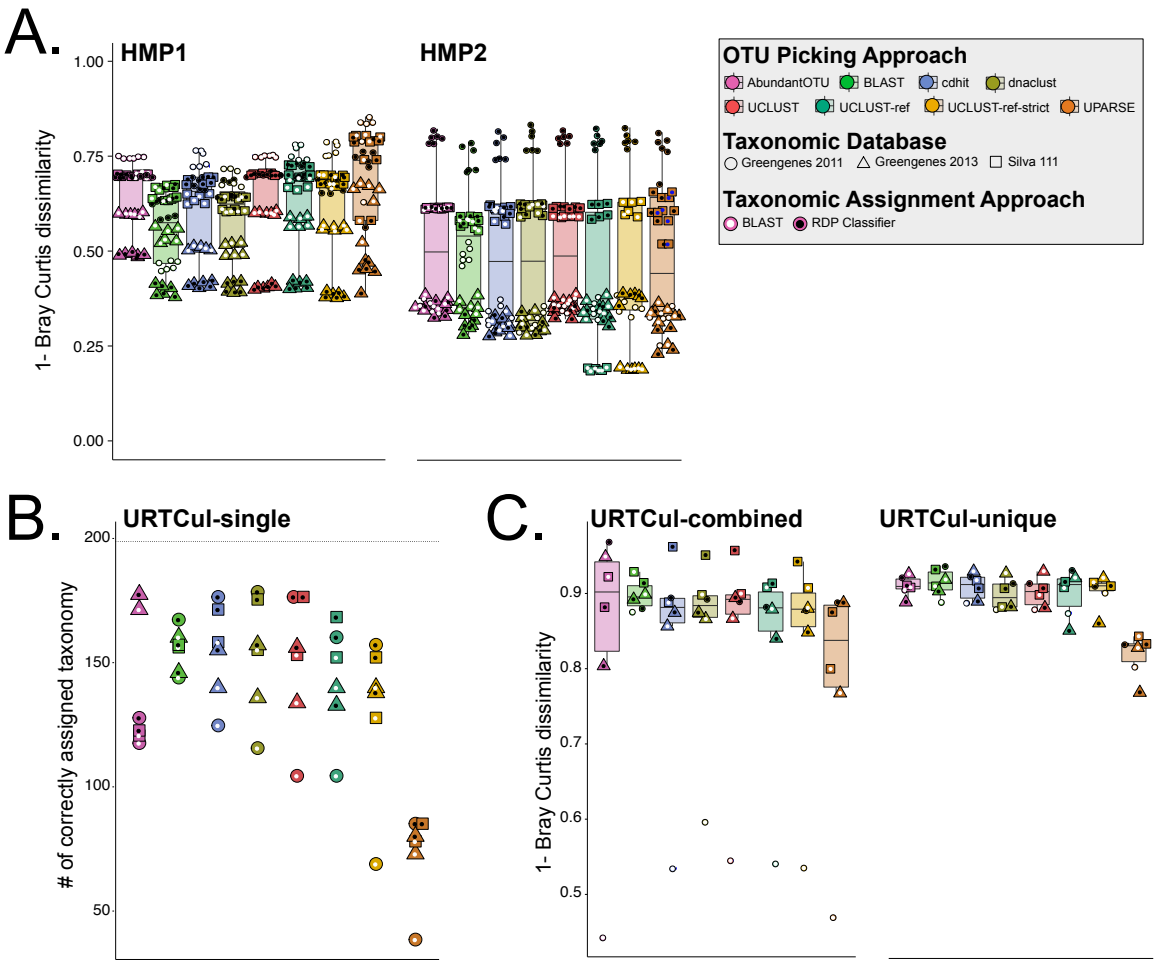


Figure 2.5: **Taxonomic assignment is dependent on up-stream choices in 16S rRNA gene processing.** sl1p implements 2 methods of taxon assignment across 3 reference databases. By running all methods, we compared taxon assignment against an expected control samples. **A.** The negated Bray-Curtis dissimilarity was used to identify which taxonomically assigned OTU sets most closely matched the known composition of the mock HMP communities **(A.)** and the combined URTCul isolates **(C.)**. **B.** In a set of 190 single isolate samples, the number of samples whose most abundant OTU correctly matched full-length 16S rRNA Sanger sequencing results is displayed.

sample within a dataset. This metric can be calculated using different indices depending on the question at hand. Popular approaches include Shannon and Simpson diversity as these indices incorporate both evenness and richness of the community into their calculations (Shannon, 1948), (Simpson, 1949). Other metrics, such as Chao1, are estimates of species richness (Chao, 1984). Using output of the sl1p processing pipeline, we calculated the Shannon, Chao1, and Simpson diversity metrics on the HMP-mock data (**Fig 2.6a & Sup Fig A.10**). Here, only the OTU clustering algorithm contributes to the estimated richness and evenness of OTU composition, except in the case of reference-based algorithms which are database-dependent (**Sup Fig A.1**). We observe that the output of α diversity metrics is dependent on the processing methods employed. The range of calculated Shannon diversity scores within the same sample processed using different commonly-used approaches is greater than 1.0 (range 1.54-2.84) (**Fig 2.6a**). Similarly, Chao1 estimates species richness anywhere from 20 to 2451 depending on data processing options employed; Simpson diversity, in contrast, has much less observed variability between OTU clustering methods and reference database choice. Interestingly, these metrics are also affected by changes in read depth as seen in the variation between sequencing replicates (**Sup Fig A.10a**); rarefaction of reads somewhat reduces this variation (**Sup Fig A.10b**).

β, or between-sample, diversity is often used as a measure of difference between $\geq$1 sample state (e.g. health and disease). Similar to α diversity, there are a variety of distance metrics one can utilize depending on the question at hand. A popular set of these metrics use the phylogenetic distances between OTUs as a contributor to the distance score. Using sl1p, we discovered that the output of these metrics

are dependent on how the accompanying phylogenetic tree is generated (**Fig 2.6b-c**). Comparisons using Procrustes analysis show substantial differences in the PCoA plots generated using the weighted UniFrac method with different phylogenetic inputs (**Fig 2.6b-c**). One approach recommended in the QIIME workflow, is the use of PyNAST (Caporaso *et al.*, 2010b) and FastTree (Price *et al.*, 2009) to create a multiple sequence alignment and phylogeny of the representative sequence from each OTU in the community (**Fig 2.6d, default phylogeny**). However, because this phylogeny is reliant on the sequence diversity within the sequenced variable region, which is often $\leq$100-300bp in length, it often does not reflect the true bacterial phylogeny but instead creates paraphyletic phyla (**Fig 2.6c**). Because of this, sl1p generates an alternate phylogeny which represents the Greengenes reference 16S rRNA gene phylogeny trimmed to those OTUs present within the given dataset. Beginning with a curated phylogeny means that the phylogenetic relationships between organisms within a given sample set are preserved. Using these phylogenies to generate the Weighted and Unweighted UniFrac metrics, summarized here as Principal Coordinate Analyses (PCoAs), results in differences in the calculated distance between the samples within this community (**Fig 2.6b-c**). These results indicate that processing options greatly affect the output and potential interpretation of 16S rRNA gene sequencing results.

2.5 Discussion

sl1p is an automated, reproducible 16S rRNA gene sequencing processing pipeline that makes 16S rRNA data processing accessible to those without formal bioinformatics training. sl1p is not restricted by variable region or choice of PCR primer set.

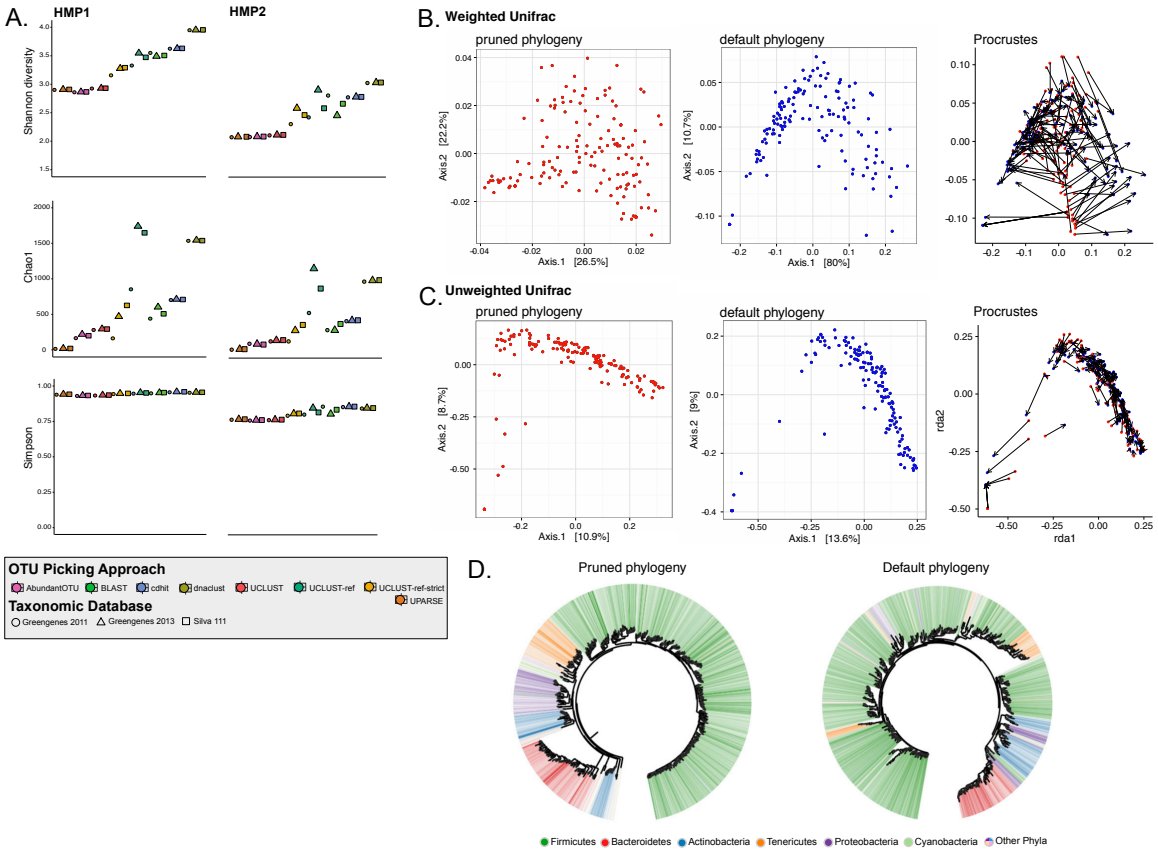


Figure 2.6: **Analyses of biologically-meaningful outputs are dependent on 16S rRNA sequence processing. A.** α diversity metrics vary greatly between OTU picking approaches, and are dependent on choice of reference database in the case of reference-based OTU clustering methods. **B-C.** Phylogeny-dependent β diversity metrics, including Weighted UniFrac **(b)** and Unweighted UniFrac **(c)**, differ depending on the method of phylogeny-generation. A comparison of the distribution of samples via a Procrustes analysis, indicates the impact that the phylogenetic tree makes on these data. **D.** sl1p generates 2 phylogenies. The default phylogeny represents the phylogeny generated as part of the default QIIME workflow. The pruned phylogeny is generated by sl1p by pruning the Greengenes reference phylogeny to those branches which are present within the sample set.

In this study, we outline the workflow of this tool, which can be broken down into 3 main steps: FASTQ quality control, OTU clustering, and taxonomic assignment (**Fig 2.1**). We also show how sl1p can aid in the comparison of multiple options and the effects they have on downstream analyses. The quality control workflow within sl1p was determined based on the parameters necessary to obtain high quality base pair assembly along the length of each paired-end sequence (**Fig 2.3**). In order to compare the effect of various OTU picking approaches and taxonomic assignment methods, mock communities were employed. Comparisons of OTU clustering algorithms displayed a wide range of predicted OTUs, generally over-estimating diversity. This, as well as the under-estimations made by UPARSE (**Fig 2.4**), have been previously shown (Kopylova *et al.*, 2016; Schloss, 2016; Westcott and Schloss, 2015). Further, the choice of taxonomic assignment algorithm and reference database greatly influenced the predicted taxonomic composition of the communities (**Fig 2.5**).

Most importantly, the use of sl1p to compare data processing outputs (OTU tables, taxonomic summaries, and phylogenies) recognizes the effect processing options have on biological analyses (**Fig 2.6**). Popular α diversity metrics such as Shannon diversity is greatly affected by OTU clustering option and sequencing depth (**Fig 2.6a**). These results have implications on the interpretation of microbiome studies across manuscripts and research groups which may process their data using different methods. Further, the differences between sequencing runs have implications for studies which are split across multiple sequencing runs due to size. Importantly, the rarefaction of these data did not fully mitigate these effects (**Sup Fig A.10b**). Further, the alternative phylogenetic representation of the OTU data generated by sl1p better describes the bacterial tree of life, allowing for more accurate β diversity

distances to be calculated between samples, furthering our knowledge of differences between varying microbial communities.

The default parameters of sl1p were carefully chosen based on the analyses presented within this study. Of course, all algorithms and tools tested have their own merits and niches within this widely growing field, and is reflected in the fact that no set of tools out performed others in all circumstances (**Fig 2.4-2.6**). We chose AbundantOTU+ as the default OTU picking approach. AbundantOTU+ most closely predicted the correct number of OTUs within HMP-mock1, HMP-mock2, URTCul-combined, and URTCul-unique, without under-estimating diversity. AbundantOTU+ was also the tool able to correctly predict the highest number of single isolate samples in the URTCul dataset. This method also performed well in tests of correctly identified taxa, including the Bray-Curtis dissimilarity comparisons. For choice of taxon assignment algorithm, we chose the RDP Classifier as sl1p's default. This tool consistently calculated the most number of accurate URTCul-singles isolates, and out-performed or tied BLAST performance on Bray-Curtis dissimilarity comparisons in all cases except for HMP-mock1. Lastly, Greengenes 2011 is sl1p's default reference database based on its superior performance in the Bray-Curtis distance comparisons of the HMP-mock communities and as one of the best choices for genus-level taxon identification.

It is important to note that these default parameters are based on mock communities of human-associated microbes and may not represent the best combination of tools in the study of other microbiota. The authors hope that by providing a pipeline in which multiple OTU picking, taxonomic assignment, and reference database options are easily accessible, that the user can choose to easily deviate from these defaults or

conduct a subset of these comparisons, as they see fit.

The field of microbiome research is growing, and with it, new approaches to data processing are in development. As such, we have written the sl1p code in a manner that will allow for the easy addition, and subsequent testing, of additional approaches to the pipeline. Because sl1p is freely available, others are free to modify the code as they wish or to request improvements from the authors. Having a non-biased method for tool comparison will be important for the maturation of this field.

In conclusion, we present a 16S rRNA gene sequence processing workflow with the aim of generating the most biologically meaningful outputs for the furthering of 16S rRNA gene sequencing techniques and microbiome research in general.

2.6 Acknowledgements

The following reagents were obtained through BEI Resources, NIAID, NIH, as part of the Human Microbiome Project: Genomic DNA from Microbial Mock Community B (Even, Low Concentration), v5.1L, for 16S rRNA Gene Sequencing, HM-782D & Genomic DNA from Microbial Mock Community B (Staggered, Low Concentration), v5.2L, for 16S rRNA Gene Sequencing, HM-783D. THE URTCul sequencing data was generously provided by Steve P. Bernier and Laura Rossi. The authors wish to thank the efforts of the HMP and the Surette laboratory for depositing sequence data in publicly available resources which allow studies like these. The authors also wish to thank the efforts of the Surette laboratory in helping to make this pipeline more effective and user-friendly.

Chapter 3

The loss of topography in the microbial communities of the upper respiratory tract in the elderly

Preface

Research presented as part of this chapter has been previously published as

Whelan FJ, Verschoor CP, Stearns JC, Rossi L, Luinstra K, Loeb M, Smieja M, Johnstone J, Surette MG, & Bowdish DME. (2014). The loss of topography in the microbial communities of the upper respiratory tract in the elderly. *Annals of the American Thoracic Society*, 11(4), 51321.

Author Contributions: FJW is the primary, first-author of this published manuscript. FJW, CPV, MGS, and DMEB conceived and designed the experimental approach. FJW, CPV, LR, MS, and KL performed experiments essential to this manuscript. FJW conducted all data analysis and prepared all Figures within this manuscript. ML and JJ performed sample collection. FJW and DMEB drafted and edited the manuscript. All authors approved the final manuscript.

The only alterations made to this publication were for thesis continuity and formatting. Supplemental material published as part of this manuscript is presented in **Appendix B**.

Title page and author list

The Loss of Topography in the Microbial Communities of the Upper Respiratory Tract in the Elderly

Fiona J. Whelan[1,2], Chris P. Verschoor[2,3,4], Jennifer C. Stearns[7], Laura Rossi[1,2], Kathy Luinstra[3], Mark Loeb[2,5], Marek Smieja[2,3,5,7], Jennie Johnstone[5], Michael G. Surette[1,2,6,7], and Dawn M. E. Bowdish[2,3,4,*]

[1]Department of Biochemistry and Biomedical Sciences, [2]Institute for Infectious Disease Research, [3]Department of Pathology & Molecular Medicine, [4]McMaster Immunology Research Centre, [5]Department of Clinical Epidemiology and Biostatistics, [6]Farncombe Family Digestive Health Research Institute, and [7]Department of Medicine, McMaster University, Hamilton, Ontario, Canada

∗ To whom correspondence should be addressed:
Dawn M. E. Bowdish, Ph.D., McMaster University, Department of Pathology & Molecular Medicine, McMaster Immunology Research Centre, M. G. DeGroote Institute for Infectious Disease Research, 1280 Main Street West, Hamilton, ON, L8S 4K1 Canada. E-mail: bowdish@mcmaster.ca

Keywords: aged; microbiome; 16S rRNA; respiratory tract infections

3.1 Abstract

Rationale: The microbial communities inhabiting the upper respiratory tract protect from respiratory infection. The maturity of the immune system is a major influence on the composition of the microbiome and, in youth, the microbiota and immune system are believed to mature in tandem. With age, immune function declines and susceptibility to respiratory infection increases. Whether these changes contribute to the microbial composition of the respiratory tract is unknown.

Objectives: Our goal was to determine whether the microbes of the upper respiratory tract differ between mid-aged adults (18−40 yr) and the elderly (>65 yr).

Methods: Microbiomes of the anterior nares and oropharynx of elderly individuals were evaluated by 16S rRNA gene sequencing. These communities were compared with data on mid-aged adults obtained from the Human Microbiome Project.

Measurements and Main Results: The microbiota of the elderly showed no associations with sex, comorbidities, residence, or vaccinations. Comparisons of mid-aged adults and the elderly demonstrated significant differences in the composition of the anterior nares and oropharynx, including a population in the anterior nares of the elderly that more closely resembled the oropharynx than the anterior nares of adults. The elderly oropharyngeal microbiota were characterized by increased abundance of streptococci, specifically, *Streptococcus salivarius* group species, but not *Streptococcus pneumoniae*, carriage of which was low (<3% of participants), as demonstrated by PCR (n = 4/123).

Conclusions: Microbial populations of the upper respiratory tract in mid-aged adults and the elderly differ; it is possible that these differences contribute to the increased risk of respiratory infections experienced by the elderly.

3.2 Introduction

In a healthy individual, the nostrils, lined with ciliated epithelia, are responsible for filtering the air we breathe for environmental particles and bacteria (Lemon *et al.*, 2010). Similarly, the oropharynx is the first line of defense against microbes from ingested foods and inhaled air (Lemon *et al.*, 2010). Even though their roles are analogous, these biogeographies possess distinct microbial communities. In mid-aged adults, the nostril, or anterior nares, is dominated by *Actinobacteria* (*Propionibacterium, Corynebacterium*) and *Firmicutes* (*Staphylococcus*), whereas *Firmicutes* (*Veillonella, Streptococcus, Staphylococcus*) are more prevalent in the oropharynx (Lemon *et al.*, 2010; Charlson *et al.*, 2011). When the defences at these two locales are bypassed, disease-causing microbes may colonize the upper respiratory tract and be aspirated into the lower respiratory tract and lung, causing severe respiratory infection (Feinberg *et al.*, 1990; Scannapieco, 1999).

Children (<5 yr) and elderly adults (>65 yr) are more susceptible to infections that originate in the upper respiratory tract, such as influenza (Ahmed *et al.*, 2007) and pneumonia (Jokinen *et al.*, 1993), than are mid-aged adults. In children, this susceptibility steadily declines as the microbial communities and immune responses of their upper respiratory tract mature (Oh *et al.*, 2012; van Benten *et al.*, 2005). In the elderly, an aging immune system has been implicated in the increase of these infections (Franceschi *et al.*, 2000); however, the effect of the upper respiratory tract microbiota has not yet been considered. To investigate this possibility, we characterized the microbial communities within the upper respiratory tract of an elderly population using 16S rRNA gene sequencing. The microbial communities of the anterior nares and oropharynx of nursing home elderly were compared with the National

Institutes of Healths publicly available Human Microbiome Project. As part of this project, 242 adults, aged 18−40 years, were sampled at up to 18 body sites (NIH HMP Working Group *et al.*, 2009). Our findings indicate that the distinct nasal and oropharyngeal microbiota present in adult populations is lost with age, and that the nasal community is replaced with an oropharyngeal-like population of microbes. Both locales are marked with a distinct increase in the relative abundance of Streptococcus with age. We hypothesized that this increase of the streptococci might include pathogenic species, such as Streptococcus pneumoniae, as the elderly have a high rate of pneumococcal infections. We found, however, that only a small number of the nursing home elderly (n = 4/123) had carriage of this species. Together, these results indicate that both the anterior nares and oropharynx microbiota differ significantly between mid-aged and elderly adults.

3.3 Methods

3.3.1 Participant selection criteria and sample collection

Elderly participants (68−96 yr old; mean = 80, n = 18 [13 females]) were recruited from four nursing homes in Ontario, Canada between October and December 2010; exclusion criteria included the use of immunosuppressive medications. Nasal swabs (Copan ESwabs; Copan Diagnostics Inc., Murrieta, CA) were obtained from right and left anterior nares, while throat swabs were obtained from the rear of the oropharynx. Swabs were immediately submerged in Liquid Amies medium (Copan Diagnostics Inc.), which was aliquoted into microcentrifuge tubes and cryopreserved at 22°C until further use. These studies were approved by the McMaster Research Ethics Board,

and informed consent was obtained for all participants.

3.3.2 DNA extraction and 16S rRNA gene amplification

DNA extraction was performed using a custom protocol and sequence amplification of the 16S rRNA gene variable 3 (v3) region was conducted as previously described (Bartram *et al.*, 2011). Briefly, 300μl of sample was resuspended in 800μl of 200 mM $NaPO_4$, 100μl of guanidine thiocyanate–ethylenediaminetetraacetic acid–Sarkosyl. The solution was homogenized using 0.2 g of 0.1−mm glass beads (Mo Bio, Carlsbad, CA). Enzymatic lysis was performed using 50μl lysozyme (100 mg/ml), 50μl mutanolysin (10 U/μl), 10μl RNase A (10 mg/ml), and incubation at 37°C for 1 hour followed by the addition of 25μl 25% sodium dodecyl sulfate, 25μl proteinase K, and 62.5μl 5M NaCl followed by incubation at 65°C for 1 hour. Samples were then pelleted via centrifugation at 12,000 X g and supernatant removed to a new microcentrifuge tube. An equal volume of phenol-chloroform-isoamyl alcohol was added and the sample centrifuged. The solution with the lowest density was transferred to a new microcentrifuge tube to which 200μl of DNA binding buffer (Zymo, Irvine, CA) was added. Solution was transferred to a DNA column (Zymo), washed, and DNA eluted using sterilized H2O.

Following this protocol, amplification of the 16S rRNA gene v3 region was performed as previously described (Bartram *et al.*, 2011). Briefly, 341F and 518R 16S rRNA primers were modified for adaptation to the Illumina (San Diego, CA) platform and included the addition of 6–base pair, unique barcodes to the reverse primer, allowing for multiplex amplification. PCR amplification, separation by gel electrophoresis, and gel extraction steps were completed as described in Reference (Bartram *et al.*,

2011), with some modifications. Briefly, the amount of primer used was decreased to 5 pmol each, a *Taq* polymerase (Life Technologies, Carlsbad, CA) was used for amplification, and the cycling times were changed to 30 seconds for each step. Products were then sequenced using the Illumina MiSeq platform.

3.3.3 Acquisition of National Institutes of Health human microbiome project data

Publicly available data collected from 242 healthy participants aged 18–40 years as part of the National Institutes of Healths Human Microbiome Project were used as a resource to compare the microbial communities of healthy mid-aged adult upper respiratory tracts to those in our nursing home cohort. It was ensured that the samples collected from the anterior nares and oropharynx of our nursing home cohort followed the same protocols as that of the Human Microbiome Project. Raw 16S rRNA gene v13 and v35 sequences of all Human Microbiome Project samples was obtained from their website (http:// hmpdacc.org) along with the available metadata (sex and sequencing location). Sequences sampled from the anterior nares and throats of individuals were extracted from publicly available sequencing results using custom Perl scripts created by F.J.W. and available upon request. The protocol and technologies used by the Human Microbiome Project allowed these samples to be sequenced from the v3 region through to v5, with an additional subset sequenced in the v13 regions. The samples obtained from our nursing home cohort, however, were only sequenced in the v3 region. To create comparable datasets, we trimmed the Human Microbiome Project data to the v3 region using Cutadapt (Martin, 2011). This trimmed dataset was used for all further analyses except where explicitly mentioned. The resulting

sequence sets were processed as described subsequently here.

3.3.4 Sequence processing and analysis

Custom Perl scripts were developed to process the nursing home cohort sequences from Illumina sequencing. First, Cutadapt (Martin, 2011) was used to trim any reads surpassing the length of the v3 region. Resulting paired-end sequences were aligned with PANDAseq (Masella *et al.*, 2012). Operational taxonomic units (OTUs) were picked using AbundantOTU+ (Ye, 2011) with a clustering threshold of 97%. Taxonomy was assigned using the Ribosomal Database Project classifier (Wang *et al.*, 2007) against the Greengenes (February 4, 2011 release) reference database to the genus level (DeSantis *et al.*, 2006). A total of 2,429,732 reads (an average of 67,492.6 reads per sample with a range of 2,189-171,376) and 1,790 OTUs (an average of 291.4 OTUs per sample with a range of 76-460) were obtained from the 36 nursing home cohort samples. Samples were rarified to 2,000 sequences for all analyses of the nursing home cohort, and 1,000 sequences in the case of comparisons of the Human Microbiome Project and nursing home cohort samples. α- and β-diversity measures were calculated using the phyloseq R package (McMurdie and Holmes, 2013). Summaries of the relative abundances of taxonomies were computed using Quantitative Insights Into Microbial Ecology (Caporaso *et al.*, 2010c). Statistical significance of the relative abundance in the mid-aged adult and elderly anterior nares, the mid-aged adult and elderly oropharynx, and the elderly anterior nares and oropharynx were computed in Quantitative Insights Into Microbial Ecology (Caporaso *et al.*, 2010c) using ANOVA methods and deemed significant if the false discover rate–corrected P value was less than or equal to 0.05.

Multiple sets of sequence processing were performed. First, the data from the nursing home cohort was processed and analyzed alone; second, the nursing home cohort was processed in conjunction with the Human Microbiome Project data. Analyses of the v13 and v35 Human Microbiome Project data with our nursing home cohort samples were conducted to ensure that trimming did not affect results (see **Figure B.1**). The v3 regions from the nursing home cohort and Human Microbiome Project were used for the comparisons in this article (unless otherwise stated). Third, to ensure that the Human Microbiome Project adult data were comparable, adult oropharynx samples collected as part of another study (n = 32) were processed and analyzed together. These samples grouped with the Human Microbiome Project oropharynx swabs when taxonomic composition was examined (data not shown).

3.3.5 *S. pneumoniae* colonization in the nursing home cohort

In December 2009, 123 nursing home elderly (≥65 yr) were prospectively enrolled in a point prevalence study to determine the proportion of *S. pneumoniae* colonization in residents without respiratory symptoms in a nonoutbreak setting. All residents had a swab of their anterior nares (ESwabs) obtained by a trained research nurse. Residents were enrolled from four nursing homes in Ontario, Canada. Baseline characteristics for all participants were prospectively collected from patient charts.

Total DNA from the anterior nares swabs was obtained using automated easyMAG (bioMerieux, Marcy l'Etoile, France) and placed in elution buffer. The real-time PCR was performed according to the methods described by McAvin and colleagues (McAvin *et al.*, 2001) using the autolysin gene (*lytA*) as a target. Positive specimens

were confirmed by a second PCR reaction using primers specific for rpoB as described in (Fazeli *et al.*, 2013) and sequencing. This study was approved by McMaster University's Research Ethics Board. Written informed consent was obtained from all participants or their legally authorized guardian.

3.3.6 Generation of phylogenetic trees of *Streptococcus* species

For the examination of the OTUs assigned to the *Streptococcus* genus, the representative sequence from each OTU assigned to the genera and contributing to 1% or greater relative abundance in any of the four groupings (mid-aged adult anterior nares, elderly anterior nares, mid-aged adult oropharynx, elderly oropharynx) were collected. The v3 regions of all *Streptococcus* sequences from known species were extracted from the Human Oral Microbiome Database using Cutadapt (Martin, 2011) and custom Perl scripts (available from F.J.W.). These sequences were aligned using MUSCLE (Edgar, 2004) and phylogenies created with MrBayes (Ronquist and Huelsenbeck, 2003) using the generalized time reversible 1 gamma evolutionary model and run for 10 million generations.

3.4 Results

3.4.1 Sex, comorbidities, housing, and prior history of vaccination do not influence the microbial communities of the upper respiratory tract in the elderly

Samples of the anterior nares and oropharynx were collected from 18 individuals aged 68−96 years residing at four nursing homes in Ontario, Canada. This nursing home cohort was analyzed for its microbial content using 16S rRNA gene profiling, as described in Section 3.3. To examine whether variables such as sex, various comorbidities, nursing home residence, or vaccination schedules contributed to the microbial composition of the elderly, β-diversity measures on OTU composition were calculated using principal coordinate analyses with the weighted UniFrac metric (**Fig 3.1**). These data demonstrate that the composition of the nursing home cohort samples was not determined by any of the metadata variables collected, including sex, age, participant comorbidities, and vaccination schedules (**Fig 3.1**, **Table B.1**). The participant's nursing home residence did not associate with sample OTU composition, indicating that any differences in daily routine, diet, or residence geography did not contribute to the upper respiratory tract microbiota (**Fig 3.1a**). In addition, analysis of β-diversity showed no correlation of the upper respiratory tract microbiota with various comorbidities (e.g., chronic obstructive pulmonary disease, congestive heart failure, coronary artery disease, asthma, dementia), or vaccinations (e.g., annual influenza vaccine 2009 and 2010, pneumococcal polysaccharide vaccine, etc.) (**Fig 3.1b** and **Fig 3.1c**, **Table B.1**). In addition, there were no differences in microbiota composition between males and females (**Fig 3.1d**). Perhaps most interestingly, **Figure**

3.1e shows no clustering based on the biogeography of the sample (anterior nares or oropharynx). These results are surprising given that previous research has shown marked differences in these upper respiratory tract locales in mid-aged adults (Lemon *et al.*, 2010).

3.4.2 There is a distinct loss of topography in the upper respiratory tract microbiotas of mid-aged and elderly adults

The lack of distinct topographies of the nursing home cohort samples by sample biogeography (anterior nares or oropharynx) was unexpected. To determine how the microbial composition of the elderly related to that of mid-aged adults, we compared the nursing home cohort swabs to mid-aged adult anterior nares and oropharyngeal samples collected as part of the publicly available National Institutes of Health Human Microbiome Project. β-diversity measures conducted using weighted UniFrac demonstrate that distinct nasal and oropharyngeal communities observed in mid-aged adults are lost with age (**Fig 3.2a**). With exception of a few individuals, the anterior nares and oropharyngeal swabs collected from elderly participants were located between the two distinct clusters of mid-aged adult nasal and oropharyngeal samples. These observations were confirmed by similar clustering results when other β-diversity measures (unweighted UniFrac, Bray-Curtis) were employed (**Fig B.2**). It is interesting to note that there is a subset of samples from the elderly that associate closely with the mid-aged adult biogeographies, indicating that the changes in these locales are dependent on the individual, and may occur at different stages in the aging process. The microbial communities of the elderly nares and oropharynx overlap in

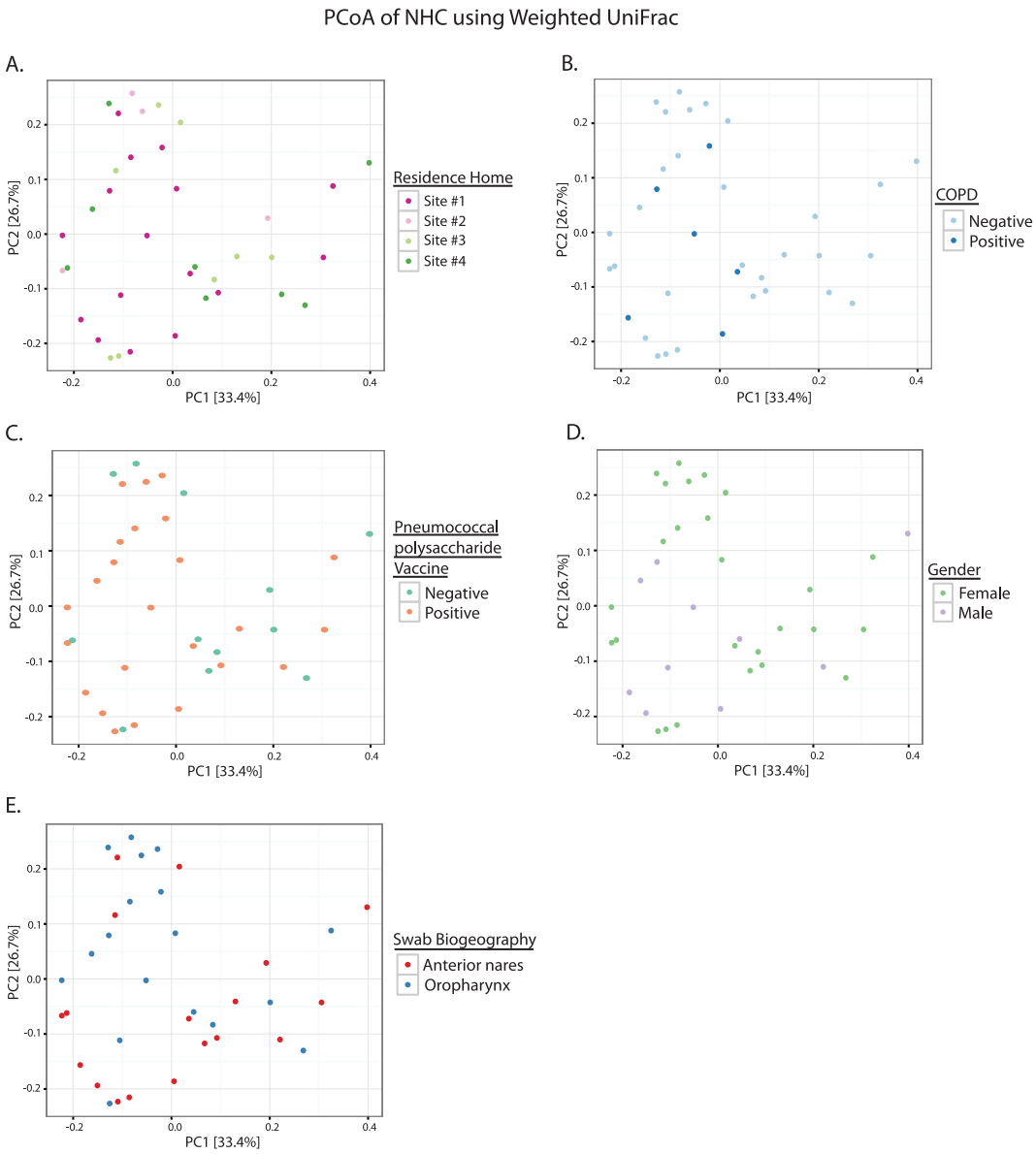


Figure 3.1: **Principal coordinate analyses (PCoA) of nursing home cohort using weighted UniFrac.** The microbial populations from samples of the anterior nares and oropharynx obtained from nursing home elderly do not associate with residence home (**A**), comorbidities (**B**), vaccination schedules (**C**), sex (**D**), or by swab geography (**E**). 16S rRNA gene profiling of the anterior nares and oropharynx were examined in relation to multiple metadata variables. No obvious association or grouping of similar phenotypic information was observed, including a lack of distinction between the biogeographies examined. (**B** and **C**) Representative examples of the lack of associations witnessed in relation to multiple comorbidities and vaccinations; the additional data that were analyzed are outlined in **Table B.1**. β-diversity measures were conducted using weighted UniFrac and visualized using PCoA. COPD = chronic obstructive pulmonary disease; PC = principal coordinate.

the principal coordinate analyses representation of β-diversity, which indicate that, in contrast to mid-aged adults, they are not composed of distinct communities (**Fig 3.2a**). However, microbial communities of the anterior nares and oropharynx of individual participants rarely cluster together (**Fig 3.2b**). This suggests that, although these nasal and oropharyngeal communities in the elderly are not distinct from each other, each individual remains heterogeneous in his or her nasal and oropharyngeal microbial composition.

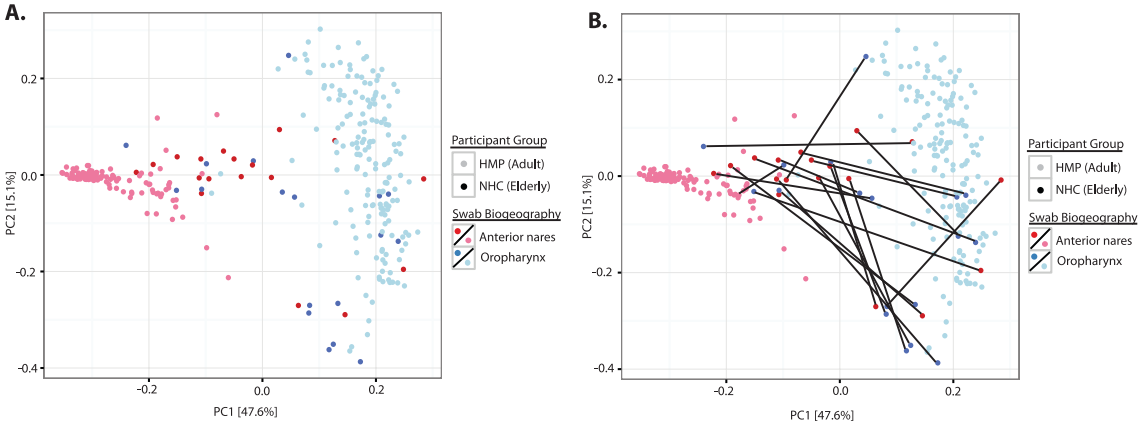


Figure 3.2: **The distinct topographies between the microbial communities of the anterior nares and oropharynx of adults are lost with age.** (**A**) β-diversity measures were used to compare upper respiratory tract biogeographies between adult and elderly individuals. The 16S rRNA gene profiles of the mid-aged adults were obtained from National Institutes of Healths Human Microbiome Project and were compared with samples from the elderly. These data suggest that there is little separation of nasal and oropharyngeal samples in the elderly, but that distinct differences between these populations exist in mid-age. (**B**) Even though there is little separation of these biogeographies in the elderly population as a whole, the microbiota of the anterior nares and oropharynx of a given individual (connected with lines) are rarely similar to each other. β-diversity measures were conducted using weighted UniFrac and visualized using principal coordinate analyses. HMP = Human Microbiome Project; NHC = nursing home cohort; PC = principal coordinate.

3.4.3 Genus-level taxonomic compositions of the nasal and oropharyngeal microbial communities

The distributions of microbial taxa were combined to produce an average representation of the taxonomic summaries for the anterior nares and oropharynx in both mid-aged and elderly adults. All taxa with abundance above 1.0% are displayed in **Figure 3.3**; a summary of all taxa present at a relative abundance below 1.0% can be found in **Figure B.3**. The oropharynx of mid-aged adults was dominated by *Streptococcus* (26.1%), *Prevotella* (14.1%), and *Veillonella* (8.9%). The elderly oropharynx was likewise dominated by these three genera; however, the relative abundance of *Streptococcus* increased to 44.0% of the total population in these samples. There was also a statistically significant increase in *Lactobacillus* and *Lactococcus* species in this population, in addition to 61 other statistical differences (**Fig 3.3**, **Table B.2 & B.3**).

The nasal microbial communities of the mid-aged and elderly adults are distinct. Although the mid-aged adults are dominated by *Corynebacterium* (28.5%), *Propionibacterium* (29%), and *Staphylococcus* (16.9%), the elderly populations are dominated instead by *Prevotella* (9.0%), *Veillonella* (4.7%), *Streptococcus* (29.6%), and by a lesser amount of *Staphylococcus* (8.2%). The taxonomic distributions of the elderly anterior nares were more similar to the mid-aged adult oropharynx than to the mid-aged adult anterior nares (**Fig 3.3**).

When the anterior nares and oropharyngeal samples from the elderly were compared with each other, there were no statistical differences between the biogeographies. This is striking given the distinct topographies of these locales in mid-aged adults, and indicates that, although mid-aged adults possess very different taxonomic

distributions in the anterior nares versus the oropharynx, the elderly are quite homogenous at these sites.

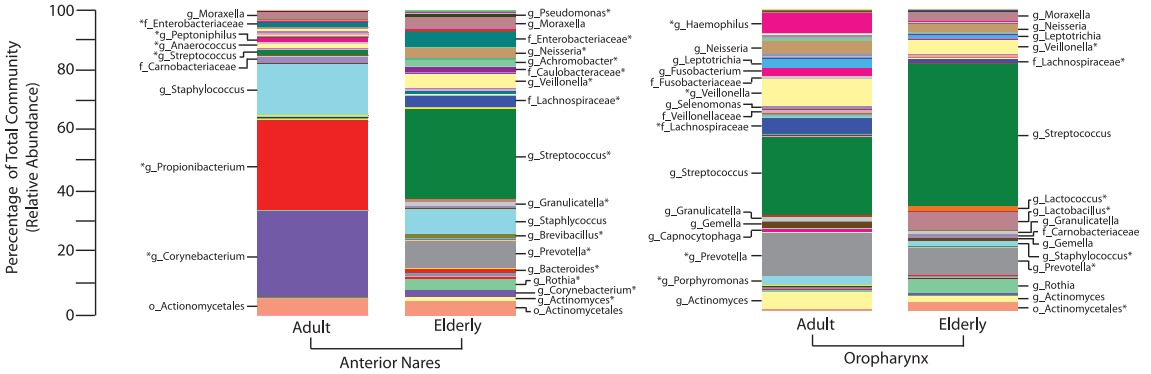


Figure 3.3: **Significant changes in the taxonomic composition of the upper respiratory tract microbial populations explain the loss of topography with age.** Taxonomic summaries of the anterior nares and oropharynx of the adult and elderly populations differ. Taxonomic information from participants in each group was summarized to assist in visual comparisons (see Section 3.3). Each bar represents the bacteria present in each group; all bacterial groups present at 1% or greater are labeled (see **Figure B.3** for lower abundant taxa). Asterisks mark those taxa that are significantly different (false discover rate-adjusted $P < 0.05$) between the adult and elderly populations in each biogeography. Taxa summaries are indicated as being resolved to the order (o), family (f), or genus (g) level where appropriate.

3.4.4 The OTU composition within the *Streptococcus* genera differs between the biogeographies of the mid-aged and elderly adults

One of the largest differences in species relative abundance between the four groups examined (anterior nares and oropharynx in mid-aged adults and elderly) was in the *Streptococcus* species. This genus contains approximately 55 human-associated species with a wide range of pathogenic phenotypes (Facklam, 2002). Species such as *S. pyogenes* and *S. agalactiae* are pathogenic, causing diseases such as impetigo, bacterial pharynitis, and neonatal sepsis, whereas others (e.g., *S. salivarius* and *S.*

oralis) often colonize, but rarely cause disease (reviewed in (Facklam, 2002)). The elderly are at increased risk of respiratory infection which may be attributed to an increased carriage of pathogenic streptococci. Consequently, we attempted to identify the species that are represented by the OTUs in this genus.

The relative abundance of each OTU assigned to the genus *Streptococcus* was examined and is displayed in **Figure 3.4a**. Two OTUs (OTU #1 and OTU #2) account for the majority of the relative abundance of the streptococci in each type of microbial community examined. In addition, across both biogeographies in both age groups, there were six OTUs that contributed 1% or greater of the relative abundance of the streptococci in any of the four groups. The mid-aged adult oropharynx possesses the most richness in *Streptococcus*, and it appears that this richness is diminished with age (**Fig 3.4a**). Interestingly, when α-diversity measures were conducted on these populations, the mid-aged adult oropharynx samples had a mean richness score well above that of the elderly oropharynx samples (**Fig B.4**), indicating that this loss in diversity may not be unique to the streptococci.

The representative sequences from each OTU were compared phylogenetically to the 16S rRNA gene v3 region of the streptococci in the Human Oral Microbiome Database (Chen *et al.*, 2010) (**Fig 3.4b**). Although species-level taxonomic identification with only the 200–bp v3 16S rRNA gene region is not possible, phylogenetic approaches allow us to define which group(s) the OTUs represent. Using this approach, OTU #1 was identified as belonging to the salivarius group (*S. salivarius, S. thermophilius, S. vestibularis*). There is an increase in the relative abundance of this OTU when the mid-aged and elderly adult oropharynx groups are compared. OTU #2, along with 55, 64, 91, and 318, clustered with a subclade within the mitis group

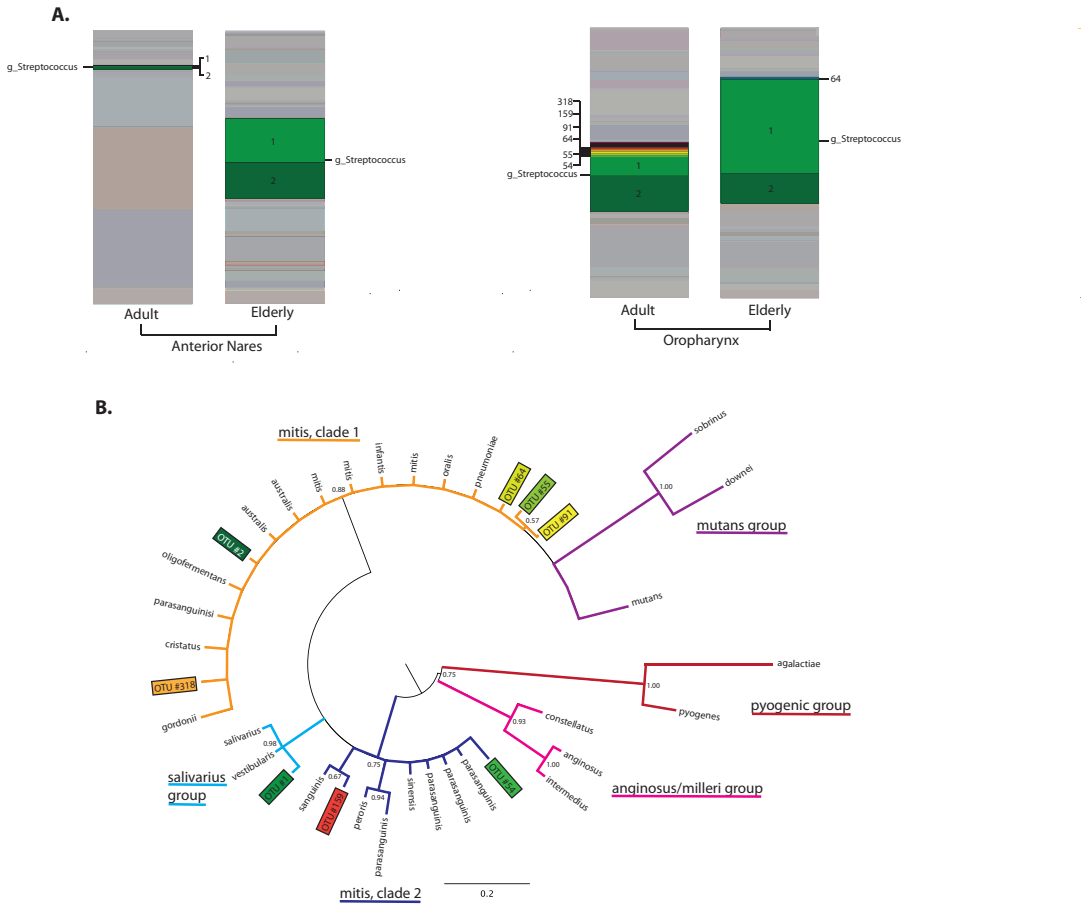


Figure 3.4: **The relative abundance of the *Streptococcus* differs depending on participant age and sample biogeography.** (**A**) When we examine the operational taxonomic unit (OTU) composition of the streptococci group, it is evident that a population of Streptococcus from OTU #1 and #2 contribute to the increased relative abundance of this genus in the anterior nares with age. In addition, an increase in OTU #1 accounts for most of the differences between the adult and elderly oropharynx locales. There is also a species richness within the Streptococcus of the adult oropharynx, which is lost in the elderly population. (**B**) Representative sequences from all OTUs with 1% or greater relative abundance in any group of samples (mid-aged adult or elderly anterior nares or oropharynx) that were analyzed phylogenetically in relation to the Streptococcus species in the Human Oral Microbiome Database (Chen *et al.*, 2010). The phylogeny identifies OTU #1 as belonging to the Streptococcus salivarius group; all other representative sequences are members of the Streptococcus mitis group. These results where verified using BLAST against National Center for Biotechnology Informations Reference rRNA database (**Table B.4**).

that includes *S. pneumoniae, S. mitis,* and *S. oralis* among others. Two OTUs (54 and 159) were identified as belonging to another subclade of the mitis group, which includes *S. sanguinis, S. parasanguinis, S. sinensis,* and *S. peroris.* It should be noted that there is no phylogenetic separation between mitis subclade 1 and the mutans group, meaning that the six OTUs assigned to this clade could represent members of the mutans group instead of the mitis group; however, additional BLAST searches of the representative sequences of these OTUs against the National Center for Biotechnology Information Reference rRNA database (Pruitt *et al.*, 2009) indicate that these sequences are most likely from the mitis group (**Table B.4**).

3.4.5 Evaluation of the presence of *S. pneumoniae* in the anterior nares of elderly nursing home residents

The lack of species-level identification of the highly diverse streptococci mean that the relatively high abundance of *Streptococcus* OTUs in the elderly nostril, compared with that of the mid-aged adult, cannot be distinguished as a high carriage of pathogenic or nonpathogenic species. Specifically, carriage of S. pneumoniae in the upper respiratory tract is a prerequisite to pneumococcal disease, to which the elderly are particularly susceptible (Bogaert *et al.*, 2004). To determine whether the increase in carriage of streptococcal species was due in part to increased carriage of *S. pneumoniae*, real-time PCR targeting *S. pneumoniae* was conducted on a separate cohort of nursing home elderly. DNA from the anterior nares were examined for *lytA*, an important virulence factor of *S. pneumoniae* (Berry *et al.*, 1989), which is often used to detect the presence of this species (McAvin *et al.*, 2001). Of the 123 participants involved in the first study, only 7 were positive for the target gene; of these, 4 were

confirmed with a secondary PCR reaction and sequencing (**Table B.5**). Of these four confirmed carriages of *S. pneumoniae*, three had received the pneumococcal vaccine in the last 5 years (**Table B.5**). Thus, it is likely that the increased abundance of OTU #1 in our nursing home cohort population does not represent an increase in *S. pneumoniae*, but instead of other mitis species.

3.5 Discussion

The complex interactions between the host and the microbial communities in which they reside are starting to be explored with improved sequencing technologies and computational approaches. However, these interactions change and evolve over our lifetimes. Babies are colonized within minutes of birth by the microbiota of their mothers (Johnson and Versalovic, 2012), and, as they grow into children, have upper respiratory tract microbial communities that fluctuate over time, and often include carriage of *Staphylococcus aureus, Haemophilus influenzae,* and *Moraxella catarrhalis* (Oh *et al.*, 2012; Bogaert *et al.*, 2011), which make children more susceptible to diseases, such as pneumonia, meningitis, and bacteraemia (Kwambana *et al.*, 2011). As children age, their immune responses and microbial communities stabilize into those characteristic of mid-aged adults (Oh *et al.*, 2012; van Benten *et al.*, 2005). These communities and a stable immune system keep upper respiratory tract infections in adults at bay. However, after the age of 65 years, respiratory infections become increasingly frequent (Mouton *et al.*, 2001). With age, immunosenescence causes an increase in proinflammatory markers and a decreased ability to handle immune stress (Franceschi *et al.*, 2000). Given the parallels in infection and dysregulated immune responses between children and the elderly, it is reasonable to hypothesize that the

microbiome of the elderly might become similarly disordered.

Swabs of the anterior nares and oropharynx of nursing home residents were collected and analyzed using 16S rRNA gene sequencing to determine the microbial composition of these biogeographies. Surprisingly, when compared using β-diversity measures, there was no grouping of samples by nursing home residence, comorbidities, vaccination, or between the anterior nares and oropharynx (**Fig 3.1**). This is in stark contrast to studies of the gut microbiome of the elderly, which demonstrated associations between residence, diet, and various comorbidities (Claesson *et al.*, 2016). This suggests that perhaps the upper respiratory tract and gut microbiota respond differently to environmental factors. The microbial composition of the anterior nares and oropharynx of the elderly were not distinct from each other, which was unexpected given the literature on the distinct topographies in mid-aged adult populations (Lemon *et al.*, 2010; Charlson *et al.*, 2011). When the taxonomic compositions of these locales in mid aged and elderly adults were compared, it was evident that these microbiota differ with age. The relative abundance of dominating organisms and, in the case of the anterior nares, the dominating organisms themselves, varied between age groups (**Fig 3.3**). The effects of age were most prominent in the anterior nares; in particular, this community in the elderly was more similar to the mid-aged adult oropharynx than to the mid-aged adult anterior nares, suggesting that the nasal community is lost with age. Interestingly, Charlson and colleagues (Charlson *et al.*, 2011) discovered that, in mid-aged adults, the nose is the only location in the respiratory tract that differs statistically in terms of its taxonomic composition. This distinct nasal microbial population is the interface between the respiratory tract and the environment; its loss in the nursing home elderly may account for increased

susceptibility to respiratory disease.

Because of the differences in the relative abundance of *Streptococcus* between the age groups at each biogeography, we compared the OTU composition within this genus. The streptococci consist of pathogenically diverse species, many of which are associated with respiratory infection (Facklam, 2002). By phylogenetically comparing the OTUs assigned to this genus to the *Streptococcus* species present in the Human Oral Microbiome Database (Chen *et al.*, 2010), we were able to narrow down the identification of each OTU to a particular group within the *Streptococcus* (**Fig 3.4**). We discovered that there was an increase in the relative abundance of OTU #1 (phylogenetically assigned to the salivarius group) between the mid-aged and elderly adult oropharynx. One of the first colonizers of the oral cavity of neonates (Rotimi *et al.*, 1985), *S. salivarius* has the ability to cause bacteraemia (Ruoff *et al.*, 1989), and meningitis in the immunocompromised (Laurila *et al.*, 1998). The increase in this organism may cause disorder to the upper respiratory tract population, contributing to an increased susceptibility of pathogen carriage in the elderly. In addition, there are numerous low-abundance streptococci OTUs in the mid-aged adult oropharynx that are not present in the elderly. It is possible that this loss of species richness opened up an environmental niche, allowing the *S. salivarius* group to flourish. OTU #2, assigned to the *Streptococcus mitis* group, was consistently abundant within the oropharynx of mid-aged and elderly adults, but was greatly increased in the anterior nares of the elderly compared with that of the mid-aged adult. The *S. mitis* group is a heterogeneous family that includes commensals and known pathogens, including *S. pneumoniae*. Further PCR analyses specific for *S. pneumoniae* indicate that the increase of streptococci in the anterior nares is likely not caused by an increase in

this species. This result was surprising given the surge of pneumonia and meningitis in the elderly (Crossley and Peterson, 1996); however, these results are in line with previous studies on the carriage rates of *S. pneumoniae* within the elderly (Ridda *et al.*, 2010; Flamaing *et al.*, 2010). Additional research has found that the elderly experience short episodes of carriage when compared with younger age groups (Ridda *et al.*, 2010) and that, on average, only a very small percentage ($\leq$10%) are colonized at any given time (Flamaing *et al.*, 2010). It is instead likely that this OTU represents non-pneumoniae members of the mitis group.

In this study, we compared publicly available data on the microbial composition of healthy mid-aged adult anterior nares and oropharyngeal samples to those of elderly residing in nursing homes. A limitation of this study is the inherent frailty and consequences of institutionalized living of these elderly individuals. We believe that the differences witnessed in these elderly participants when compared with healthy mid-aged adults are representative of the elderly population as a whole, given that the microbial communities of these individuals did not associate with the 12 comorbidities analyzed as part of this study (**Table B.1**); however, we suggest that further research conducted on community-dwelling elderly individuals should be completed, and compared with that of healthy adults and nursing home-dwelling individuals. Although we did not measure the impact of geographical diversity in this study (all individuals lived within 1,000 km^2), when data from an independent study of the oropharynx was sampled from mid-aged adults residing in southern Alberta and compared with those oropharyngeal sequences generated as part of the Human Microbiome Project (i.e., across the United States), we did not see any effects of geography (data not shown). Thus, we believe that the changes that we observed in this study were not influenced

by geographical location; however, further research is needed to confirm this.

We discovered that the distinct microbial topography of the mid-aged adult anterior nares and oropharynx is lost with age. In the elderly, the microbial population in the anterior nares appears to be displaced by that of the oropharynx, possibly explaining the differences in respiratory infection rates between these populations.

3.6 Acknowledgements

The authors wish to thank the individuals who participated in this study as well as the HMP for their contributions to open-source science.

Chapter 4

Longitudinal sampling of the lung microbiota in individuals with cystic fibrosis

Preface

Research presented as part of this chapter has been previously published as

Whelan FJ, Heirali AA, Rossi L, Rabin HR, Parkins MD, & Surette MG. (2017). Longitudinal sampling of the lung microbiota in individuals with cystic fibrosis. *PLOS ONE*, 12(3): e0172811. doi:10.1371/journal.pone.0172811

Author Contributions: FJW is the primary, first-author of this published manuscript. FJW, MDP, and MGS conceptualized and designed the project. Data curation of patient records was performed by HRR and MDP. FJW performed all data analysis, including the generation of all figures. Funding for the project was acquired by MDP and MGS. Laboratory preparation of the samples was conducted by AAH and LR. All authors contributed to the writing and editing of the manuscript.

The only alterations made to this publication were for thesis continuity and formatting. Supplemental material published as part of this manuscript is presented in **Appendix C**.

Title page and author list

Longitudinal sampling of the lung microbiota in individuals with cystic fibrosis

Fiona J. Whelan[1], Alya A. Heirali[2], Laura Rossi[1], Harvey R. Rabin[2,3], Michael D. Parkins[2,3], Michael G. Surette[1,2,4*]

[1]Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Canada, [2] Department of Microbiology, Immunology and Infectious Diseases, The University of Calgary, Calgary, Canada, [3] Department of Medicine, The University of Calgary, Calgary, Canada, [4] Department of Medicine, McMaster University, Hamilton, Canada

∗ To whom correspondence should be addressed:
surette@mcmaster.ca

4.1 Abstract

Cystic fibrosis (CF) manifests in the lungs resulting in chronic microbial infection. Most morbidity and mortality in CF is due to cycles of pulmonary exacerbations episodes of acute inflammation in response to the lung microbiome - which are difficult to prevent and treat because their cause is not well understood. We hypothesized that longitudinal analyses of the bacterial component of the CF lung microbiome may elucidate causative agents within this community for pulmonary exacerbations. In this study, 6 participants were sampled thrice-weekly for up to one year. During sampling, sputum, and data (antibiotic usage, spirometry, and symptom scores) were collected. Time points were categorized based on relation to exacerbation as Stable, Intermediate, and Treatment. Retrospectively, a subset of samples were interrogated via 16S rRNA gene sequencing. When samples were examined categorically, a significant difference between the lung microbiota in Stable, Intermediate, and Treatment samples was observed in a subset of participants. However, when samples were examined longitudinally, no correlations between microbial composition and collected data (antibiotic usage, spirometry, and symptom scores) were observed upon exacerbation onset. In this study, we identified no universal indicator within the lung microbiome of exacerbation onset but instead showed that changes to the CF lung microbiome occur outside of acute pulmonary episodes and are patient-specific.

4.2 Introduction

Cystic fibrosis (CF) is caused by mutations in the cystic fibrosis transmembrane conductance regulator (CFTR) gene (Kerem *et al.*, 1989; O'Sullivan and Freedman, 2009), which leads to impairments in pancreatic and liver function, and intestinal obstruction (Elborn, 2016; Andersen, 1938). However, it is the effect that this disease has on the lungs that has the greatest clinical burden. Repeated cycles of airway infection, mucous impaction, and bronchiectasis results in the majority of morbidity and mortality in the patient population (Elborn, 2016; Ferkol *et al.*, 2006). This chronic lung disease is progressive, manifesting as persistent lung function decline and diminishing quality of life (Corey *et al.*, 1997; Sanders *et al.*, 2011).

Pulmonary exacerbations are respiratory perturbations characterized by increased respiratory symptomatology, systemic inflammation, fatigue, and weight loss (Fuchs *et al.*, 1994), symptoms which are potentially rescued by airway clearance and antimicrobial therapy directed against chronically infecting pathogens (Ferkol *et al.*, 2006; Lam *et al.*, 2015; Flume *et al.*, 2009; Döring *et al.*, 2012). These events are critical in CF as they cause permanent loss of lung function; however, the mechanisms underlying these events remain largely unknown. Exacerbations have been associated with viral infections (Goss and Burns, 2007; Hiatt *et al.*, 1999) as well as changes in density of primary bacterial pathogens within the lungs (Goss and Burns, 2007; Carmody *et al.*, 2013) perhaps due to a clonal expansion of pre-existing strains (Aaron *et al.*, 2004). However, the true cause of pulmonary exacerbations is likely multi-factorial in nature, including interactions between the immune system, lung microbiota, airway physiology, and the environment (Ferkol *et al.*, 2006; Goss and Burns, 2007), complicating the understanding and treatment of these events.

The CF airways have long been known to harbor certain primary pathogens such as *Pseudomonas aeruginosa, Burkholderia cepacia* complex, and *Staphylococcus aureus* (LiPuma, 2010). More recently, as sequencing technologies and laboratory culture techniques advance, it has become appreciated that there are many additional bacterial members of the CF lung community which have the propensity to contribute to disease. For example, *Stentrophomonas maltophilia, Achromobacter* spp., *Mycobacterium abscessus*, Methicillin-resistant *Staphyloccocus aureus* (MRSA), and the *Streptococcus* Anginosus/Milleri group have been described as emerging CF pathogens (LiPuma, 2010; Sibley *et al.*, 2008, 2010b; Whelan and Surette, 2015; Surette, 2014). Similarly, other non-bacterial members of the CF lung microbiome have been implicated in worsened prognosis such as the fungus *Aspergillus fumigatus* (LiPuma, 2010; Surette, 2014).

To date, many studies of the CF lung microbial population, or microbiome, include comparisons of sputum samples collected during pulmonary exacerbation and clinical stability (for e.g. (Coburn *et al.*, 2015; Carmody *et al.*, 2013)). While these sampling methods can be very informative, they cannot determine daily dynamics of the CF lung microbiome during exacerbation onset. There are two notable exceptions; Carmody *et al.* collected daily sputum samples from 4 participants over a 25-day period which included the onset of pulmonary exacerbation (Carmody *et al.*, 2015). In this study, the authors identified changes in the CF microbiome at exacerbation onset in a subset of participants by examining the beta diversity dissimilarity between longitudinal bacterial communities (Carmody *et al.*, 2015). Second, Cuthbertson *et al.* studied 10 CF patients at baseline, 30 days prior to treatment, treatment for exacerbation, 30 days post treatment, and post-exacerbation baseline (Cuthbertson *et al.*, 2015).

The authors determined that the core microbiota were resistant to exacerbation and associated antimicrobial treatments (Cuthbertson *et al.*, 2015).

In this study, we expand on the above by examining relative changes to the CF lung bacterial community over the course of one year in 6 participants in the context of clinical status (exacerbation treatment versus stability), changes in participant reported symptom scores and spirometry values, and antibiotic treatments. We discovered no consistent indicator of exacerbation onset and instead confirm the patient-specific nature of the CF lung microbiome.

4.3 Methods

4.3.1 Participant recruitment and sputum collection

From July to October of 2012, 6 knowledgeable and compliant cystic fibrosis (CF) patients were recruited for this study from the Southern Alberta Adult Cystic Fibrosis Clinic. The median age of participants was 32.5 (IQR 26-36), and all were homozygous for the F508del mutation except for one who was a compound heterozygote, F508del/621+1G-T (**Table 4.1**). Median lung function for participants was 1.72L (IQR 1.55L-2.66L), 66.9% predicted (IQR 58.3-85.0). Participants self-collected sputum samples 3 times a week (Monday, Wednesday, and Friday) into clinical laboratory collection jars (which were then immediately stored in their home freezers). During periods of absence from home (e.g. holidays/work trips) some samples were omitted. Participants self-reported data including symptom scores adapted from Jarad *et al.* Jarad and Sequeiros (2012). Symptoms with respect to Cough, Sputum Production, Shortness of Breath, Wheezing, Nasal Irritation, Throat Irritation, Fatigue, and

Appetite were independently scored relative to an individual's norm/baseline (=0) with increased symptomatology scored as 1=mild, 2=moderate, or 3=severe deterioration. Additionally, study participants were outfitted with PIKO-6 (nSpire Health; Longmont, CO) home spirometers to measure spirometry. Prior to enrolment, all participants were trained by a study investigator in the use of the PIKO-6 device. Participants were taught to perform expiratory maneuvers three times and record each value. Values used represent the best of each three attempts. Values collected at enrolment were correlated with complete pulmonary function testing performed during the clinic visit. Lung function values were reported as forced expiratory volume in one (FEV1) second. Any antibiotics, including those for chronic suppression of lung disease and acute management of pulmonary exacerbation, were similarly recorded. All collected data and records of antibiotic usage were made available to the study authors. Ethical approval for this study was given by the Calgary Health Region Ethics Board (REB-24123). At the enrolment visit, each patient provided written informed consent (with an REB approved document) after detailed discussion with research/clinic staff regarding what the study entailed. Of the 6 participants who began the study, 3 completed the full 1-year term with the remaining 3 participants dropping out of the study due to poor health or non-compliance (**Table C.1**); however, all 6 contributed serial samples and are included in subsequent analyses.

At the end of the study, participants returned a study log which included the metadata and their stored sputum using -20°C freezer packs and insulated transport bags to ensure samples were kept frozen. Upon receipt, samples were immediately transferred and stored at -80°C.

Table 4.1: **Clinical and methodological information about the study participants (* = did not complete study).**

Participant	Age at Study Onset	CFTR Mutation	Clinically cultured & treated primary CF pathogen(s)	% predicted lung function (FEV1/ FVC) at study onset	# of Exacerbations
A*	36	F508del/F508del	*P.aeruginosa, M.abscessus*	54.0	1
B	38	F508del/F508del	*P.aeruginosa*	73.5	1
C	26	F508del/F508del	*P.aeruginosa, S. agalactiae*	105.0	0
D*	36	F508del/F508del	*P.aeruginosa, Streptococcus* Anginosus group	58.3	1
E	23	F508del/ 621+1G-T	*P.aeruginosa*	60.2	4
F*	29	F508del/F508del	*P.aeruginosa, S.aureus, Cupriavidus.sp*	85.0	1

Samples were assigned into one of the three categories based on the time of collection: Treatment if the sample was collected during a pulmonary exacerbation (as defined by Fuchs *et al.* (Fuchs *et al.*, 1994)) and after any intravenous antibiotics were administered; Intermediate if the sample was collected in the month prior to or following a pulmonary exacerbation; Stable otherwise. At the end of the study, we retrospectively chose a subset of samples for marker gene analysis. Where possible, we chose a subset that included tri-weekly samples from the Treatment stage, weekly samples during Intermediate stages, and monthly samples during Stable periods. Using this schema, 121 of the 508 available samples were chosen for 16S rRNA gene sequencing (**Table C.1**).

4.3.2 Clinical microbiology

Standard clinical microbiology was performed during regular clinic visits as has been previously described (Sibley *et al.*, 2010b; Lam *et al.*, 2015). Quantitative analysis of sputum was carried out by plating on Columbia blood agar (CBA), chocolate

agar (CHOC), MacConkey agar (MAC), mannitol salt agar (MSA), and oxidation-fermentation polymyxin bacitracin lactose agar (OFPBL). These solid media plates were incubated at 35°C, 5% CO_2 for 2 days with the following exceptions: OFPBL was incubated at 30°C; CHOC which was incubated anaerobically.

4.3.3 DNA isolation and Illumina sequencing

Genomic DNA was isolated as previously described (Whelan *et al.*, 2014; Bartram *et al.*, 2011). Methods of genomic DNA extraction differed from (Whelan *et al.*, 2014) only in that the starting material was 300μl of sputum which had been homogenized by repeated passage through a 18 gauge needle and syringe. Barcoded universal primers adapted from (Bartram *et al.*, 2011) were used to amplify the variable 3 region of the 16S rRNA gene. The PCR reaction consisted of 5pmol of each primer, 50ng template DNA, 200μM dNTPs, 1.5mM $MgCl_2$, 4mg/mL BSA, 1x reaction buffer, and 1 U Taq polymerase. The PCR protocol was as follows: 94°C for 5 minutes, followed by 30 cycles of 94°C for 30 seconds, 50°C for 30 seconds, and 72°C for 30 seconds, with a final 72°C for 7 minutes. Presence of a PCR product was verified by electrophoresis (2% agarose gel). PCR products were normalized for quantity using the SequalPrep Normalization kit (ThermoFisher #A10510-01) and sequenced using the Illumina MiSeq platform using 2x250 paired-end reads.

4.3.4 16S rRNA sequence processing and analysis

The resulting sequencing data were processed using a custom in-house pipeline as in (Whelan *et al.*, 2014) with some modifications (**Fig C.1**). Briefly, primers and/or

read-through of the variable 3 region was trimmed using cutadapt (Martin, 2011), low-quality reads were culled using sickle with a quality threshold of 30 (https://github.com/najoshi/sickle), and chimeras were removed using USEARCH as written into QIIME (Edgar, 2010; Caporaso *et al.*, 2010c). Operational taxonomic units (OTUs) were generated using AbundantOTU+ (Ye, 2011) and each was given a taxonomic assignment using the RDP Classifier (Wang *et al.*, 2007) against the Greengenes reference database (February 4th 2011 release) (DeSantis *et al.*, 2006). OTU tables were generated via QIIME (Caporaso *et al.*, 2010c). Any OTU consisting of only one read across the dataset (i.e. singleton) was removed. After processing, there was a mean of 105,884 reads per sample (range: 33,940-215,072) and a mean of 216 OTUs per sample (range: 73-491). The 16S rRNA gene sequencing data and clinical metadata that make up this dataset are available via NCBIs Short Read Archive, (BioProject PRJNA360332).

All analyses of the resulting OTU table were performed in R (R Core Team 2016) using packages phyloseq (McMurdie and Holmes, 2013) and vegan for beta diversity calculations, vegan for tests of community-wide significance (i.e. PERMANOVA), and pheatmap to generate heatmap figures. A UPGMA representation of the Bray-Curtis dissimilarity between samples was generated using QIIME. Phylogenetic representations of participant's OTU diversity were generated by trimming the reference phylogeny provided with the 2011 release of Greengenes to those taxa which matched taxonomic assignments of each participant's OTUs. Visual representations of these phylogenies were created using graphlan (Asnicar *et al.*, 2015). Correlations of core OTUs, as defined by any OTU present in all samples from a particular participant with a sum of >1.0% over the study period, and key collected data (symptom scores,

FEV1, antibiotic usage, alpha, and beta diversity) were calculated using eLSA (Xia *et al.*, 2011) and were considered significant if p-values < 0.05 and q-values < 0.05, and the length of the observed correlation spanned the full dataset.

4.4 Results

4.4.1 Participant information and samples collected

In total, 6 individuals took part in this study. All 6 participants were chronically colonized with *Pseudomonas aeruginosa*; additionally, 4 of the 6 participants were chronically infected with additional organisms being targeted with antibiotic therapy: *Mycobacterium abscessus, Streptococcus agalactiae, Staphylococcus aureus, Cupriavidus* sp., or *Streptococcus* Anginosus/Milleri group (**Table 4.1**). *Streptococcus* Anginosus/Milleri group members have been previously reported as common CF pathogens in this clinic (Sibley *et al.*, 2008). These individuals experienced a total of 8 exacerbations (**Table 4.1**).

Of the 6 participants who began the study, 3 completed the year-long term with an average of 150 samples over the study period (**Fig 4.1, Table C.1**). The remaining participants dropped out of the study after an average of 116 days, and 20 samples per patient (**Table C.1, Fig 4.1**). A subset of this collection were retrospectively chosen for 16S rRNA gene sequencing with a focus on tri-weekly Treatment samples, weekly Intermediate samples, and monthly Stable samples (**Fig 4.1**).

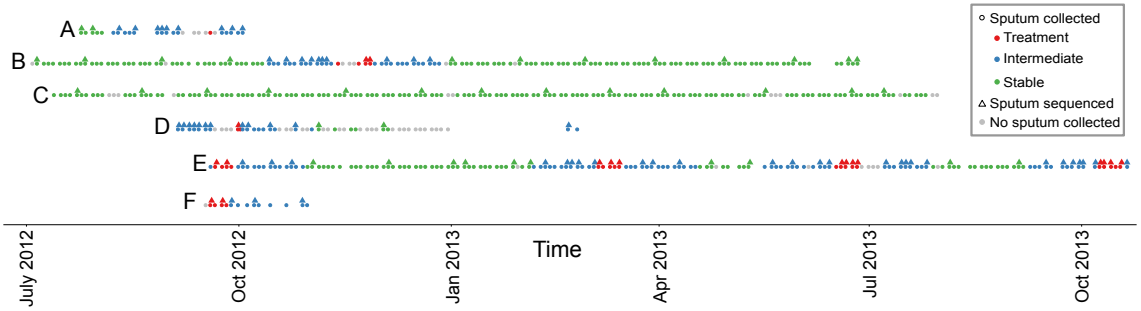


Figure 4.1: **Outline of sputum collection and samples chosen for sequencing.** Participants self-collected sputum 3 times a week while simultaneously recording clinical symptoms. On occasion, sputum could not or was not collected yet participant information was recorded (grey dots). Samples were retroactively chosen for microbiome analysis based on the sample type, aiming to follow Treatment more closely (1 sample/per 2-3 days) then Intermediate (1 sample per 1 week) and Stable (1 sample per 1 month) samples. All but one participant, C, experienced an exacerbation during the study period. Exact dates and length of sample collection for each participant is provided in **Table C.1**.

4.4.2 The cystic fibrosis lung microbiome is patient-specific

First, we aimed to examine the study-wide diversity amongst samples at the community level. Using the Bray-Curtis dissimilarity metric, which takes the relative abundance of individual OTUs into account, it was shown that the lung microbiota was significantly different between participants (PERMANOVA, p=0.001). This result is visualized using a Principal Coordinates Analysis (**Fig 4.2a**). In a few cases, such as between Participant B and C, these participant-specific clusters overlap, indicating shared elements of their microbial composition. This is further examined via a genus-biplot of the PcoA (**Fig C.2**) which indicates that *Pseudomonas* contributes to the separation of samples from Participants B and C; similarly, *Staphylococcus* and *Cupriavidus* separates samples from Participant F, and *Fusobacterium* separates samples from Participant D. Further, a UPGMA phylogeny of the Bray-Curtis dissimilarity between samples shows almost perfect inter-participant separation (**Fig 4.2b**).

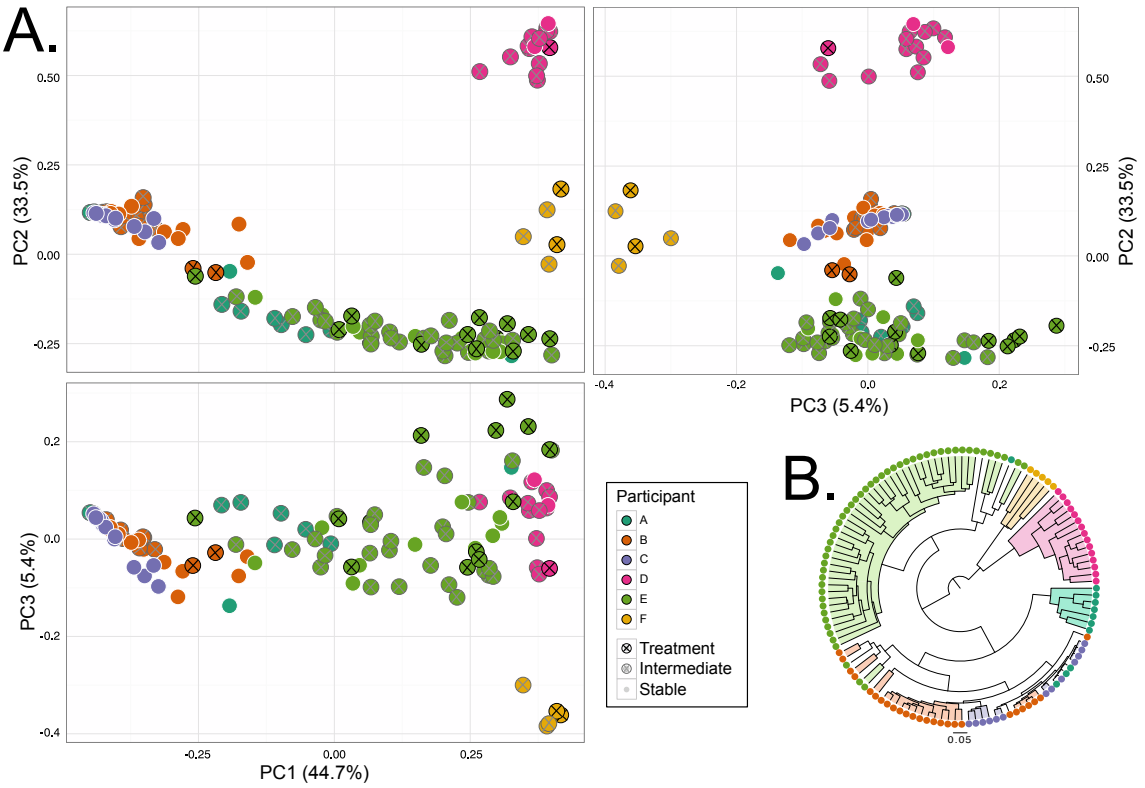


Figure 4.2: **The CF lung microbiome is distinguished by individual. A.** PCoA plots of all participants illustrate the clustering of participant samples, indicated as significant by PERMANOVA (p=0.001). Health state within participants, as defined as Stable, Intermediate (<1 month pre- or post-Treatment), and Treatment was significant (PERMANOVA, p=0.016), but was highly confounded by the participant (p=0.042 of Participant:Health interaction term). **B.** UPGMA phylogeny depicting the Bray-Curtis dissimilarity between samples. It is apparent that the principle driver of similarity between samples are inter-individual microbial lung composition due to the almost complete separation of participant samples.

Additionally, we investigated the effect of sample type (Treatment, Intermediate, Stable) on the microbiota at the community level by PERMANOVA (p=0.016). Although sample type was found to have a significant effect on microbial composition, this result was confounded by the participant (p=0.042 of the Participant:Health interaction term). This indicates that the composition of the microbiome is influenced more by the individual then by the sample type as has been previously shown (for e.g. Coburn *et al.* (2015); Kramer *et al.* (2015)).

4.4.3 Exacerbation does not consistently associate with community-wide changes to the microbiome

Next, we examined each participant's microbiota independently. Taxonomic summaries were used to visualize changes in the microbiota over the course of the study and in relation to the health state of the individual (**Fig 4.3a**). These community-wide profiles display unique communities in each individual, corresponding to the results in **Figure 4.2**. For example, while the microbial communities of participants A, B, and C are dominated by *Pseudomonas*, participants D, E, and F have more diverse communities consisting of dominant organisms such as *Prevotella, Streptococcus*, and *Fusobacterium* (**Fig 4.3a**). These communities, generated using 16S rRNA gene sequencing, mirror the selective culturing performed by the clinical microbiology laboratory associated with the clinic (**Table 4.1**); however, greater diversity is evident via 16S rRNA gene sequencing approaches.

During the study period, all participants except for C experienced a pulmonary exacerbation (**Fig 4.3a, red triangles**). Unfortunately, no samples were obtained from participant A during a pulmonary exacerbation that occurred between samples

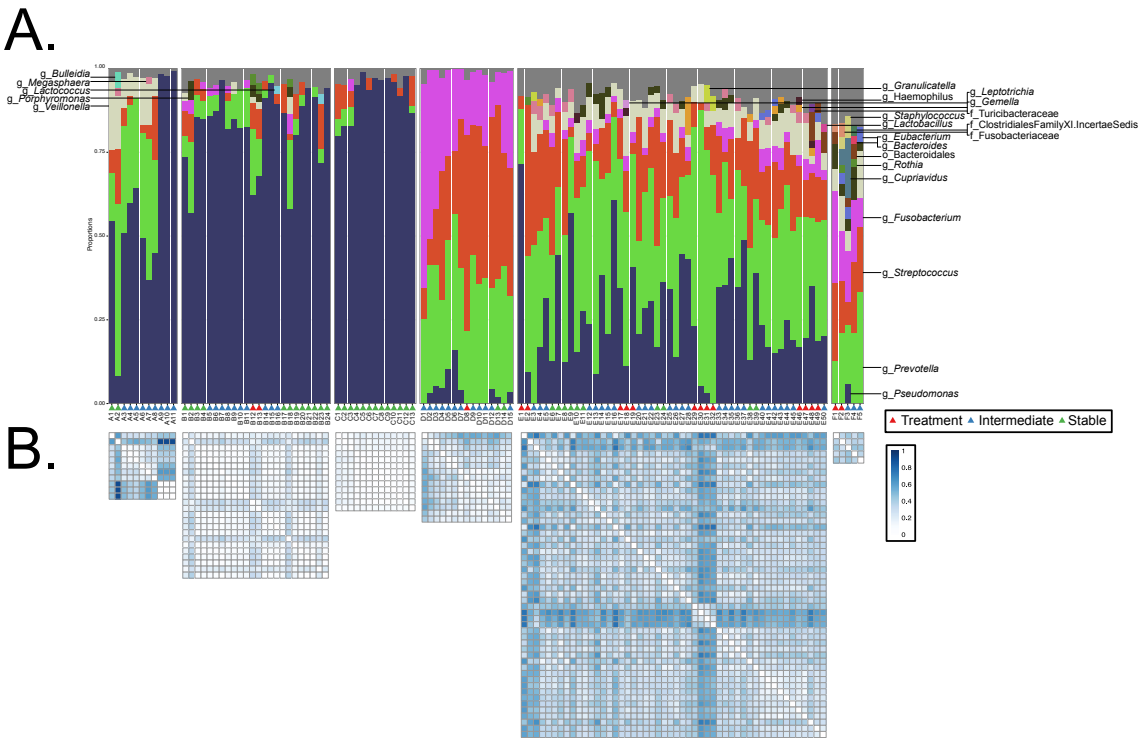


Figure 4.3: **The effects of exacerbation on the lung microbiome are not consistently seen at the community level. A.** Taxonomic summaries of all samples sequenced. These summaries indicate that changes to the lung microbiome upon exacerbation are not often obvious when examining the community-wide taxa composition. Taxa present at <2% are summarized in the grey bar. Participant E experienced 4 exacerbations during the study period which are indicated with black lines. **B.** Heatmaps indicate the Bray-Curtis dissimilarity between each sample. Here, we can see that samples taken during some exacerbations are more dissimilar to those collected during stability; however, this is not true for every exacerbation. These observations are qualified by statistical measures (**Table C.2**) and were independent of FEV1 (**Table C.3**).

A8 and A9. Visually, we observe from these taxa summaries that there are sometimes, but not always, observable changes in the lung microbiota preceding, during, or following pulmonary exacerbations.

To quantify these observations, the Bray-Curtis dissimilarity between samples (**Fig 4.3b**) and statistical measurements between categories were calculated (**Table C.2**). These metrics indicate that there are statistically significant changes in the lung microbiota between non-Treatment (Intermediate and Stable) and Treatment time points in 2 of the 4 participants (**Table C.2**; participants A and C were omitted due to no Treatment samples). These community-wide changes are seen between participant B's Intermediate and Treatment time points (p=0.045) as well as in participant E (Stable vs. Treatment, p=0.022; Intermediate vs. Treatment, p=0.009). However, results from participants D and F indicate no statistically significant changes to the microbiome with Treatment (**Table C.2**). Importantly, in participant E who had 4 exacerbations in the study period, only 1 of the 4 was accompanied with statistical changes to the microbiota (**Table C.2**). None of these observed alterations in the microbiota were accompanied with statistically significant changes to FEV1 (**Fig C.3, Table C.3**).

Further, there are observable disturbances to the lung microbiota within treatment categories. **Figure 4.3b** indicates changes in the lung community between a number of sequentially collected samples, taken at least 1 month outside of any exacerbation. Examples include changes in Bray-Curtis dissimilarity scores between Samples A1 and A2 as well was Samples B2 and B18 when compared to other Stable time points. Together, these findings indicate that some but not all exacerbations (2 of 7) result in or are preceded by a discrete, measurable change in the microbiome and that

observable shifts in these communities also occur independent of exacerbation onset.

4.4.4 Exacerbation is not linked with changes in within-sample diversity

Each sample within this study was examined independently to determine the within-sample diversity by calculating Shannons diversity index, an alpha diversity metric measuring both richness and evenness. Previous research has reported a decrease in alpha diversity with declining lung function and age (Cox *et al.*, 2010; Goddard *et al.*, 2012), affected by antibiotic therapy (Zhao *et al.*, 2012) and exacerbation treatment (Fodor *et al.*, 2012; Smith *et al.*, 2014); increased diversity of the lung microbiota is associated with stable lung function (Filkins *et al.*, 2012). In this study, differences in Shannon diversity were measured between Treatment, Intermediate, and Stable samples in each individual (**Fig 4.4 & C.4**). While a significant increase in diversity was observed between Intermediate and Treatment samples in Participant B (**Fig C.4**), the majority of samples showed no significant differences between Treatment, Intermediate, and Stable samples (**Fig 4.4 & C.4**).

Furthermore, Shannons diversity index was patient-specific (**Fig 4.4**). A range of values (0.077-3.345) were observed across participants (**Fig 4.4**). Interestingly, the participant with the lowest mean diversity score was the only participant who did not experience an exacerbation during the study period (**Fig 4.4, participant C, purple line**) and who maintained the highest FEV1 over the course of the study (**Table 4.1, Fig 4.5-4.6**). Although this sample size is small, these findings indicate that the use of alpha diversity metrics to assess the relative health status and stability of the lung microbiota may be complex at the patient-level even though low alpha

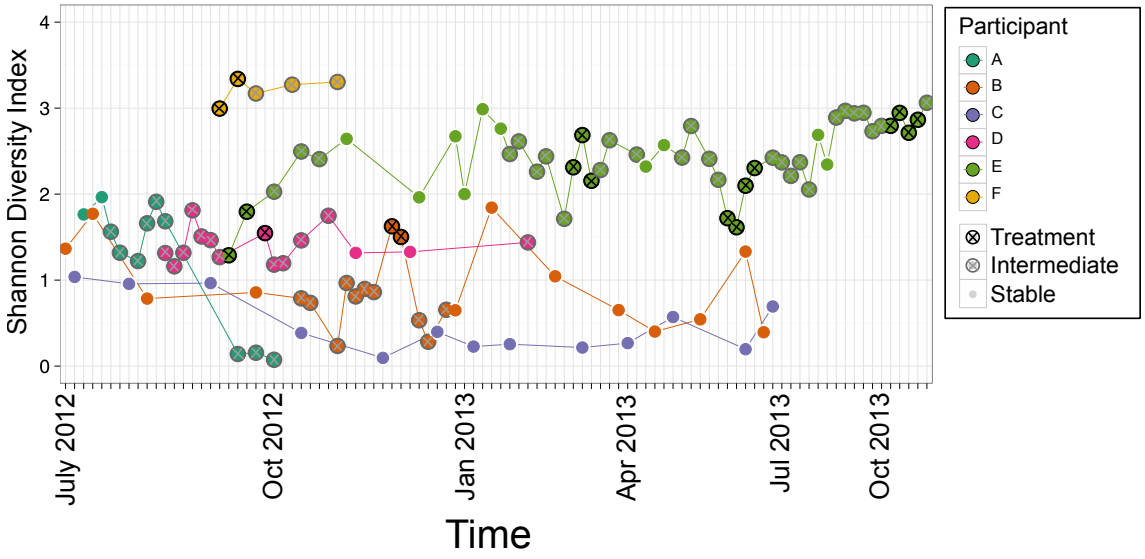


Figure 4.4: **Diversity within the lung community does not consistently decrease with exacerbation.** A longitudinal representation of the evenness and richness of the CF lung microbiota across study participants indicates patient-specific levels of within-patient diversity.

diversity has been associated with poor lung function at the population-level.

4.4.5 Longitudinal dynamics of the cystic fibrosis lung microbiota

To further understand elements of the patient-specific dynamics of the CF lung microbiota, we focused on 2 participants who completed the full study period. We chose participants C and E because they represented the individuals who had the least (n=0, C) and most (n=4, E) observed exacerbations.

As has been shown above, participant C demonstrated fairly uniform alpha and beta diversity across the study period (**Fig 4.3b & Fig 4.4**). This individual was on alternating 4-week tobramycin inhalation powder (TIP) therapy throughout the year, and had a consistent FEV1 in the range of 2.2-2.69L (**Fig 4.5a**). Overall, this individual's symptom scores were low (i.e. close to baseline), although there were periods

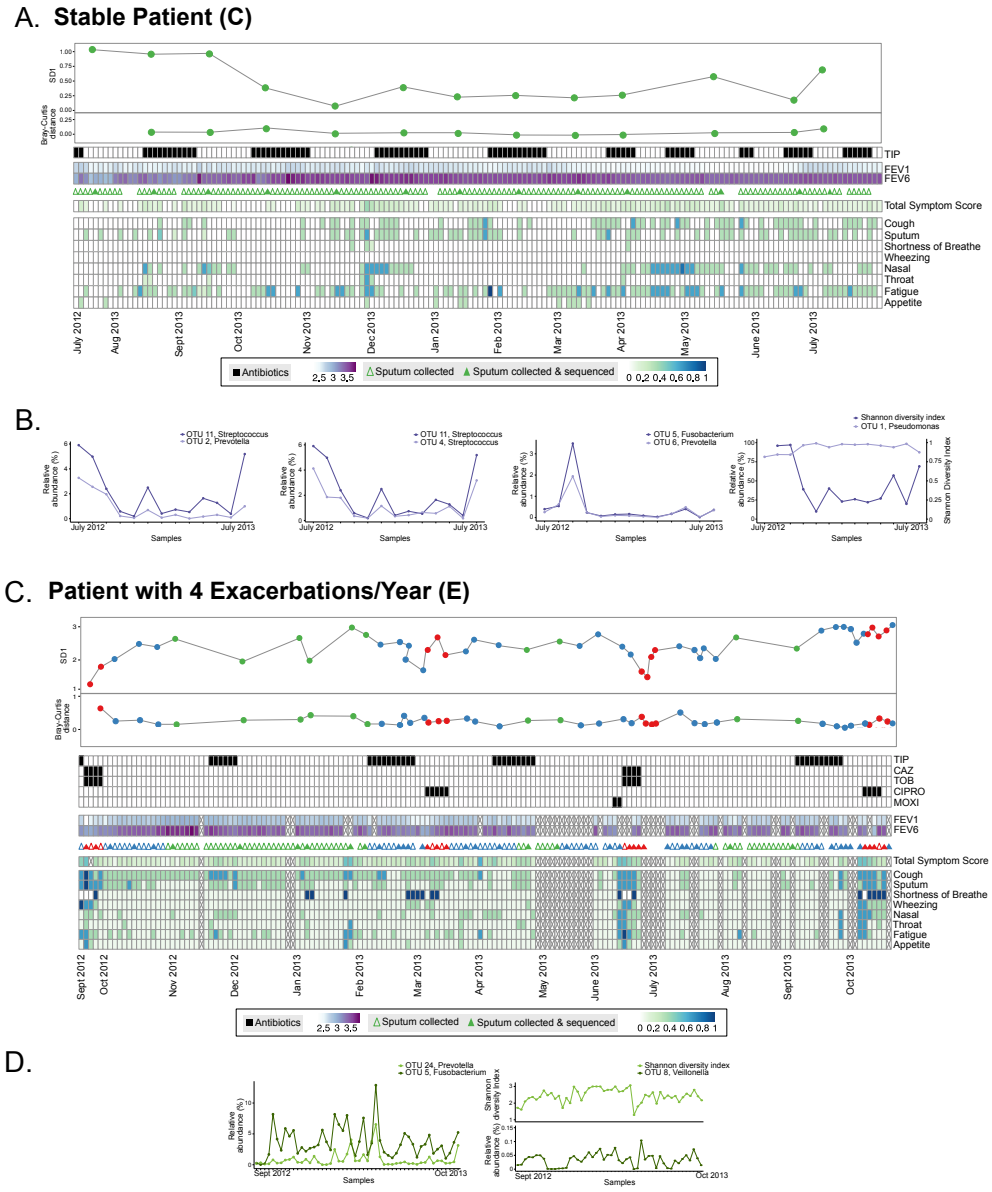


Figure 4.5: **Longitudinal Dynamics of two select participants (C and E).** Two participants who were the outliers in terms of the number of pulmonary exacerbations experienced over the course of the study period were chosen for closer examination. **A.** Sample collection for participant C is shown in relation to, antibiotic use, FEV1, and symptom scores. **B.** Correlations between collected data, diversity metrics, and OTU relative abundance were calculated and significant correlations were reported (**Table C.4**); a subset of these significant correlations are plotted. **C.** Sample collection for participant E in relation to antibiotic use, FEV1, and symptom scores. **D.** Correlations between these collected data and the OTUs present within the microbiome were calculated and significant correlations were reported (**Table C.5**); a subset of these significant correlations are plotted.

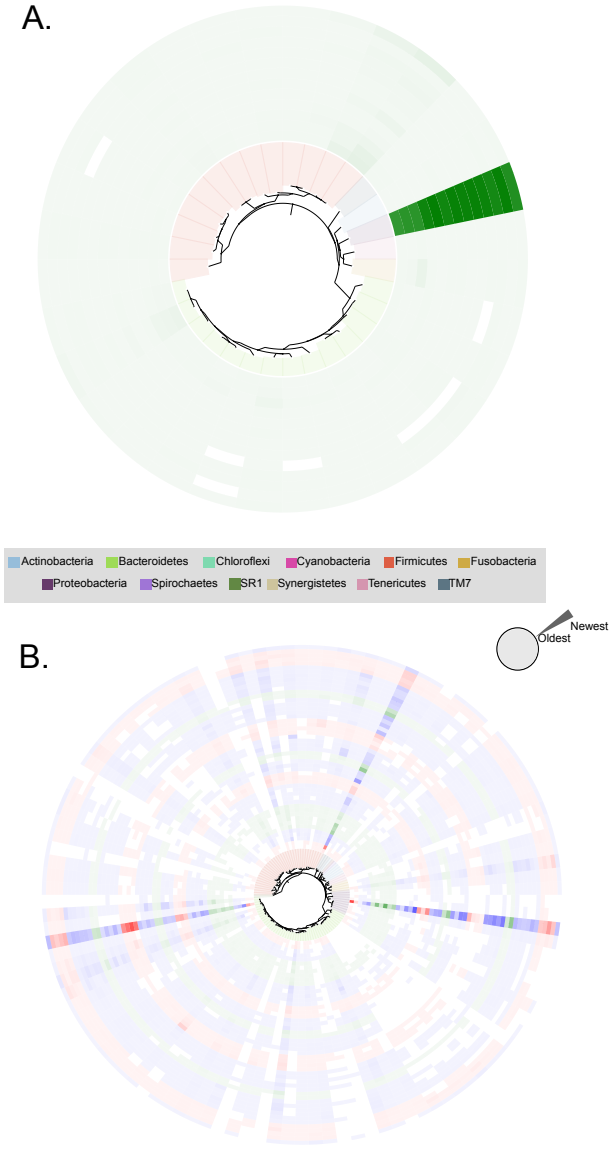


Figure 4.6: **Examples of stability and variability in the CF lung microbial communities of two select participants (C and E). A.** Visualization of the stability of participant C's lung microbial community over the study period. Each OTU is presented as a terminal node on the phylogeny; its presence in each sample evaluated using 16S rRNA gene sequencing is shown extending outwardly from the inner phylogeny in chronological order. The density of the colour indicates the relative abundance of the OTU; when the OTU is not identified, the space is left blank. **B.** Participant E, who experienced 4 exacerbations over the course of the year, has a much more variable lung microbiota than participant C. Similar to Fig 5c, OTUs are represented as nodes in the phylogeny whose relative abundance is indicated with varying colour density. Rings in the phylogeny are coloured to indicate the sample type (Treatment red, Intermediate blue, Stable green). Density of the colour indicates relative abundance of the OTU and time periods are coloured according to the health state.

of increased sinus congestion and fatigue during the study period. Correlations were calculated using eLSA between all core OTUs (sum > 1.0% relative abundance across all samples from participant), collected data (antibiotic use, FEV1, symptom scores), and diversity metrics (Shannon diversity index, Bray-Curtis dissimilarity scores) (**Fig 4.5b, Table C.4**). None of the collected data correlated with individual components of the microbiota; of the diversity metrics tested, Shannon diversity was negatively correlated with OTU 1 (**Fig 4.5b**). Instead, correlating OTUs within the microbiome were observed (**Fig 4.5b, Table C.4**). For example, *Prevotella* OTU 2 was positively correlated with *Streptococcus* OTU 11 (**Fig 4.5b**); additionally, OTU 11 was correlated with another *Streptococcus* (OTU 4). Further, *Prevotella* (OTU 8) was positively correlated with *Fusobacterium* OTU 5. However, all of these correlations were observed amongst OTUs with low (<10%) relative abundance. When the relative abundance of each OTU was examined longitudinally, we observed a remarkably stable lung microbial community dominated by a single *Pseudomonas* OTU (**Fig 4.6a**). These results suggest a community within the lung whose composition is highly dependent on its microbial membership, but less on external factors such as antibiotic use.

However, when we examine participant E, we see a very different picture of CF lung disease. Participant E was also on alternating TIP therapy over the study period; however, this treatment was supplemented with further antibiotics upon exacerbation onset including ceftazidime (CAZ), tobramycin (TOB), ciprofloxacin (CIPRO), and moxifloxacin (MOXI) (**Fig 4.5c**). FEV1 decreased over the study period and was within a range of 1.13-2.11L. Similar to participant C, when correlations between OTUs, collected data, and diversity metrics were calculated, we found no correlations

between OTUs and collected data such as antibiotic use, FEV1, and symptom scores (**Table C.5**). *Fusobacterium* OTU 5 was positively correlated with *Prevotella* OTU 24 over the study period (**Fig 4.5d**). Additionally, as observed with participant C, Shannon diversity was correlated with multiple OTUs (**Fig 4.5d, Table C.5**). Participant E's lung community was consistently dominated by 3 OTUs corresponding to *Pseudomonas, Prevotella,* and *Streptococcus* (**Fig 4.6b**). In contrast to the stable microbiome seen in participant C, the community in participant E contained many members which fluctuated over the study period (**Fig 4.6b**).

4.5 Discussion

Our current understanding of the pathophysiology of pulmonary exacerbations in CF is limited. Understanding the mechanisms underlying pulmonary exacerbation, and thus being able to mitigate symptom onset and/or severity would have important implications for individuals with CF. Pulmonary exacerbations likely have many triggers including elements of the inflammatory response, lung microbiota, and extrinsic factors such as pollution, allergen exposure and medication compliance (Ferkol *et al.*, 2006). Because antimicrobial therapies often control and resolve the symptoms associated with pulmonary exacerbations, it is important that we understand the longitudinal dynamics of the CF lung microbiota with respect to onset of pulmonary symptoms.

In this study, when examined as discrete groups, samples of the CF lung microbiota obtained during Treatment, Intermediate, and Stable periods were identified as significantly different from each other (PERMANOVA of Bray-Curtis distance, p=0.016) though highly confounded by the originating participant. However, when

samples from each participant were examined independently, it was evident that discrete changes in microbial composition only accompanied some pulmonary exacerbations (**Fig 4.3**). Further, longitudinal analyses did not provide statistically significant correlations between respiratory symptoms and elements of the lung microbiota (**Fig 4.5-4.6**). Notably, although changes are seen during some participant's pulmonary exacerbations, some individual's lung microbiota also undergo large compositional changes during periods of clinical stability. These types of changes may result from changes in antimicrobial therapy (Sibley *et al.*, 2008; Coburn *et al.*, 2015), changes in pulmonary function (Surette, 2014; Lynch and Bruce, 2013), or other undetermined factors.

When we focused on the 2 participants in the study who had the most (n=4) and least (n=0) number of exacerbations during the study period, longitudinal analyses were unable to provide general microbiome patterns predicting exacerbation; there were no correlations between exacerbation and alpha or beta diversity, FEV1, antibiotic use, or symptom scores. Previous longitudinal analyses of the CF lung microbiota's role in pulmonary exacerbation onset have drawn similar conclusions to those observed within this study (Carmody *et al.*, 2015; Cuthbertson *et al.*, 2015). Instead, statistically significant correlations between alpha diversity and microbial membership (i.e. OTUs) were identified, as well as correlating OTUs. In both participants there was a negative correlation found between dominating members of the microbiota and alpha diversity. However, these correlations are likely a result of the compositional and relative nature of the 16S rRNA gene sequencing approaches employed. In participant C, positive correlations were observed between *Prevotella*

and *Streptococcus*, as well as between 2 *Streptococci*. In both participants, correlations were observed between *Prevotella* and *Fusobacterium*. These species are often found in the lungs of individuals with CF, but haven't been previously correlated. However, *Streptococcus salivarius* and *Prevotella intermedia* have been implicated in coaggregation in periodontal disease (Levesque *et al.*, 2003), and oral streptococci and *Prevotella* have been isolated together from dentoalveolar abscesses (Sakamoto *et al.*, 1998), indicating that organisms within these genera may correlate in a variety of infectious diseases. While these correlations did not differ before, during, or after pulmonary events, they may be important microbe-microbe interactions in this environment which should be further investigated. It is important to note that because of imperfections in OTU clustering approaches (Westcott and Schloss, 2015; He *et al.*, 2015), that the 2 correlating streptococci OTUs may in fact be sequences from the same organism which were misclustered into 2 separate OTUs.

In this study, we identified inter-individual differences of the CF lung microbiota in terms of taxonomic composition (**Fig 4.3**), alpha (**Fig 4.4**), and beta diversity (**Fig 4.2**). The results of this study help us to consider the goal of this research: to better understand and improve the lives of those suffering from CF. Studying 6 participants longitudinally has identified that conclusions which have been made in the literature which apply at the population-level are not necessarily meaningful to the individual. For example, we report that periods of exacerbation were not consistently correlated with an increase in Shannon Diversity (**Fig 4.4**). This is in contrast to previous results which have shown increases in alpha diversity during exacerbations when compared to surrounding time points at the population-level (Fodor *et al.*, 2012;

Smith *et al.*, 2014). It has been previously suggested that a patient-specific, cross-sectional use of alpha diversity to predict state of disease would not be of use (Surette, 2014), especially since measures of alpha diversity cannot be acted on in the clinic via a specified treatment or pharmacological aid. Because of the unique nature of microbial acquisition in the lungs, CFTR modulators, and patient environments and actions, individuals with CF represent unique patients who should be assessed in a case-by-case basis.

The most important limitation of this study are the short-comings of using 16S rRNA sequencing of sputum as a measure of the CF lung microbiome. First, 16S rRNA sequencing does not distinguish nonviable from viable cells. Second, the onset of exacerbations may be triggered by a small proportion of the total community or by non-bacterial members of the microbiota; however, this method does not differentiate between metabolically active and inactive members and does not capture non-bacterial components (Whelan and Surette, 2015). Third, although conflicting studies exist (Rogers *et al.*, 2006), expectorated sputum may be subject to contamination by oral microbes (Goddard *et al.*, 2012; Muzanye *et al.*, 2009). These important shortcomings of our ability to fully understand the CF lung microbiome may mean that with the advantages that 16S rRNA sequencing of sputum affords (total community profiling with relative abundance information), that the associated disadvantages may be masking an important microbial component to these events.

A small sample size of participants enrolled and completed the study (**Table 4.1**). Prospectively collecting and storing sputum samples is tedious and difficult in a large patient cohort. Previous studies were similarly limited to small patient numbers (Carmody *et al.*, 2015; Cuthbertson *et al.*, 2015). Additionally, requesting tri-weekly

symptom score profiles and self-administered spirometry measurements furthers the participant burden on individuals with a disease that already requires time-consuming pharmacologic and physical therapies (Sawicki *et al.*, 2009). However, longitudinal studies of the dynamics within the CF lung microbiota are important in determining the bacterial component of pulmonary exacerbation.

By studying 6 people with CF for up to a year in a prospective, longitudinal study of the microbiota preceding, during, and following exacerbation, we conclude no discernible, participant-wide dynamics which explain the onset of pulmonary exacerbation. Some hypothesized causes of pulmonary exacerbations may not have been measurable in this study; for example, we have previously hypothesized that strain dynamics would be very difficult to determine from community-wide studies of the 16S rRNA gene (Whelan and Surette, 2015). Further, elements other than the microbiome, such as host inflammatory defenses, may be the driving force behind these events. This study also supports the growing data that suggest that the lung microbiome in CF is highly patient-specific and that it should be investigated as such.

4.6 Acknowledgements

The authors would like to acknowledge the tireless efforts of the participants involved in this study for their dedication to the scientific process.

Chapter 5

Culture-enriched metagenomic sequencing of the cystic fibrosis lung microbiota

Preface

Research presented as part of this chapter has been prepared for publication.

Whelan FJ, Waddell B, Syed SA, Rabin HR, Parkins MD, & Surette MG. Culture-enriched metagenomic sequencing of the cystic fibrosis lung microbiota. *Written for publication.*

Author Contributions: FJW is the primary, first-author of this prepared manuscript. HPR, and MDP collected patient information and willing participations for this study. BW cultured sputum samples. FJW and SAS and performed laboratory experiments for this study. FJW and MGS and conceptualized the experimental outline. FJW conducted all data analyses and wrote this manuscript draft.

Supplemental material prepared for this manuscript is presented in **Appendix D**.

5.1 Abstract

Next generation sequencing technologies have afforded the field of microbiome research with the ability to profile communities without the need to culture their inhabitants. While the power of these technologies is clear, combining these approaches with classical, culture-dependent methods can provide us with greater resolution of human microbiota. In this work, we have combined advances in marker gene and metagenomic sequencing with culture-enriched molecular profiling to study the microbiota of the cystic fibrosis lung. Culture-enrichment consistently recovered the CF lung microbiota, culturing an average of 81.21% of OTUs representing 99.15% of the relative abundance of those sequenced from sputum lung samples directly. Further, culture-enrichment retrieved 65.5% more species than culture-independent methods, reflecting the selective ability of media to enrich for low abundance organisms. Using a novel plate coverage algorithm, we conducted metagenomic sequencing on a minimal subset of culture plates in a patient-specific manner. When compared to culture-independent methods, culture-enriched metagenomics consistently recovered more organisms with better genome coverage when compared to metagenomics conducted on sputum samples. However, the interpretation of these results differ depending on which software approach is utilized. Culture-enrichment of sputum reconstructs the CF lung microbiota for a better understanding of this disease.

5.2 Introduction

The study of microscopic organisms is dependent on the physical attainment of these organisms via culture isolation. The field of microbiology began when van Leeuwenhoek became the first to visualize microbes using the first microscope (Van Leeuwenhoek, 1683), and continued once we learned to control the growth and propagation of these organisms. However, the advent of next generation sequencing technologies has allowed for the study of microbial communities without the requirement of culture, expanding our understanding of the diversity of microbial communities. Specifically, sequence-based studies of the human microbiota have identified the effect that environment, diet, and host genetics can have on the formation and persistence of these communities (Turnbaugh *et al.*, 2009; The Human Microbiome Project Consortium, 2012b; Spor *et al.*, 2011).

These community-wide, culture-independent studies have made crucial contributions to the field of microbiome research and our knowledge of host-microbiota interactions. However, without being able to culture these microbes, we lose the ability to conduct classical microbiology, leading to the current lack of mechanistic studies within the microbiome field and instead a focus on "dysbiosis" and "diversity" (Olesen and Alm, 2016; Shade, 2017).

While it is commonly cited that the majority of the human microbiota is unculturable, a number of studies suggest that this is not the case. Before the term "microbiome" was commonplace, Finegold *et al.* cultured up to 300 different species from 40 fecal specimens using both aerobic and anaerobic culture (Finegold *et al.*, 1974). Goodman *et al.* cultured almost half of the human gut microbiota, recovering 316 Operational Taxonomic Units (OTUs) from the 631 OTUs identified by

culture-independent techniques (Goodman *et al.*, 2011). Lagier *et al.* used the term "culturonomics" to describe their ability to recover 340 species of bacteria, 31 of which were novel, from 3 stool samples (Lagier *et al.*, 2012). Two more recent studies recovered 88% of family-level OTUs (Rettedal *et al.*, 2014) and 95% of all OTUs identified in fecal specimens (Lau *et al.*, 2016); importantly, both of these studies also identified more OTUs via culture-dependent than culture-independent approaches. Besides the gut microbiota, other human-associated communities have been cultured, including urine (Hilt *et al.*, 2014), skin (Myles *et al.*, 2016), the oral cavity (Thompson *et al.*, 2015), and the CF lung microbiota (Sibley *et al.*, 2011).

Cystic fibrosis (CF) is the most prevalent genetic disease in caucasians (Elborn, 2016). Among the most severe symptoms are those which affect the lungs. A decrease in mucus viscosity and a deregulation of the host immune response provides a microbial niche that allows for the chronic colonization of the lungs (O'Sullivan and Freedman, 2009). A thorough understanding of this microbial community is vital to allow proper treatment with anti-microbial agents for improved clinical care.

This study presents the merger of culture-enrichment and next generation sequencing technologies in order to obtain a more thorough understanding of the genomic content of the CF lung microbiota. First, 16S rRNA gene sequencing was used to establish that 81.21% of all OTUs - representing 99.15% of the relative abundance - in the CF lung microbiota are culturable using commonly available agar media. In fact, culture-enrichment of sputum samples resulted in increased OTU recovery when compared to culture-independent investigations of the sputum sample itself. Following, culture-enriched metagenomic profiling was performed, which improved genomic recovery by taking advantage of the natural ability of culture conditions

to biologically bin organisms by growth conditions. As part of this approach, we present a plate coverage algorithm which allows the user to direct culture-enriched metagenomic sequencing efforts based on the diversity identified in each culture condition via marker gene analysis. We identify the advantages of culture-enrichment by the increased recovery of OTUs in the 16S rRNA gene sequencing, the decrease in host contamination in metagenomic approaches, and the increase in genetic content obtained from culture-enriched metagenomic sequencing.

5.3 Methods

5.3.1 Sputum collection and culture-enrichment

Sputum samples were collected from December 4th 2013 to October 6th 2014 from willing participants visiting the Calgary Adult CF Clinic (ethical approval granted by the Calgary Health Region Ethics Board, REB-24123). Two samples were collected from each patient: one at the onset of pulmonary exacerbation (as defined by Fuchs *et al.* (Fuchs *et al.*, 1994)) upon hospitalization but prior to intravenous antibiotic therapy, and a second during a follow up appointment 1 week to 4 months following the resolution of symptoms.

Samples were transported to an anaerobic environment within 60 seconds of expectoration and plated within 4 hours of sputum production. Samples were homogenized by passage through a 18 gauge needle and 1 mL syringe. Once homogenous, 300μL was set aside for culture-independent sequencing. The remainder was used for culture enrichment. Thirteen solid agar media were prepared: Actinomycetes isolation agar (AIA; BD), brain heart infusion agar (BHI; BD), cooked meat broth with 1.5%

agar (Beef; Fluka), Columbia agar base with 5% sheep's blood (CBA; BD), GC powder (BD) with 5% hemoglobin, and 0.5% IsoVitaleX (CHOC; BD), Columbia CNA agar with 5% sheep's blood (CNA; BD), fastidious anaerobe agar (FAA; Acumedia), tryptic soy agar with 0.1μg/mL kanamycin, 7.5μg/mL vancomyin, 10μg/mL Vitamin K, 0.05ng/mL hemin, and 5% laked blood (KVLB), MacConkey agar (MAC; BD), mannitol salt agar (MSA; BD), McKay media (Sibley *et al.*, 2010b), phenylethyl alcohol agar with 5% sheep's blood (PEA; BD), and tryptic soy agar with 1.5% yeast extract (TSY; BD). To Beef, BHI, and TSY, the following additional additives were included: 10μg/mL colisitin sulfate, 0.5mg/mL L-Cysteine, 1.0ng/mL Vitamin K, and 10ng/mL hemin.

Culture enrichment was performed by placing 100μL of sputum diluted in BHI with 0.05% L-Cysteine to 10^{-3} and 10^{-5} on to each of the above media. Two sets of plating were performed, one which was incubated aerobically (5% CO_2, 37°C) and another anaerobically (5% CO_2, 5% H_2, 90% NO_2, 37°C). This method resulted in 52 plates per sample.

After 3-5 days (aerobic) and 5-7 days (anaerobic) of growth, plates were imaged and growth acquired by adding 2-3mL of BHI broth to each plate and lifting colonies. Part of this solution was frozen directly for DNA extraction while the rest was frozen in 20% skim milk for any potential growth or re-isolation. Any plate which resulted in no visible bacterial colonies was discarded and omitted from all downstream processing. Comparisons of this culturing method to that employed by a typical microbiology lab in **Figure 5.5** included the following media types: aerobic CBA, MAC, MSA, and anaerobic CHOC.

5.3.2 DNA isolation and Illumina sequencing

Genomic DNA was isolated from culture-enriched plates and sputum as previously described (Whelan *et al.*, 2014) with the exception of use of lifted colonies/homogenized sputum as input instead of Copan Swabs as performed in (Sibley *et al.*, 2011). Dilutions resulting from the same media/environment pairing were combined into one genomic DNA isolation for a maximum of 26 culture-enriched plates per sputum sample. The variable 3 region of the 16S rRNA gene was amplified using universal primers as adapted from (Whelan *et al.*, 2014; Bartram *et al.*, 2011). The PCR reaction consisted of 5pmol of each primer, 1ng template DNA, 200μM dNTPs, 1.5mM $MgCl_2$, and 1 U Taq polymerase. The PCR protocol employed is as follows: 95°C for 5 minutes, followed by 30 cycles of 95°C for 30 seconds, 50°C for 30 seconds, and 72°C for 30 seconds, with a final 72°C for 7 minutes. Presence of a PCR product was verified by electrophoresis (2% agarose gel). PCR products were sequenced using the Illumina MiSeq platform using 2x250 paired-end reads.

Using the plate coverage algorithm developed by the authors and discussed in detail in the Results (**Fig 5.6**), a subset of samples were prepared for metagenomic sequencing. Select culture-enriched samples and all sputums were sonicated to 300bp and library preparations were made using the NEBNext DNA Library Prep Master Mix Set for Illumina (New England Biolabs) and sequenced using the Illumina HiSeq platform with 2x250 paired-end reads.

5.3.3 16S rRNA sequence processing and analysis

16S rRNA paired-end reads were processed using a custom, in-house pipeline as previously described (Whelan *et al.*, 2014). Reads were trimmed of any remaining primers

using cutadapt (Martin, 2011) and discarded using sickle based on a quality threshold of 30 (`https://github.com/najoshi/sickle`). Paired-end reads were assembled using PANDAseq (Masella *et al.*, 2012). OTUs were picked using AbundantOTU+ with a 97% clustering threshold (Ye, 2011) and chimeras removed using USEARCH (Edgar, 2010) as implemented in QIIME (Caporaso *et al.*, 2010c). The Ribosomal Database Project classifier (Wang *et al.*, 2007) was used to assign taxonomy against the 4th February 2011 release of the Greengenes database (DeSantis *et al.*, 2006). OTU tables were created with QIIME (Caporaso *et al.*, 2010c). Any OTU which was not assigned a bacterial taxonomy or which only had one instance across the full dataset (singleton) was culled. Any sample with < 1000 reads was discarded. The result of this culling process, in combination with only sequencing plates with visual growth, resulted in a total of 531 samples (20 sputum samples and 511 plates). The mean sequence depth across this dataset was 68,160 reads per sample (range 2,032-159,381), with a mean number of OTUs of 94.1 (range 10-311).

Taxonomic summaries over multiple samples were performed by calculating the maximum relative abundance across samples, and normalizing to 100%. Principal Coordinate Analysis (PCoA) plots were calculated using phyloseq (McMurdie and Holmes, 2013) and ggplot2 (Wickham, 2009) in R after proportional normalization (McMurdie and Holmes, 2014). An OTU was considered present in a sputum or cultured sample if it had a relative abundance of $> 0.01\%$ (all exceptions noted). Phylogenies were decorated with GraPhlAn (Asnicar *et al.*, 2015). Heatmaps were calculated using pheatmap (Kolde, 2012).

5.3.4 Recovery of isolates from frozen culture-enriched stocks

Improved isolation of *Stenotrophomonas maltophilia* from frozen skim milk stocks of select plates was performed using a selective medium as described in (Denton *et al.*, 2000). Isolates were Sanger sequenced using the 8F (5'-AGAGTTTGATCCTGGCTC AG-3') and 926R (5'-CCGTCAATTCCTTTRAGTTT-3') primers to the 16S rRNA gene, resulting in a 900nt product. The identity of the isolates were confirmed by comparisons to the Human Oral Microbiome Database (HOMD) and to NCBI's 16S ribosomal RNA sequences (Bacteria and Archaea) Database (**Table D.1**).

5.3.5 Metagenomic sequence processing and analysis

Resultant paired-end reads were processed using available software and tools. Cutadapt was used to trim Illumina adaptors and primers (Martin, 2011), and sequences were removed based on quality by using sickle with a threshold of 30 (`https://github.com/najoshi/sickle`). Culture-independent sputum samples were decontaminated using DeconSeq (Schmieder and Edwards, 2011). 16S rRNA sequences were extracted out of the metagenomic data by alignment to the 2011 Greengenes database (DeSantis *et al.*, 2006) using Bowtie2 (Langmead and Salzberg, 2012). The trimmed, quality-filtered FASTQ culture-enriched metagenomic reads were assembled using RayMeta (Boisvert *et al.*, 2012) and binned using CONCOCT (Alneberg *et al.*, 2014) and MaxBin (Wu *et al.*, 2014). Visualizations of the CONCOCT binning procedure was done with an auxiliary script distributed with CONCOCT (Alneberg *et al.*, 2014). Taxonomic assignments for each bin were conducted by using BLASTn (Altschul *et al.*, 1990) to align the contigs of each bin to NCBI's RefSeq database. Analyses of these bins including calculations of genome coverage were performed using

samtools (Li *et al.*, 2009), and bamtools (Barnett *et al.*, 2011).

5.4 Results

In this study, we collected 20 sputum samples from 10 patients for both culture-independent and culture-dependent profiling (**Fig 5.1**). Samples were homogenized before 300μl was set aside (culture-independent) and the rest plated onto 13 different media types under aerobic and anaerobic conditions (culture-dependent). 16S rRNA gene sequencing was performed on the sputum sample as well as on the collective organisms grown in each media/environment pairing for a total of 26 culture-enriched samples per sputum obtained. Following, based on the distribution of OTUs identified via 16S rRNA gene sequencing, a representative subset of plates (and the original sputum) were chosen for metagenomic sequencing based on a novel plate coverage algorithm.

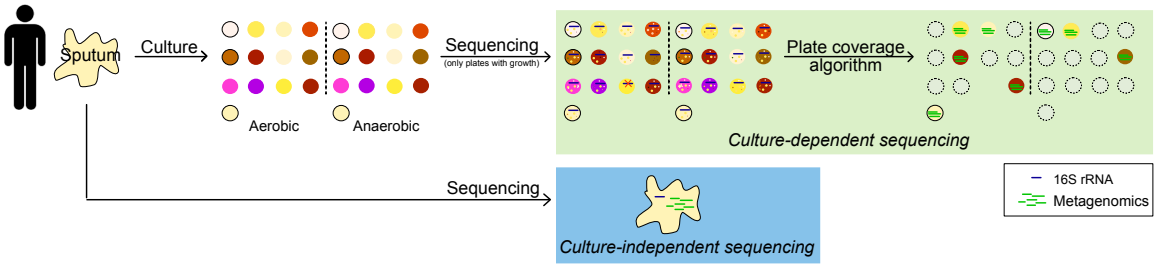


Figure 5.1: **Culture-enriched metagenomic sequencing workflow.** Sputum samples collected from cystic fibrosis (CF) patients were plated onto 13 selective and non-selective media and incubated either aerobically or anaerobically. 16S rRNA gene sequencing was conducted on the sputum sample (direct profiling) as well as on each media type (culture-enriched profiling). Metagenomic sequencing was conducted on the sputum sample as well as on an appropriate subset of plates as indicated by the plate coverage algorithm (see **Figure 5.6**).

5.4.1 The majority of the cystic fibrosis lung microbiota is culturable

Using the conditions outlined above, we first chose to study the sum of the culture-enriched microbiota per sample in relation to those OTUs recovered from sequencing of the sputum directly (mean of 55 OTUs per sample). For each sample, the majority of OTUs identified by culture-independent sequencing were also found in the culture-enriched samples, indicating that most OTUs are culturable (**Fig 5.2a, "Shared"**; average 81.21% shared, range 64.62% - 100%). Further, when the relative abundance of these cultured OTUs was examined, they constituted 99.15% of the microbiota of the sputum sample (**Fig 5.2a, green line**, range 97.62% - 99.8%), indicating that those OTUs which were not recovered by culture were at low abundance within the originating community.

Examining all recovered OTUs (mean of 168 OTUs per sample) across both culture-independent and -dependent sequencing in all 20 samples across 10 patients, we see that few OTUs are never cultured (**Fig 5.2b, blue ring**; OTUs $\geq$ 0.01% relative abundance in culture-dependent sequencing) when compared to those which are seen in both the culture-independent and -dependent sequencing (**Fig 5.2b, grey ring**; OTUs $\geq$ 0.01% relative abundance in at least one culture-independent and -dependent sample). The unculturable groups include Spirochaetes (7 of 7 OTUs identified in culture-independent sequencing) and SR1 (1 of 1), and many members of the Tenericutes (7 of 22), and TM7 (2 of 3) phyla.

These results still hold if a more stringent OTU cutoff of $\geq$0.1% relative abundance is used (**Fig D.1**).

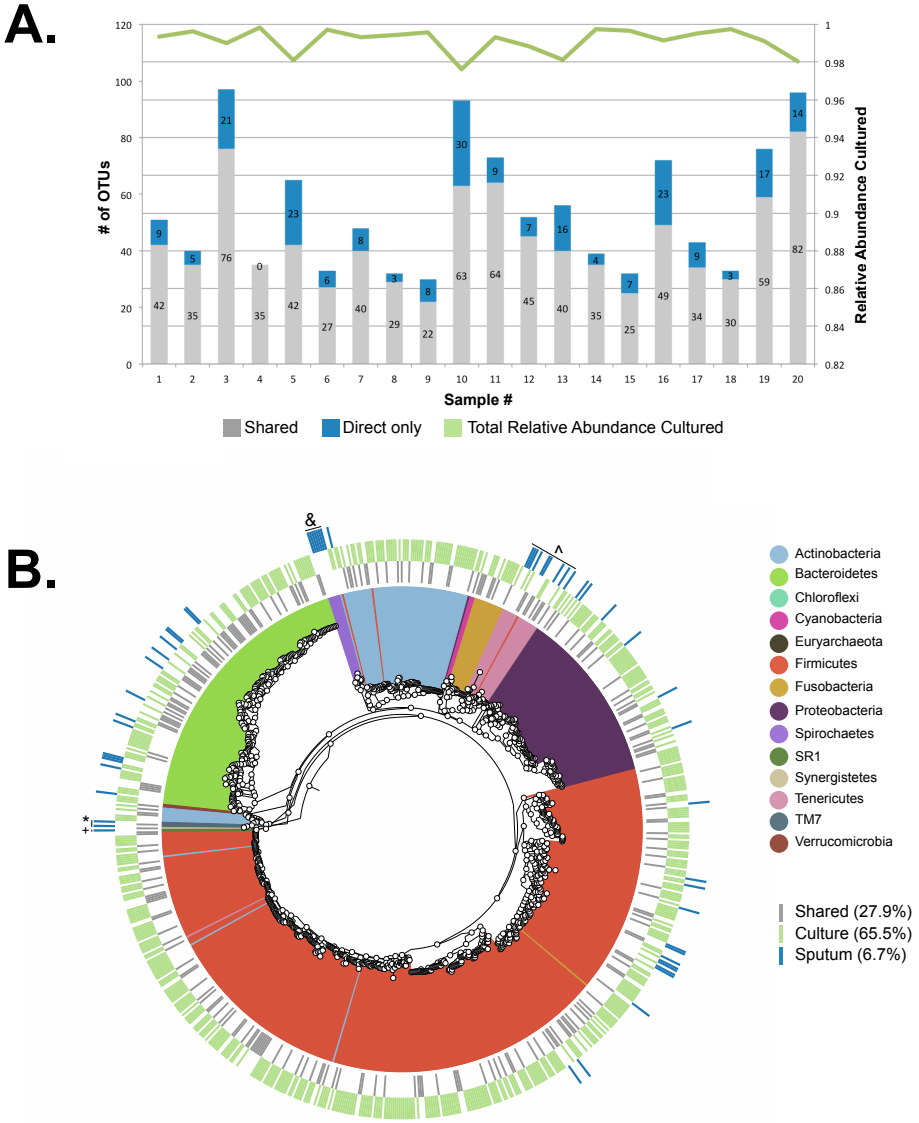


Figure 5.2: **The majority of the CF lung microbiota is culturable. A.** Of the OTUs identified in the direct profiling of each sputum sample, most are also cultured. The number of OTUs which are identified only by direct profiling (i.e. are not cultured) differs between samples; however, the culturable proportion of culture-independent OTUs is 81.21% on average and correspond to 99.15% of the relative abundance of the lung microbiome (green line; range: 97.62%-99.8%). **B.** Across the dataset, 6.7% of OTUs were unculturable, including many Tenericutes (^), and TM7 (*), and all Spirochaetes (&), and SR1 (+). Otherwise, uncultured OTUs were not restricted to a particular clade or bacterial family. Similar results are also true when a more stringent relative abundance cutoff is used (**Fig D.1**).

5.4.2 Culture-enrichment increases OTU recovery

What is perhaps most interesting is the frequency of organisms present only in the culture-enriched sequencing (**Fig 5.2b, green ring**; OTUs $\geq 0.01\%$ relative abundance in ≥ 1 culture-dependent sample but $\leq 0.01\%$ in all culture-independent samples). In fact, for each sputum sample examined in this study, there was a greater number of OTUs recovered from the associated culture-enrichment than from the sample itself (**Fig 5.3a**). On average across the dataset, 6.7% of OTUs were never cultured, 27.9% of OTUs were identified via both methods, and 65.5% of OTUs were identified via culture-enrichment alone (**Fig 5.3b**). We hypothesized that this may be a result of the ability of culture-enrichment to recover low abundance taxa which may be below the detection threshold of our sequencing depth (average of 68,160 reads per sample), but when given the opportunity to proliferate in an environment which is, perhaps, amenable to growth, is detectable in the culture-enriched sequencing results.

To test this hypothesis, we re-sequenced one of our sputum samples to a depth 24x deeper than the original (41,199 versus 972,834 reads) and rarefied it at decreasing depths (range: 500,000-1,000 reads). Comparisons across these sequencing depths (**Fig 5.4**) indicated that the number of culture-only OTUs decreases as the sequencing depth increases (**Fig 5.4, inset**). This result is made apparent by visualizing rank abundance curves at each depth which shows the distribution of cultured OTUs along the long tail distribution of these sputum samples, the rarest of which appear as culture-only OTUs when lower depth sequencing is used. Thus, culture-dependent sequencing allows for the enrichment of low abundance taxa present in the sputum sample but not necessarily sequenced by the depths typically employed.

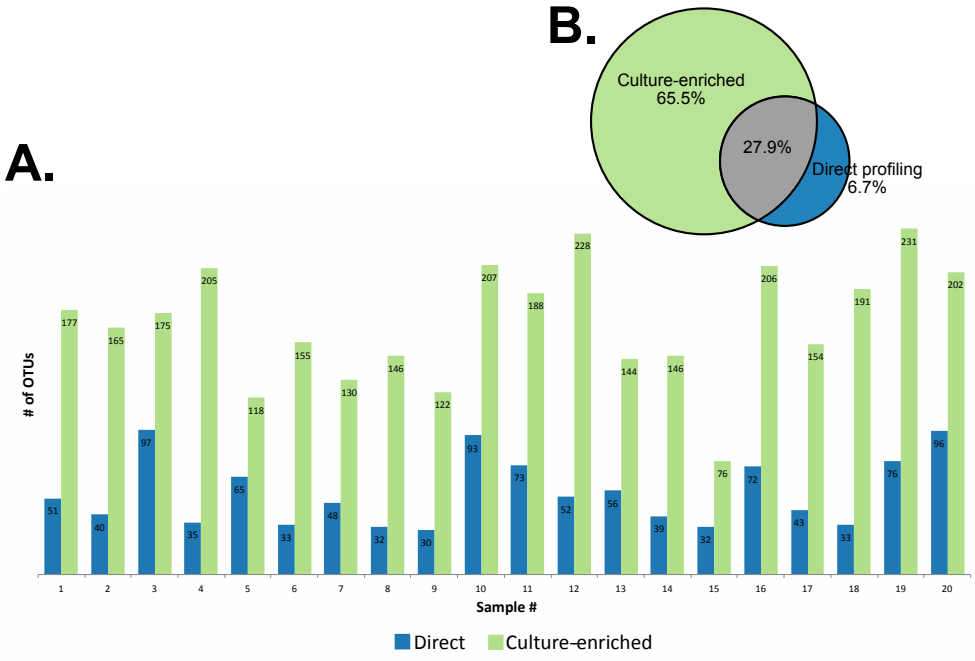


Figure 5.3: **Culture enrichment results in an increase in OTU recovery when compared to direct profiling.** **A.** In each sample, more OTUs were recovered via extensive culture than were found in the direct profiling of the sputum sample itself. **B.** Overall, an average of 65.5% of all OTUs were identified only via culture-enrichment.

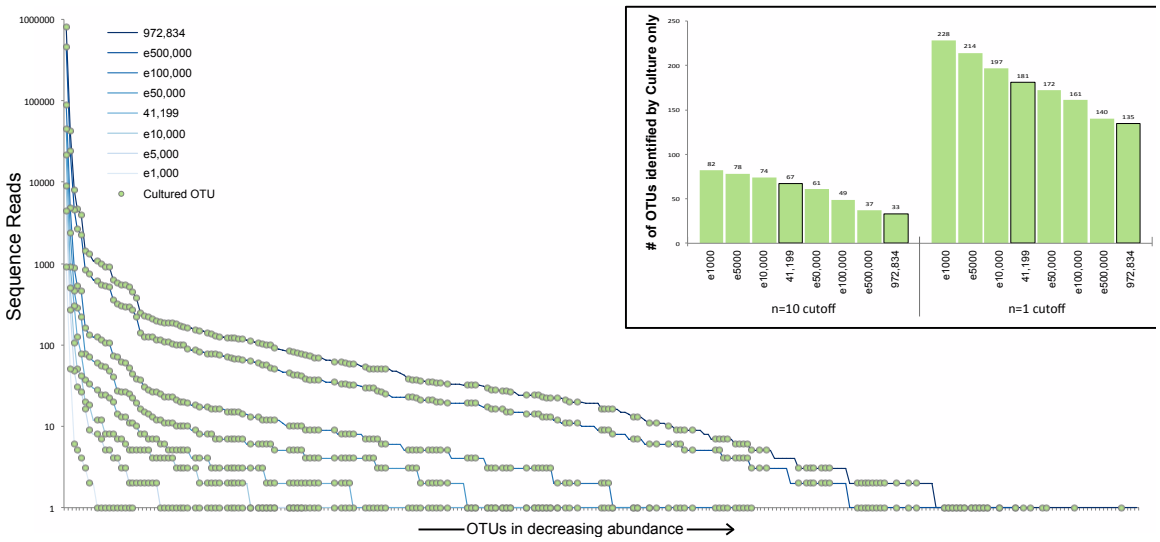


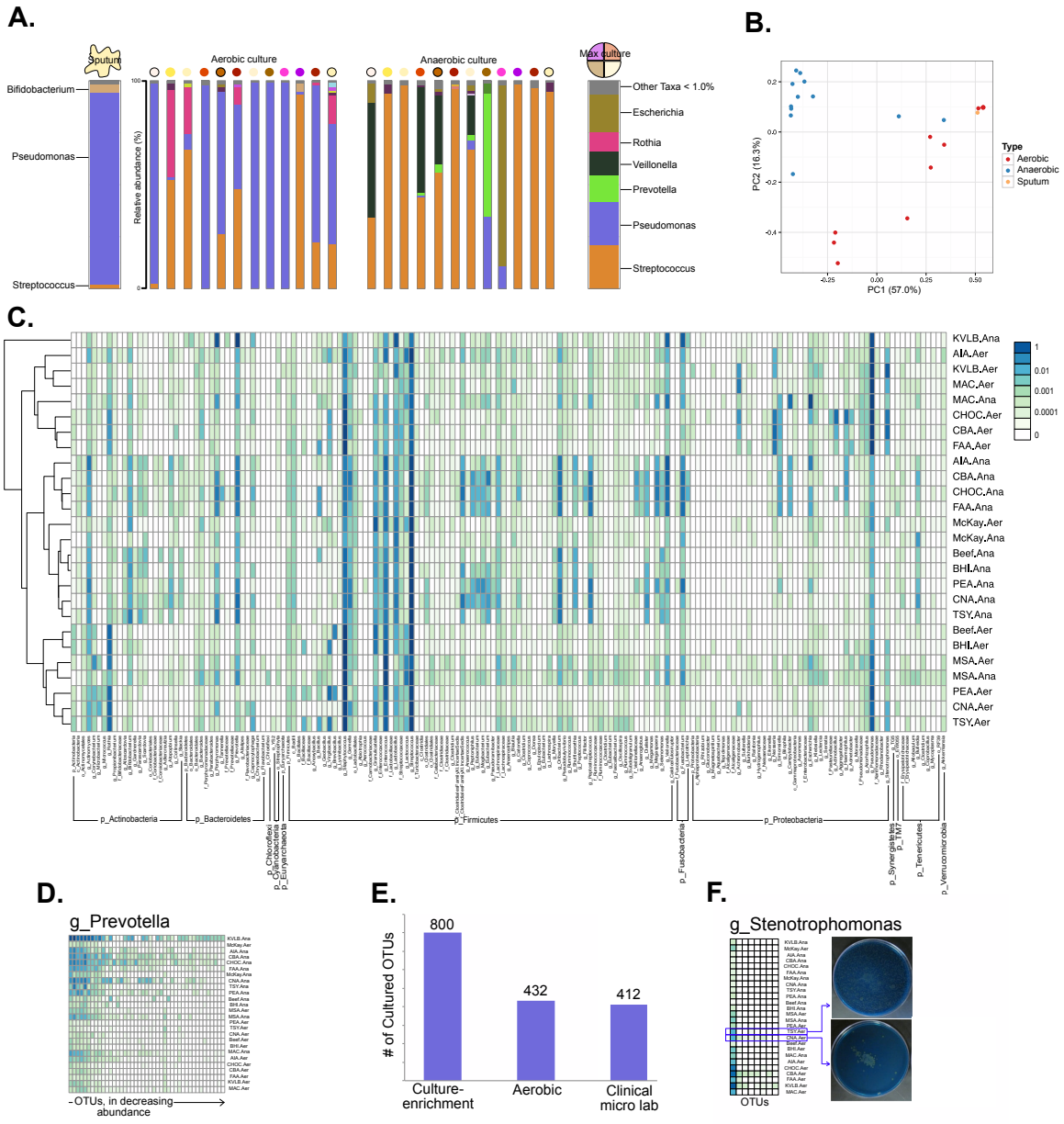
Figure 5.4: **Culture-enrichment enriches for low abundance taxa.** A single sputum sample was sequenced to a depth of 41,199 reads and again 24x deeper 972,834 reads. A rank order curve illustrating the number of OTUs present at both sequencing depths. With 972,834 reads, more Shared (direct and culture-enriched) OTUs are discovered that were originally only seen in the culture-enriched sequencing, demonstrating that culture-enriched for low abundance taxa. The number of OTUs recovered via culture but not seen in the direct profiling decrease when a deeper sequencing depth is used, indicating that enriched culturing allows for the recovery of low abundance organisms not captured with typical sequencing depths (**inset**).

5.4.3 Culture-enrichment's increase in OTU recovery is dependent on media type and oxygen availability.

Next, we focussed on the OTU diversity within each culture-enriched plate pool individually. Each plate was incubated in one of two environments (aerobic or anaerobic) and consisted of 1 of 13 media types. In order to demonstrate the taxonomic distribution across culturing conditions, we first focussed on one sputum sample and its associated culture-dependent sequencing results. By this approach, the importance of using both anaerobic and aerobic conditions is evident in the differences between taxonomic composition (**Fig 5.5a**). For example, *Veillonella*, and *Prevotella* species are recovered exclusively under anaerobic conditions; conversely, *Rothia* and *Pseudomonas* species are obtained at greater abundances in aerobic culture (**Fig5.5a**). The community-wide differences between these environmental conditions can be visualized by a PCoA analysis (**Fig 5.5b**); here, the sputum sample clusters tightly with the aerobic samples due to the prominence of *Pseudomonas* (and not due to a lack of bacterial growth) (**Fig D.2**). A hierarchical clustering of the media/environment pairings employed across all genera cultured from 20 sputum samples shows the importance of the variety of conditions utilized (**Fig 5.5c**). While some taxa are less discrete as to which media they can be cultured on (e.g. *Streptococcus*), others, such as *Neisseria, Rothia,* and *Stenotrophomonas* can only proliferate on a subset of the conditions employed. Further, a similar analysis at the OTU level (**Fig D.3**) shows that there are also OTU-dependent differences in growth patterns (e.g. *Prevotella* OTUs, **Fig 5.5d**). Importantly, across the full dataset, anaerobic culture was responsible for almost half of the total recovery of OTUs, and the expansion of our culture conditions from the 4 typically employed in the clinical lab to our 13 nearly

doubles the number of OTUs cultured (**Fig 5.5e**).

A further advantage to this approach is that it allows for the post-hoc recovery of particular organisms of interest from the frozen skim milk stocks made from each individual plate. As an example of this method and its usefulness, *Stenotrophomonas* was isolated from two mixed communities in which it accounted for 1.3 and 1.5% of the relative abundance in each cultured community. To do so, the skim milk stocks were replated on a media type designed to isolate this genera (**Fig 5.5f**). Colonies were confirmed via full-length 16S rRNA gene sequencing (**Table D.1**). This result is important as it shows the viability of organisms as part of these mixed communities, and that we can recover and isolate organisms of interest from the original culture post-hoc for further phenotypic, functional, and/or mechanistic analyses.



Caption follows on next page.

Figure 5.5 *(previous page)*: **Increased OTU recovery seen with culture-enrichment is dependent on the variety of media types and environments employed.** The variety in selective and non-selective media types, and aerobic and anaerobic environments is important in capturing the diversity of the CF microbiome. The use of both anaerobic and aerobic conditions encourage the recovery of very different taxa, as indicated by examining the taxonomic profiles of a sputum and associated culture (**A**) as well as the distinct difference in β-diversity scores (**B**). **C.** A heatmap showing the maximum observed relative abundance (displayed between 0-1) of each genus across culturing conditions indicates the necessity of such extensive plating in order to culture the diversity within the CF lung microbiota. **D.** This necessity is further evident by the differences in suitable culture conditions necessary for OTUs within the same genus. Here, *Prevotella* is shown as an example; **Fig D.3** shows a heatmap indicating suitable culture conditions for each OTU. **E.** Across the full sample set, the number of OTUs obtained from culture-enrichment is compared to the number obtained if only aerobic culture was employed, or if culture was restricted to that of a standard clinical microbiology laboratory. **F.** Cultured organisms can be recovered from frozen plate pool stocks. Here, a medium designed for growth of Stenotrophomonas species was used to select for a Stenotrophomonas OTU present at 1.3% on CNA.Aer and 1.5% on TSY.Aer. Max Culture = the maximum relative abundance for each OTU observed across culture enrichment and normalized to within 0-1.

5.4.4 The plate coverage algorithm defines the optimal subset for culture-enriched metagenomic sequencing

The media/environment pairings of plates used for culture-enrichment are necessarily broad in order to capture a wide range of organisms. The CF lung microbiota, like other human-associated communities, can host a wide range of organisms, from common pathogens such as *Pseudomonas, Staphylococcus*, and *Haemophilus* (Surette, 2014), to recently-appreciated anaerobes (e.g. *Prevotella, Fusobacterium*, and *Veillonella* (Tunney *et al.*, 2008)), and emerging pathogens (e.g. *Stenotrophomonas*, and *Achromobacter* (Parkins and Floto, 2015)). While the lung could be home to any of these plethora of organisms, an individual's lung microbiota is a patient-specific subset of these possibilities. As such, this means that while the variety of conditions we employ as part of culture-enriched profiling is necessary to capture the diversity

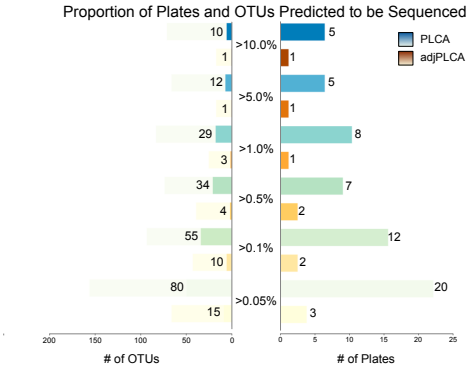
across patients, not every plate is needed to enrich each patient sample. This information is not known *a priori*; however, using 16S rRNA gene sequencing results as a proxy, the minimum number of plates needed to capture the diversity within a given sputum sample can be determined and applied to metagenomic sequencing. As such, we wrote the PLate Coverage Algorithm (PLCA) which, based on 16S rRNA gene sequencing of each plate, determines the optimal subset of plates that are needed to capture the totality of the culture-enriched microbiota (**Fig 5.6**). Briefly, this algorithm identifies all OTUs present in only one media/environment pairing, and includes that plate in the optimal subset. Following, any OTU and plate not already covered by this subset will be examined such that, recursively, the plate with the most remaining OTUs is added to the subset until all OTUs are covered.

There are two versions of the PLCA. The *denovo PLCA* (**Fig 5.6a, PLCA**) is designed to capture the total culture-enriched diversity of a sample, independent of the relative abundance or content of the associated culture-independent sequencing. This version allows the interrogation of the totality of the cultured community. However, the relative abundance within the originating sample is often important, especially when answering research questions surrounding clinically associated samples such as those from the CF lung microbiota. For these cases, an *adjusted PLCA* (**Fig D.4, adjPLCA**) generates the subset of plates on which all OTUs from the originating sample have been cultured. The adjusted PLCA often produces a smaller set of plates then the denovo PLCA.

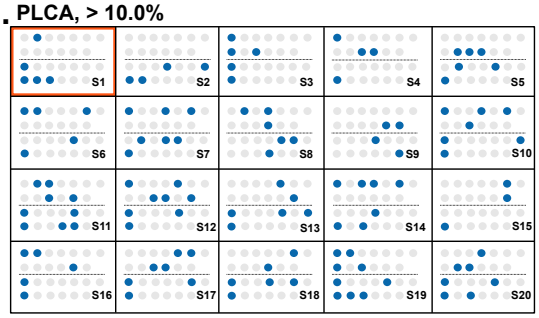
A.

```
pre:
  abund = abundance cutoff
  plateList = ()
  allOTUs = all cultured OTUs

for each OTU a in allOTUs
   count the # of plates with OTU a at abundance > abund
   if count == 1
      plate p = plate w/ OTU a on it
      plateList += plate p
      allOTUs -= any OTU on plate p
while allOTUs is not empty
   find plate q with the greatest # of OTUs remaining in all OTUs
   plateList += q
   allOTUs -= any OTU on plate q

post:
  plateList = list of plates for metagen seqing
  allOTUs = ()
```

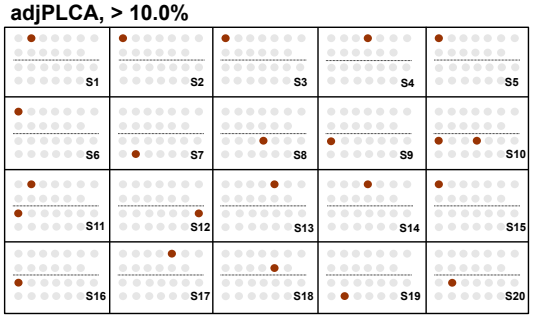
B.



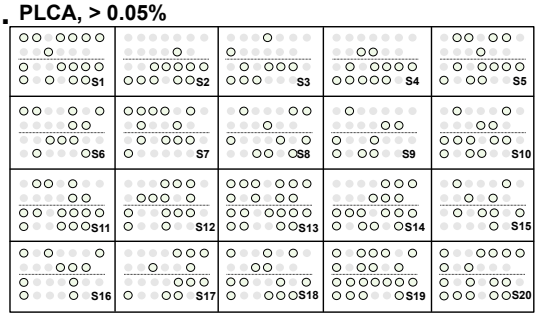
C. **PLCA, > 10.0%**



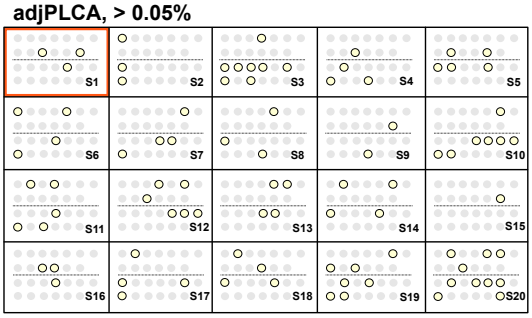
adjPLCA, > 10.0%



D. **PLCA, > 0.05%**



adjPLCA, > 0.05%



Caption follows on next page.

Figure 5.6 *(previous page)*: **A novel plate coverage algorithm determines a sample-specific plateset needed to focus culture-enriched metagenomic sequencing. A.** Pseudocode of the denovo plate coverage algorithm (PLCA) illustrates the method of subsetting a culture-enriched set into the minimal number of plates needed to cover all OTUs above a user-defined threshold. A variation on this algorithm, the adjusted PLCA, is presented in **Fig D.4**. **B.** The number of plates predicted by the denovo and adjusted PLCAs to sequence all OTUs in a single sputum sample with varying abundance cutoffs. The stacked bar chart represents the number of OTUs above the threshold (coloured bars), as well as OTUs obtained below the threshold (light yellow bars) by consequence of being present on plates within the PLCA plateset. A similar output for all sputum samples is available in **Fig D.5**. The plate subsets for the denovo and adjusted PLCA, indicated by the coloured subsets, at a high (**C**) and low (**D**) threshold indicate the importance of each media/environment pairing employed.

The PLCA can be used at user-identified thresholds of OTU relative abundance (**Fig 5.6b**). The denovo PLCA can provide a subset of plates which includes all OTUs greater than a provided relative abundance in the culture-enriched sequencing. In the adjusted PLCA, a second threshold determines the desired depth within the culture-independent sputum sample (**Fig D.4**). Using different thresholds results in a different number of plates and OTUs recovered (both those above the threshold, which must be included, and those below the threshold which are included by consequence of being present on a plate which is part of the optimal plateset) (**Fig 5.6b**). By examining the results of using the denovo PLCA at the upper and lower thresholds of 10% and 0.05% across the 20 sputum samples within this study (**Fig 5.6c-d, PLCA**), we see that the plates needed to capture such diversity varies across samples. Examining the results of applying the adjusted PLCA at similar thresholds, we observe that less plates are needed in order to obtain the diversity within the sputum sample, consistent with the enrichment of diversity with culture-enriched methods (**Fig. 5.6c-d, adjPLCA**). Further, both aerobic and anaerobic environments are essential; specifically, every media/environment pairings employed as part

of this culture-enriched sequencing approach is necessary in at least one sample (**Fig 5.6c-d**).

5.4.5 Culture-enriched metagenomic sequencing provides similar bacterial taxonomic classifications as 16S rRNA gene sequencing

Next, we displayed the capacity of these algorithms by employing the upper (10.0%) and lower (0.05%) thresholds to the denovo and adjusted PLCA in the first sputum sample in our collection (**Fig 5.6c-d, orange outlines**). This involved metagenomic sequencing of 5 and 3 culture plates, respectfully, in comparison with the original sputum sample. In order to verify our metagenomic sequencing techniques, we first extracted 16S rRNA gene reads from the metagenomic sequencing and compared these to the corresponding 16S rRNA amplicon data (**Fig 5.7**). High concordance of these sequencing methods is observed between the samples in both taxonomic summaries and PCoA analyses (**Fig 5.7**), indicating the reproducibility of these results. The CHOC plate sequenced as part of the adjusted PLCA plateset is an exception to this observation, seeing an expansion of the *Lachnospiraceae* family in the metagenomic sequencing and a decrease in *Veillonella spp.* (**Fig 5.7c**).

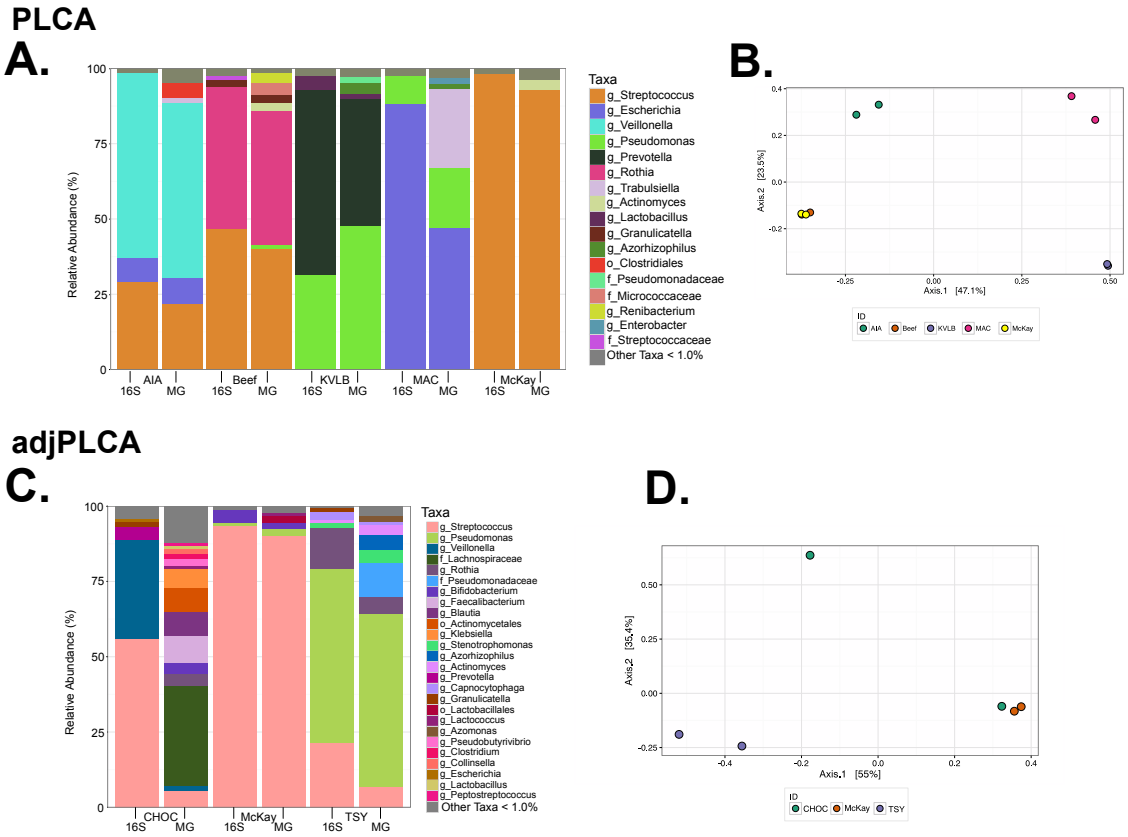
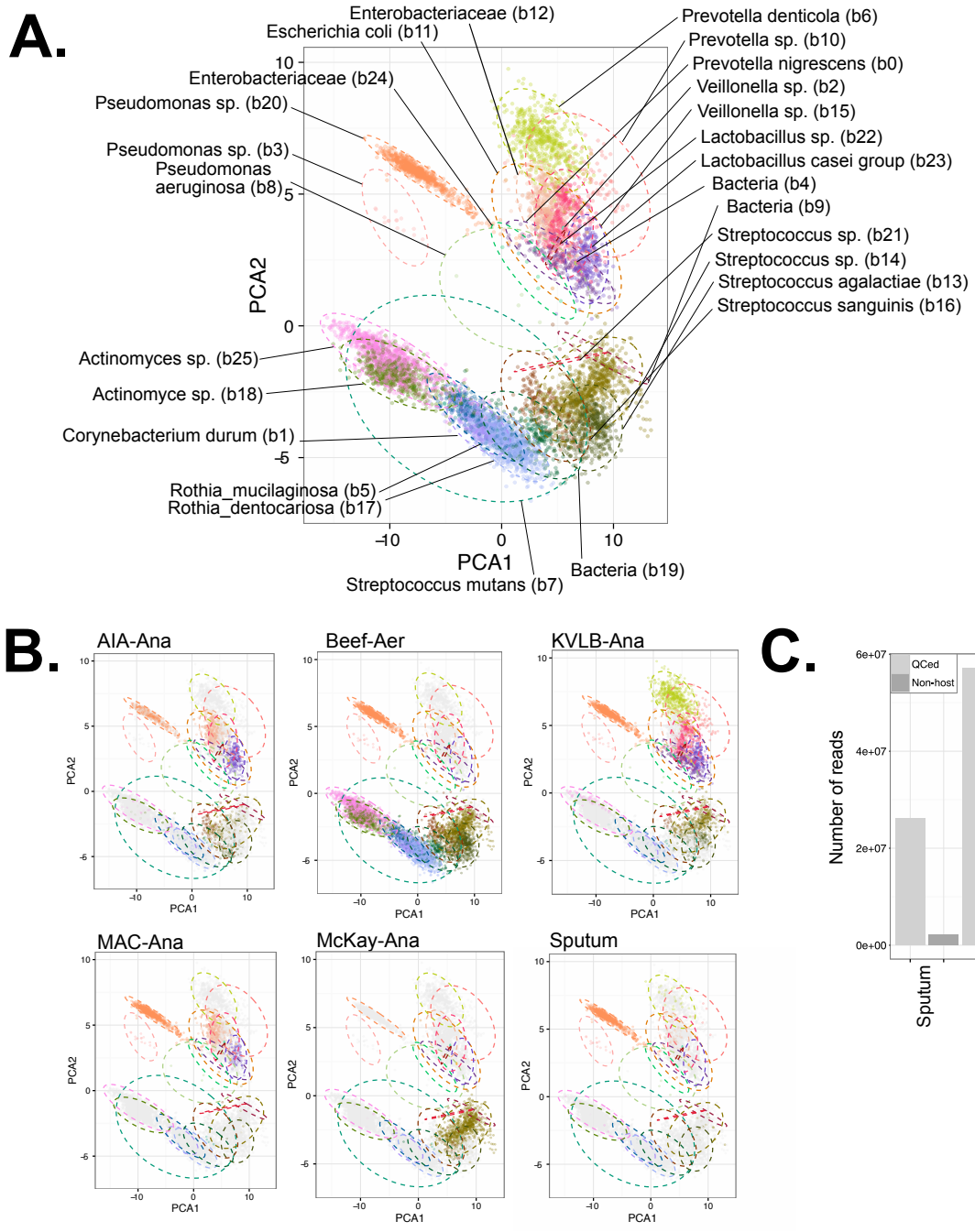


Figure 5.7: **Metagenomic sequencing reveals similar bacterial communities to 16S rRNA gene sequencing.** Comparisons of the bacterial composition of amplicon 16S rRNA gene communities to the 16S rRNA gene sequences obtained via whole-genome metagenomics reveal similar communities in the culture conditions amplified as part of the PLCA (**A-B**) and adjPLCA platesets. (**C-D**). Communities are compared visually using taxonomic summaries (**A,C**) and quantitatively using Principal Coordinates Analysis (**B,D**).

5.4.6 Culture-enriched metagenomic sequencing provides greater bacterial diversity when compared to metagenomic sequencing of the sputum sample directly.

With this verification complete, we conducted a co-assembly of the metagenomic reads from each plateset into a set of contigs. The denovo PLCA contig set consisted of 79,160 contigs with an mean length of 752.4bp (range 100-708,587bp) (adjusted PLCA: 138,871 contigs, mean length 611.2bp, range 100-653,474bp). Further, we binned the resulting contigs using available methods. CONCOCT is one commonly used approach for binning contigs based on composition and coverage in the input samples (Alneberg *et al.*, 2014). Binning the denovo PLCA plateset with CONCOCT identified 26 bins (**Fig 5.8a**). A Principal Components Analysis (PCA) of the contigs, taking composition and coverage across the plateset into account, shows some bins clustering away from others, such as bin (b) 20, as well as other bins which have a sizeable amount of overlap, such as b13, 14, 16, and 21 (**Fig 5.8a**). Upon assigning taxonomy to the contigs within each bin, we identify that bins that overlap in PCA space, such as those of *Streptococcus* and *Prevotella spp.*, belong to the same phylogenetic groups, whereas others are more taxonomically distinct. Importantly, some bins contained small numbers of contigs for which a taxonomic identity could not be defined, perhaps indicating mis-assembly of reads into chimeric contigs.



Caption follows on next page.

Figure 5.8 *(previous page)*: **Binning of culture-enriched metagenomic contigs reveals the diversity of this approach when compared to sputum metagenomics. A.** Using CONCOCT, culture-enriched metagenomic reads from 5 plates amplified according to the PLCA were grouped into 26 bins. These bins are displayed here as a PCoA in which each dot represents a contig coloured according to its bin classification. A greater diversity of organisms were obtained via culture-enriched approaches as evident by separating each contig based on which sample they originated from (colour) compared to which were not identified in a given sample (grey) (**B**); culture-enrichment contributes to the biological binning of such organisms. Further, a greater number of bacterial reads were obtained from culture-enriched sequencing (**C**) when compared to direct sputum sequencing due to host-contamination in sputum metagenomics. The results of binning the 3 plates sequenced in accordance with the adjPLCA is shown in **Figure D.6**.

Because of the ability of culture-enrichment to "biologically bin" organisms based on selective media and the varied growth requirements of different bacteria, we can use our knowledge of biological growth patterns to compare the contig bin identifications across culture conditions. **Figure 5.8b** divides the co-assembly to a per-plate basis, displaying which contigs were constructed using reads from each plate. For example, MacConkey agar, a known gram-negative selection media (MacConkey, 1905), contributed reads which almost exclusively were assigned to a bin identified taxonomically as *Pseudomonas spp.* Similarly, non-selective agar mediums such as Beef agar contributed contigs from a variety of bacterial species (**Fig 5.8b**).

Importantly, the diversity identified in the co-assembly of the denovo PLCA plateset is not recapitulated in the sputum sample, which contains only contigs from *Pseudomonas spp.* (**Fig 5.8b**). Although this sputum sample was sequenced to a depth 50% deeper than the average plate pool sample (**Sup Table D.2**), the majority of these aligned to the human genome and thus were subsequently culled (**Fig 5.8c**).

Similar results were obtained with the adjPLCA plateset (**Fig D.6a**). Binning of this plateset with CONCOCT resulted in 28 bins, only 1 of which (taxonomically assigned to *Pseudomonas*) was identified in the sputum sample (**Fig D.6b**).

5.4.7 The biological implications of culture-enriched metagenomic sequencing differ depending on binning strategy

Recently, MaxBin (Wu *et al.*, 2015) was identified as a more biologically-relevant binning approach in an exhaustive benchmarking study of metagenomic software (Sczyrba *et al.*, 2017). Applying this binning strategy to our co-assembled denovo PLCA contig set resulted in 12 bins, 14 less than the binning generated with CONCOCT; similarly, binning on the adjusted PLCA contig set generated 17 bins versus CONCOCT's 28. In order to validate which approach was more biologically meaningful, we compared the rank abundance curves of the maximum abundance of OTUs from the 16S rRNA gene sequencing across the culture-enriched dataset (**Fig 5.9a**) and the sputum sequencing (**Fig 5.9b**) to the metagenomic bins from denovo PLCA and adjusted PLCA, respectively. MaxBin only identified 5 of the 10 expected OTUs above the 10.0% threshold employed with denovo PLCA; however, CONCOCT identified all 10. Similarly, CONCOCT identified 11 of the 13 OTUs expected above the 0.05% threshold used in the adjusted PLCA algorithm whereas MaxBin identified only 8 (**Fig 5.9b**).

Further statistical comparisons of the two approaches identify how different their interpretations of the data are (**Fig 5.9c-d**). The number and concatenated length of the contigs across the MaxBin bins generated for the PLCA plateset are very evenly distributed (**Fig 5.9c**), with the expectation for a large number of short reads which were not binned (**Fig 5.9c, Unb**). Further, there is an average of 75.1% percent coverage of each bin to the closest reference genome (range: 34.2%-93.2%) (**Fig 5.9d**). This is in contrast to CONCOCT which has a very uneven distribution of contigs and their concatenated lengths across the bins (**Fig 5.9c**). This is reflected

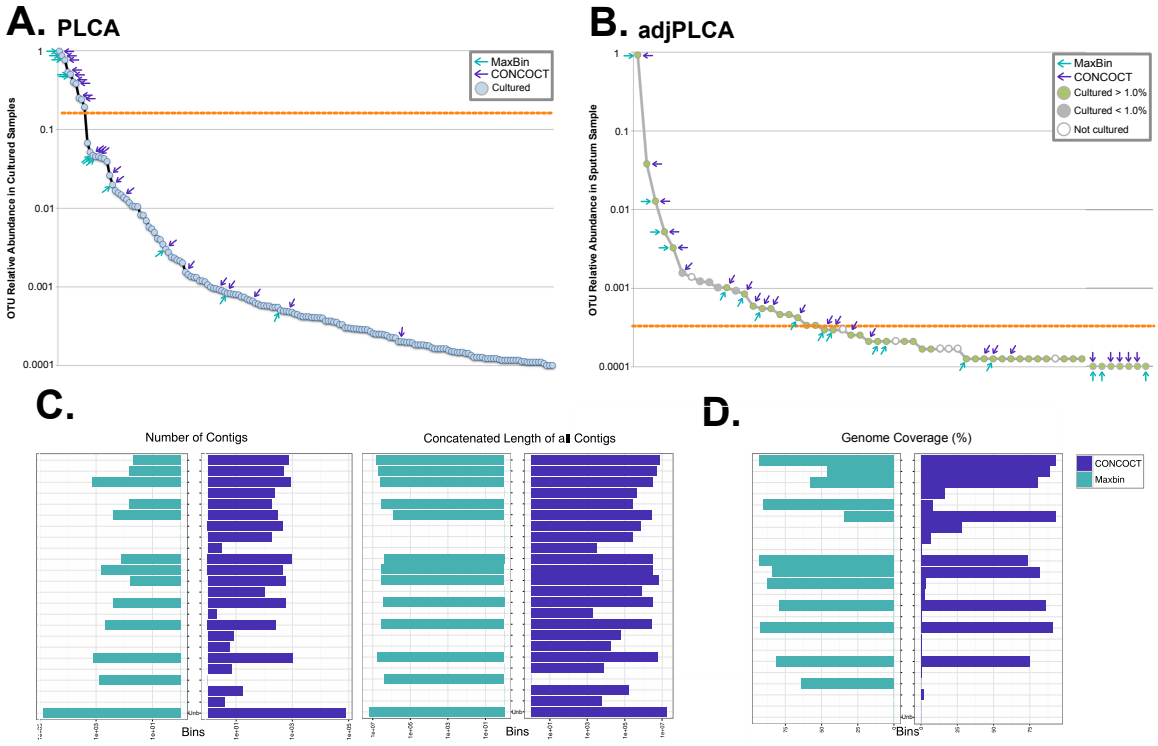


Figure 5.9: **The results of culture-enriched metagenomic sequencing are dependent on processing techniques.** Both CONCOCT (purple arrows) and MaxBin (blue arrows) were used to bin contig sets generated from the denovo PLCA and adjusted PLCA algorithms. Bins were assigned a taxonomy and compared to OTUs which were targeted with PLCA (OTUs above the orange dotted line) in the cultured OTU (**A**) and sputum OTU (**B**) datasets. CONCOCT consistently created bins which more readily matched the 16S rRNA gene sequencing and identified all OTUs targeted by the PLCA algorithm and 11 of the 13 expected OTUs in the sputum sample. Differences between MaxBin and CONCOCT bins, ordered here by rank abundance curve, can be seen in the distribution of contigs and their lengths among the bins (**C**) as well as the percent genome coverage of the closest reference genome (**D**).

in the 15 bins who have a percent coverage of $\leq 10\%$ compared to the closest reference genome.

The high taxonomic similarity between 16S rRNA gene sequencing with CONCOCT versus MaxBin binning suggest that CONCOCT has more biologically accurate results; however, the binning statistics indicate inconsistencies in CONCOCT bin content which are not present in the MaxBin output. These results make any biological interpretation of these data difficult to elucidate.

5.5 Discussion

The decrease in cost and increase in massively parallelized sequencing technology has revolutionized the way that the research community studies the human microbiota. Our understanding of microbial communities and how they relate to health and a wide variety of diseases and disorders are still being elucidated using these approaches. The power of next generation sequencing technology cannot be argued with; however, in this study, we show that complementing culture-independent approaches with culture enrichment can increase our understanding of human-associated communities, in particular those of the lower respiratory tract.

In this study, we show that the majority of the CF lung microbiota is culturable. More specifically, an average of 81% of OTUs identified in sputum were recovered by culture across a 20 sample dataset; these OTUs represented an average of 99.15% of the relative abundance of organisms within these sputum samples (**Fig 5.2**). These findings follow from the pinnacle results of Sibley *et al.* who, using T-RFLP and 454 sequencing were also able to identify a culturable majority within the CF lung microbiota (Sibley *et al.*, 2011). These results are unsurprising even though the

human microbiota is often described as unculturable; conventional CF pathogens, including *Pseudomonas aeruginosa, Staphylococcus aureus, Haemophilus influenzae*, and *Burkholderia cepacia* complex (Surette, 2014) are all readily culturable organisms, which has contributed to the history of their treatment in CF lung disease (Sibley *et al.*, 2011). Further, the top OTUs identified in microbiota studies of the CF airways, including *Streptococcus, Prevotella, Veillonella,* and *Rothia*, all have a strong culturable history (Shah and Collins, 1990; Sibley *et al.*, 2011; Gronow *et al.*, 2010; Georg and Brown, 1967). There were a few organisms which were consistently found to be unculturable across this 20 sample dataset. Many of these organisms, including those of the *Spirochaetes*, and *TM7* phyla consist of organisms which are notoriously difficult to culture (Chi *et al.*, 1999; Marcy *et al.*, 2007). Further, none of these phyla have been previously identified as common organisms in CF lung disease and thus were not targeted in our culture-enrichment set of 13 media.

An increase in OTU diversity (**Fig 5.2**) and number (**Fig 5.3**) was obtained from culture-dependent sequencing approaches when compared to 16S rRNA gene sequencing of the sputum sample alone. These results indicate the advantage of culture. General growth, high-nutrient media ensure that a wide-range of organisms are able to grow in different niches across the plate surface area. Selective media allow for the proliferation of low abundance organisms which may not be identified via direct sequencing approaches. Anaerobic culture alone increases the number of OTUs identified by 54%. When we consider the density of the human microbiota, it is not surprising that low abundance organisms are missed by conventional sequencing; it is not uncommon, for example, for the cystic fibrosis lung microbiota to reach a density of 10^8 CFUs/mL (Meyer *et al.*, 1997; Stressmann *et al.*, 2011b). If marker

gene sequencing produces 50,000 reads per sample, an organism identified by a single read would equate to 0.002% relative abundance or $2x10^4$CFUs/mL. Thus, many organisms present within this microbial community at a low abundance could easily be missed using culture-independent methods and conventional sequencing depths. Further, anaerobic bacteria are known to occupy niches in the gastrointestinal tract (Savage, 1977), vagina (Bartlett *et al.*, 1977), and cystic fibrosis lung microbiota (Tunney *et al.*, 2008), explaining the increased diversity in OTUs identified from media incubated under these conditions.

The plate coverage algorithms (PLCA) described within represents a way to use 16S rRNA (or other) marker gene analysis as a way to minimize the number of plates that need to be interrogated with metagenomic sequencing in order to recapitulate the functional diversity of the totality of the environment (denovo PLCA) or the diversity as a function of the abundance within the originating sputum sample (adjusted PLCA). These approaches, utilized here at the highest (10%) and lowest (0.05%) thresholds available, were able to recapitulate the majority of targeted OTUs, dependent on the metagenomic binning approaches used (**Fig 5.9**). A fraction of the OTUs missed by the adjusted PLCA algorithm with a threshold of 0.05% were not enriched for in the culture above the secondary threshold of 1% (**Fig 5.9a, grey dots**). The use of $\geq$ 1% relative abundance as a secondary threshold is an estimate based on a number of metagenomic sequencing studies which have assembled draft genomes (Sangwan *et al.*, 2016; Hugerth *et al.*, 2015; Nielsen *et al.*, 2014). Future attempts at this enrichment could correct for this by (a) expanding the media selection to include more selective media which would allow for the proliferation, and thus greater relative abundance, of these microbes, or (b) employ deeper metagenomic sequencing

then the average read depth of 14,688,868 reads used within this study.

When the taxonomic diversity of the 16S rRNA gene within the metagenomic sequencing was compared to the amplicon data, the taxonomic assignment and diversity were comparable across both the denovo and adjusted PLCA platesets (**Fig 5.7**) with the exception of an expansion of *Lachnospiraceae* on the anaerobic CHOC plate employed. Interestingly, when these data were subsequently binned and compared to the closest reference genomes, no representatives of the *Lachnospiraceae* family were observed. This indicates a potential mis-alignment of metagenomic sequences to the 16S rRNA gene reference database or to the existence of chimeric contigs.

Culture-enriched metagenomic sequencing reads were co-assembled, and binned into groups whose taxonomy closely mirrored that of the 16S rRNA gene sequencing results (**Fig 5.9a-b**). However, the most appropriate bioinformatic workflow for dealing with these reads has not been established. This is evident in the drastically different results that were obtained by two commonly used binning methods. Most troubling are the differences between percent coverage of the closest reference genome between MaxBin and CONCOCT. CONCOCT's binning process appears to create spurious OTUs containing < 10 contigs with an average concatenated length of 15,358bp; these contigs do not map well to NCBI's RefSeq database and do not provide high coverage of the closest reference genome. However, this may not be an issue with CONCOCT and the binning process; instead, CONCOCT could be identifying chimeric contigs that have been created as part of the metagenomic assembly across the plateset, and, given that their composition would not match any other set of contigs, are being binned independently of others. Further benchmarking with real data is necessary to compare metagenomic assemblers and binning techniques which

can provide more biologically relevant interpretations of metagenomic data before conclusive insights can be drawn.

In conclusion, the CF lung microbiota is a culturable majority. The combination of culture-enrichment with commonly used culture-independent approaches enhances the diversity of either approach alone and, baring future improvements in metagenomic bioinformatic tools, will provide greater insight into human-associated microbial communities.

5.6 Acknowledgements

The authors would like to thank the patients and health care professionals at the Calgary Adult CF Clinic for their participation and assistance with this study.

Chapter 6

Conclusions[1]

Within this body of work, I present the use of culture-independent and -dependent methods to study the human respiratory tract in health and compare it to the effects of age and disease. Together with the co-authors of these studies, I have shown how next generation sequencing can contribute to our understanding of the microbial communities within the anterior nares, oropharynx, and lung. After establishing the most biologically accurate bioinformatic tools for the processing of 16S rRNA gene sequencing data (**Chapter 2**), we used this approach to study the upper respiratory tract in elderly individuals and compared these to publicly available data on mid-aged adults (**Chapter 3**). In this study, we concluded that these communities differ drastically as we age in terms of membership and diversity. Next, we applied these approaches to the study of the lower respiratory tract in cystic fibrosis (CF) (**Chapter 4**). Participants longitudinally collected sputum samples 3 times a week

[1]Some of the opinions and ideas presented within this Chapter have been previously published as Whelan FJ & Surette MG (2015). Clinical Insights into Pulmonary Exacerbations in Cystic Fibrosis from the Microbiome: What Are We Missing? *Ann Am Thorac Soc* **12**(Supp2) S207-S211.

for a year in order to test whether these communities are altered preceding the onset of pulmonary exacerbation (PE). In this study, we concluded that this was not the case and that no consistent disturbances to this community were observed preceding PE. Following this conclusion, we wished to devise methods which could be used to interrogate these communities in greater detail. As such, we built on the culture-enrichment work of Sibley *et al.* (2011) to study the diversity of the CF lung microbiota and to follow up this culture with metagenomic sequencing (**Chapter 5**). This methodology allows for the recovery of an increased number of organisms while mitigating the contamination of host DNA, a problematic contaminant in the direct sequencing of sputum. While we concluded that the interpretation of metagenomic sequencing results is software-dependent and problematic, this methodology paves the way for future studies of the CF lung microbiota.

The general, over-arching hypothesis of my thesis research was that through next generation sequencing technologies, we can study the totality of the microbial communities in the upper and lower airways, allowing a better understanding of health and disease. Specifically, I generated and tested two specific hypotheses; in the first, I hypothesized that the microbial communities of the URT of elderly individuals is altered in comparison to adults, making these individuals more susceptible to infection. By collecting nasal and oral samples from a set of elderly individuals, we were able to show that these communities differ statistically from data of publicly available mid-aged adults; this result answers the first half of this hypothesis. However, we were unable to show a causal link between these alterations and the elderly's increased risk for respiratory infection. We did identify an increase in *Streptococcus*

spp. in these environments; however, PCR specific for *Streptococcus pneumoniae*, the leading causative agent of pneumonia, did not identify an increase in carriage within this population. Thus, I was unable to definitively answer the second part of this hypothesis from the results of this study.

The conclusions which we obtained from this study of the elderly URT open up many avenues for future research. The nasopharynx is the only community sampled along the respiratory tract whose composition is unique (Charlson *et al.*, 2011). It could be hypothesized that this uniqueness contributes to this community acting as a barrier between the environment and the rest of the respiratory tract. In this study, we discovered that the anterior nares community enters a dysbiotic state as we age. Taking this conclusion together with the fact that elderly are at higher risk of respiratory infections, it could be suggested that this change in the microbiota is crucial to the health of the elderly. Alternatively, as we age the epithelial cells in our nasal passages develop a new physiology and dry out (Lindemann *et al.*, 2008); an alternative hypothesis could be posed that the alterations in the nasal microbiota that we observed in this study are unrelated to the increased risk of infection in the elderly but instead is the effect of a physiological change to this environment.

It should be noted that culture-enrichment of the elderly URT microbiota would be predicted to increase recovered diversity as has been shown in the CF lung. However, because there were few prior investigations of this community before we conducted the study presented in **Chapter 3** from which we could devise a culture-enrichment scheme, and samples had already been collected (in an aerobic environment), we

decided to address this hypothesis using culture-independent approaches. Applying culture-enrichment to this environment could aid in answering a few important questions. The PCR reactions for the nasal swabs that were collected for this study were difficult to perform; in our experience this often means that these samples had very small amounts of total DNA or have large amounts of non-bacterial DNA. Culture-enrichment of these communities in parallel to those of healthy mid-aged adult nasal communities would not only provide us with a deeper resolution as to the membership of these microbiota, but would also allow us to compare the bacterial load in this locale and how it changes with age. Perhaps respiratory infections are more able to establish in the elderly simply because colonization, which is necessary to preclude infection, is easier given the greater available real estate.

Further, we mention in the text of this study (see Discussion of **Chapter 3**) that our investigations were limited to elderly which resided in nursing homes within close geographical proximity to each other. Future studies of the elderly URT should include community dwelling elderly as the underlying differences in health of those individuals who are able to live alone versus those in need of assisted living may influence these results. In addition, these results may help further understand the physiological differences between these populations. Further, it would be interesting to compare individuals from different parts of the country (or world) to each other. Studies in the cystic fibrosis lung have identified geographical differences in the lung microbiota (Stressmann *et al.*, 2011a), and seasonal differences have already been identified in the nasal cavity (Bogaert *et al.*, 2011) indicating that the external environment plays a large role in the colonization of these locales.

As in all studies of the microbiota, it is important to consider the cause and effect relationship between the URT and respiratory infection in the elderly. It could be the case that dysbiotic regulation of this community contributes to a loss in an environmental barrier which then allows for increased rate of infection. However, it just as equally could be the case that changes to the immune system effect how pathogens such as *Streptococcus pneumoniae* are dealt with as well as effecting the microbiota of the nose. If the results of future studies are able to solidify a causative effect of change in the microbiota of the URT in the elderly on incidence of respiratory infection, these results could help mitigate these changes. Some sort of nasal probiotic could be used to help maintain these communities in their adult-like state; however these therapeutics would be useless if these alterations were simply an effect of larger, body-wide dynamics.

My second specific hypothesis related to the cystic fibrosis lung microbiota. Specifically, I hypothesized that using culture-dependent methods in conjunction with culture-independent advancements in sequencing technologies would improve the taxonomic and functional resolution of the CF lung microbiota in order to elucidate microbial processes within the CF lung which contributes to the onset of PE. We conducted two separate investigations in order to test this hypothesis. The first involved, longitudinal sampling of the CF lung microbiota, which did not indicate taxonomic changes within these communities that could be consistently recognized during the onset of PE; this study suggested that our hypothesis would be rejected on the grounds of having found no differences in these communities at the resolution

of 16S rRNA gene sequencing. However, our investigations using culture-enrichment and metagenomic sequencing identify a surplus of organisms which are not identified via 16S rRNA gene sequencing of sputum alone, identifying that these microbial processes within the lung community may have been missed in our longitudinal study. Unfortunately, conflicting results from bioinformatic processing of the metagenomic sequencing forced us to spend more time benchmarking new software rather than being able to fully elucidate the biological implications of these communities. While it has been exciting being on the forefront of these technological advancements, it has also been a frustrating process. Future studies which follow the next wave of improvements in metagenomic sequencing technologies will be needed in order to fully answer this hypothesis.

Our studies of the cystic fibrosis lung microbiota aimed to get at the root of the cause and effect issue of microbiota studies by collecting sputum samples prior to onset of symptoms related to PE. In most studies, samples are collected once treatment for PE has already begun, meaning that any microbial changes in these communities which could be initiating such an event would have already occurred and antibiotic treatment, which would greatly affect these communities, has begun. Through the amazing collaboration that our laboratory has had with the Alberta Adult Cystic Fibrosis Clinic, we were able to obtain sputum at hospital admission and before escalation of antibiotic therapy for treatment of PE.

Our use of 16S rRNA gene sequencing technologies to longitudinally study the CF

lung microbiota in 6 Participants concluded that there were no consistent, identifying changes in these communities preceding PE onset. While disappointing that this study's conclusions included negative results, I think this and other similar studies (Carmody *et al.*, 2015; Cuthbertson *et al.*, 2015) are important to keep in mind as this field progresses.

Dr. Surette and I have theorized that PEs may be brought on by a small, active community within the lung that constitutes a minority of the sputum sample being profiled (Whelan and Surette, 2015). In this model, the microbiota of the CF lung during stability would signify the "stable state" community; upon PE, a small active community may break away from this stable community and change in composition, function, or virulence in a manner that activates an immune response. However, because the sputum sample represents multiple areas of the lungs, this small active community may not be distinguishable from sequencing noise (**Fig 6.1**).

Unfortunately, it is difficult to test this theory directly using 16S rRNA gene sequencing technologies as the least invasive way of sampling this community is via the expectorated sputum that patients produce during physical therapy. More invasive procedures such as BALs similarly would not be able to distinguish stable from active communities. One possibility would be to separate the heterogeneous chunks present within an expectorated sputum sample and isolate the DNA separately from each; however, even if this approach was able to distinguish populations, there would be no way of knowing which population corresponded to the active driver of PE.

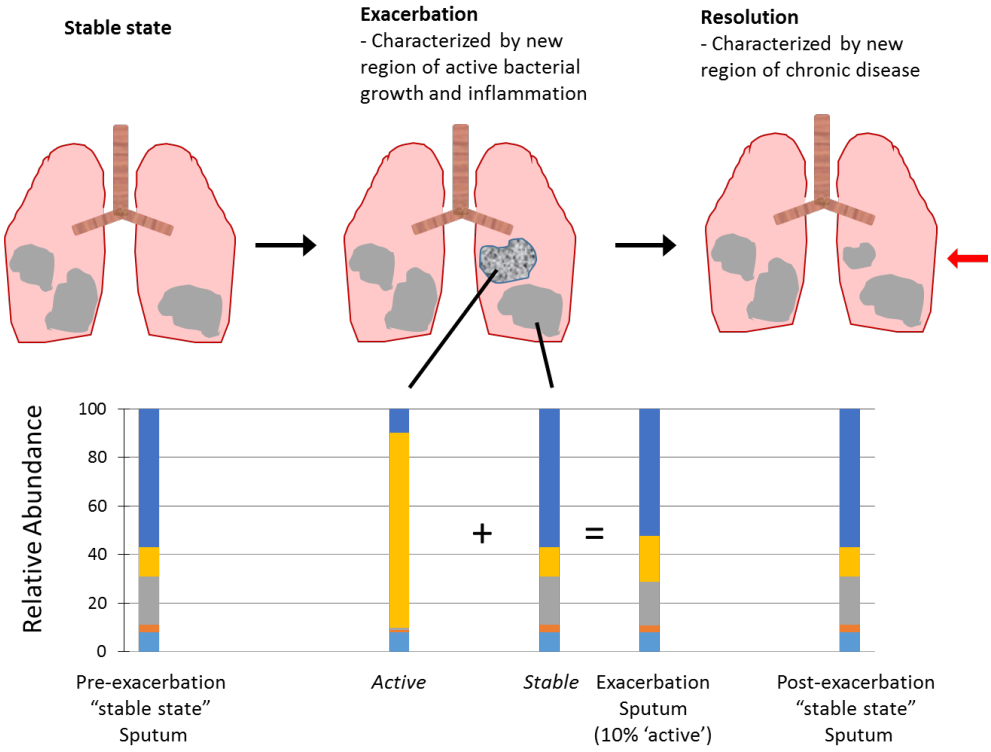


Figure 6.1: **A simple, hypothetical model of microbiota distribution in stable and exacerbating lungs.** In this model, sputum produced during an exacerbation is a combination of the stable subpopulation present before exacerbation onset and the active subpopulation driving disease. Because this active community may be only a small proportion of the total sample, it is difficult to elucidate using molecular profiling techniques. Each colour represents a different bacterial species. This figure and accompanying label originally appeared in Whelan and Surette (2015) and has been reproduced here with permission. Copyright © the American Thoracic Society.

Luckily, the answer may lie in other technologies. Sequencing of the bacterial RNA within the lung during stability and exacerbation should be able to distinguish these populations from each other; in order to encourage an immune response, it could be presumed that the active population is proliferating and transcribing DNA to eventually translate into proteins. Conducting bacterial RNA sequencing of sputum immediately after production would be an interesting future study. This technology has been used in a few previous studies with limited results; similar to issues with shotgun metagenomic sequencing, RNA sequencing of sputum samples would be plagued with host RNA. In fact, we did try this approach with a sputum sample collected as part of the culture-enrichment study and identified $\geq$97% host RNA in our results (data not shown). Unlike metagenomic sequencing which we can use culture-enrichment to biologically diversify and rid of host contamination, bacterial RNA sequencing must be done on the sputum sample directly in order to best capture the signature of this active population. Continual improvements in sequence capability and reductions in sequencing costs should soon make it feasible to sequence these populations at great depth, to extract the bacterial RNA *in silico*. These results would determine whether the stable/active population model of the CF lung microbiota which we propose is supported.

Answering both of these specific hypotheses aids in answering the more general question of this thesis which is whether these approaches are viable options for gaining a better understanding of the respiratory tract in health and disease. I believe that the data and results presented within support this hypothesis. Our research made great strides for the understanding of the elderly URT, and outlined bioinformatic

tools and culturing approaches necessary for the use of culture to inform metagenomic sequencing. However, in general, I believe that microbiota studies have to better understand the limitations of next generation sequencing approaches. Without these methods being used in conjunction with others, it is impossible to know the actual load of bacteria within an environment, and whether the DNA being sequenced is from viable organisms. The samples that we use to study these communities are a proxy; geographical information is lost and hints at the inhibition or synergistic relationships between organisms cannot be inferred. Instead, the field of microbiome research needs to combine the power of next generation sequencing with microbiology to better understand the dynamics within these communities.

Ultimately, what is most important is how these studies of the respiratory microbiota can affect patient care. There is an element of self-interest and basic science to wanting to understand the dynamics and ecology of these communities, but personally I am more interested in seeing these results translated into clinical care. I think that the research presented within has laid the groundwork for such translations and that future research in these areas, as outlined above, can truly help us understand and treat implications of age and disease in the upper and lower respiratory tracts.

Chapter 7

Bibliography

Aaron, S. D., Ramotar, K., Ferris, W., Vandemheen, K., Saginur, R., Tullis, E., Haase, D., Kottachchi, D., St Denis, M., and Chan, F. (2004). Adult cystic fibrosis exacerbations and new strains of *Pseudomonas aeruginosa*. *Am. J. Respir. Critical Care Medicine*, **169**(7), 811–5.

Accurso, F. J., Rowe, S. M., Clancy, J., Boyle, M. P., Dunitz, J. M., Durie, P. R., Sagel, S. D., Hornick, D. B., Konstan, M. W., Donaldson, S. H., Moss, R. B., Pilewski, J. M., Rubenstein, R. C., Uluer, A. Z., Aitken, M. L., Freedman, S. D., Rose, L. M., Mayer-Hamblett, N., Dong, Q., Zha, J., Stone, A. J., Olson, E. R., Ordoñez, C. L., Campbell, P. W., Ashlock, M. A., and Ramsey, B. W. (2010). Effect of VX-770 in Persons with Cystic Fibrosis and the G551D- *CFTR* Mutation. *New Engl. J. Medicine*, **363**(21), 1991–2003.

Ahmed, R., Oldstone, M. B. A., and Palese, P. (2007). Protective immunity and susceptibility to infectious diseases: lessons from the 1918 influenza pandemic. *Nat. Immunol.*, **8**(11), 1188–93.

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F., and Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nat. Methods*, **11**(11), 1144–6.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, **215**(3), 403–410.

Andersen, D. (1938). Cystic fibrosis of the pancreas and its relation to celiac disease. *Am. J. Dis. Child.*, **56**(2), 344.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Apprill, A., McNally, S., Parsons, R., and Weber, L. (2015). Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquatic Microb. Ecol.*, **75**(2), 129–137.

Armstrong, D. S., Grimwood, K., Carlin, J. B., Carzino, R., Gutierrez, J. P., Hull, J., Olinsky, A., Phelan, E. M., Robertson, C. F., and Phelan, P. D. (1997). Lower Airway Inflammation in Infants and Young Children with Cystic Fibrosis. *Am. J. Respir. Critical Care Medicine*, **156**(4), 1197–1204.

Arrieta, M., Stiemsma, L., Dimitriu, P., Thorson, L., Russell, S., Yurist-Doutsch, S., Kuzeljevic, B., Gold, M., Britton, H., Lefebvre, D., Subbarao, P., Mandhane, P., Becker, A., McNagny, K., Sears, M., Kollmann, T., Mohn, W., Turvey, S., and Brett Finlay, B. (2015). Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Sci. Transl. Medicine*, **7**(307), 307ra152–307ra152.

Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C., and Segata, N. (2015). Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ*, **3**, e1029.

Bacci, G., Paganin, P., Lopez, L., Vanni, C., Dalmastri, C., Cantale, C., Daddiego, L., Perrotta, G., Dolce, D., Morelli, P., Tuccio, V., De Alessandri, A., Fiscarelli, E. V., Taccetti, G., Lucidi, V., Bevivino, A., and Mengoni, A. (2016). Pyrosequencing Unveils Cystic Fibrosis Lung Microbiome Differences Associated with a Severe Lung Function Decline. *PLOS ONE*, **11**(6), e0156807.

Barnett, D. W., Garrison, E. K., Quinlan, A. R., Stromberg, M. P., and Marth, G. T. (2011). BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics.*, **27**(12), 1691–1692.

Barriuso, J., Valverde, J. R., and Mellado, R. P. (2011). Estimation of bacterial diversity using next generation sequencing of 16S rDNA: a comparison of different workflows. *BMC Bioinformatics.*, **12**(1), 473.

Bartlett, J. G., Onderdonk, A. B., Drude, E., Goldstein, C., Anderka, M., Alpert, S., and McCormack, W. M. (1977). Quantitative Bacteriology of the Vaginal Flora. *J. Infect. Dis.*, **136**(2), 271–277.

Bartram, A. K., Lynch, M. D. J., Stearns, J. C., Moreno-Hagelsieb, G., and Neufeld, J. D. (2011). Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end illumina reads. *Appl. Environ. Microbiol.*, **77**(11), 3846–52.

Bassis, C. M., Tang, A. L., Young, V. B., and Pynnonen, M. A. (2014). The nasal cavity microbiota of healthy adults. *Microbiome*, **2**, 27.

Beck, J. M., Young, V. B., and Huffnagle, G. B. (2012). The microbiome of the lung. *Transl. Res.*, **160**(4), 258–266.

Berkebile, A. R., McCray, P. B., and Jr. (2014). Effects of airway surface liquid pH on host defense in cystic fibrosis. *The Int. J. Biochem. & Cell Biol.*, **52**, 124–9.

Berry, A., DeVault, J. D., and Chakrabarty, A. M. (1989). High osmolarity is a signal for enhanced algD transcription in mucoid and nonmucoid *Pseudomonas aeruginosa* strains. *J. Bacteriol.*, **171**(5), 2312–7.

Bilton, D., Pye, A., Johnson, M. M., Mitchell, J. L., Dodd, M., Webb, A. K., Stockley, R. A., and Hill, S. L. (1995). The isolation and characterization of non-typeable *Haemophilus influenzae* from the sputum of adult cystic fibrosis patients. *The Eur. Respir. J.*, **8**(6), 948–53.

Bittar, F. and Rolain, J.-M. (2010). Detection and accurate identification of new or emerging bacteria in cystic fibrosis patients. *Clin. Microbiol. Infect.*, **16**(7), 809–820.

Bogaert, D., De Groot, R., and Hermans, P. W. M. (2004). *Streptococcus pneumoniae* colonisation: the key to pneumococcal disease. *The Lancet. Infect. Dis.*, **4**(3), 144–54.

Bogaert, D., Keijser, B., Huse, S., Rossen, J., Veenhoven, R., van Gils, E., Bruin, J., Montijn, R., Bonten, M., and Sanders, E. (2011). Variability and Diversity of

Nasopharyngeal Microbiota in Children: A Metagenomic Analysis. *PLOS ONE*, **6**(2), e17035.

Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F., and Corbeil, J. (2012). Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol.*, **13**(12), R122.

Boyle, M. P., Bell, S. C., Konstan, M. W., McColley, S. A., Rowe, S. M., Rietschel, E., Huang, X., Waltz, D., Patel, N. R., Rodman, D., and VX09-809-102 Study Group (2014). A CFTR corrector (lumacaftor) and a CFTR potentiator (ivacaftor) for treatment of patients with cystic fibrosis who have a phe508del CFTR mutation: a phase 2 randomised controlled trial. *The Lancet Respir. Medicine*, **2**(7), 527–538.

Browne, H. P., Forster, S. C., Anonye, B. O., Kumar, N., Neville, B. A., Stares, M. D., Goulding, D., and Lawley, T. D. (2016). Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nat.*, **533**, 543–546.

Burns, J., Gibson, R., McNamara, S., Yim, D., Emerson, J., Rosenfeld, M., Hiatt, P., McCoy, K., Castile, R., Smith, A., and Ramsey, B. (2001). Longitudinal Assessment of *Pseudomonas aeruginosa* in Young Children with Cystic Fibrosis. *The J. Infect. Dis.*, **183**(3), 444–452.

Button, B., Cai, L.-H., Ehre, C., Kesimer, M., Hill, D. B., Sheehan, J. K., Boucher, R. C., and Rubinstein, M. (2012). A Periciliary Brush Promotes the Lung Health by Separating the Mucus Layer from Airway Epithelia. *Sci.*, **337**(6097), 937–941.

Campodónico, V. L., Gadjeva, M., Paradis-Bleau, C., Uluer, A., and Pier, G. B.

(2008). Airway epithelial control of *Pseudomonas aeruginosa* infection in cystic fibrosis. *Trends Mol. Medicine*, **14**(3), 120–133.

Cantin, A. M., Hartl, D., Konstan, M. W., and Chmiel, J. F. (2015). Inflammation in cystic fibrosis lung disease: Pathogenesis and therapy. *J. Cyst. Fibros.*, **14**(4), 419–430.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., Fierer, N., and Knight, R. (2010a). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci.*, **108**(Supplement 1), 4516–22.

Caporaso, J. G., Bittinger, K., Bushman, F. D., DeSantis, T. Z., Andersen, G. L., and Knight, R. (2010b). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics.*, **26**(2), 266–7.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R. (2010c). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**(5), 335–6.

Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., Knights, D., Gajer, P., Ravel, J., Fierer, N., Gordon, J. I., and Knight, R. (2011). Moving pictures of the human microbiome. *Genome Biol.*, **12**(5), R50.

Carmody, L. A., Zhao, J., Schloss, P. D., Petrosino, J. F., Murray, S., Young, V. B., Li, J. Z., and LiPuma, J. J. (2013). Changes in Cystic Fibrosis Airway Microbiota at Pulmonary Exacerbation. *Ann Am Thorac Soc*, **10**(3), 179–187.

Carmody, L. A., Zhao, J., Kalikin, L. M., LeBar, W., Simon, R. H., Venkataraman, A., Schmidt, T. M., Abdo, Z., Schloss, P. D., and LiPuma, J. J. (2015). The daily dynamics of cystic fibrosis airway microbiota during clinical stability and at exacerbation. *Microbiome*, **3**(1), 12.

Castellani, C., Cuppens, H., Macek, M., Cassiman, J., Kerem, E., Durie, P., Tullis, E., Assael, B., Bombieri, C., Brown, A., Casals, T., Claustres, M., Cutting, G., Dequeker, E., Dodge, J., Doull, I., Farrell, P., Ferec, C., Girodon, E., Johannesson, M., Kerem, B., Knowles, M., Munck, A., Pignatti, P., Radojkovic, D., Rizzotti, P., Schwarz, M., Stuhrmann, M., Tzetis, M., Zielenski, J., and Elborn, J. (2008). Consensus on the use and interpretation of cystic fibrosis mutation analysis in clinical practice. *J. Cyst. Fibros.*, **7**(3), 179–196.

Centers for Disease Control and Prevention (CDC) (1995). Pneumonia and influenza death rates–United States, 1979-1994. *MMWR Morb Mortal Wkly Rep.*, **44**(28), 535–7.

Chao, A. (1984). Nonparametric Estimation of the Number of Classes in a Population. *Scand. J. Stat.*, **11**(4), 265–270.

Charlson, E. S., Bittinger, K., Haas, A. R., Fitzgerald, A. S., Frank, I., Yadav, A., Bushman, F. D., and Collman, R. G. (2011). Topographical continuity of bacterial populations in the healthy human respiratory tract. *Am. J. Respir. Critical Care Medicine*, **184**(8), 957–63.

Chen, T., Yu, W.-H., Izard, J., Baranova, O. V., Lakshmanan, A., and Dewhirst, F. E. (2010). The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database*, **6**(baq013).

Cheng, K., Smyth, R. L., Govan, J. R., Doherty, C., Winstanley, C., Denning, N., Heaf, D. P., van Saene, H., and Hart, C. A. (1996). Spread of β-lactam-resistant *Pseudomonas aeruginosa* in a cystic fibrosis clinic. *The Lancet*, **348**(9028), 639–642.

Chi, B., Chauhan, S., and Kuramitsu, H. (1999). Development of a system for expressing heterologous genes in the oral spirochete *Treponema denticola* and its use in expression of the *Treponema pallidum* flaA gene. *Infect. Immun.*, **67**(7), 3653–6.

Cho, I. and Blaser, M. J. (2012). The human microbiome: at the interface of health and disease. *Nat. Rev. Genet.*, **13**(4), 260–70.

Claesson, M. J., Jeffery, I. B., Conde, S., Power, S. E., O'Connor, E. M., Cusack, S., Harris, H. M. B., Coakley, M., Lakshminarayanan, B., O'Sullivan, O., Fitzgerald, G. F., Deane, J., O'Connor, M., Harnedy, N., O'Connor, K., O'Mahony, D., van Sinderen, D., Wallace, M., Brennan, L., Stanton, C., Marchesi, J. R., Fitzgerald, A. P., Shanahan, F., Hill, C., Ross, R. P., and O'Toole, P. W. (2016). Gut microbiota composition correlates with diet and health in the elderly. *Nat.*, **488**(7410), 178.

Clunes, M. T. and Boucher, R. C. (2007). Cystic fibrosis: the mechanisms of pathogenesis of an inherited lung disorder. *Drug Discov. Today: Dis. Mech.*, **4**(2), 63–72.

Coburn, B., Wang, P. W., Diaz Caballero, J., Clark, S. T., Brahma, V., Donaldson,

S., Zhang, Y., Surendra, A., Gong, Y., Elizabeth Tullis, D., Yau, Y. C. W., Waters, V. J., Hwang, D. M., and Guttman, D. S. (2015). Lung microbiota across age and disease stage in cystic fibrosis. *Sci. Reports*, **5**, 10241.

Cohn, J. A., Friedman, K. J., Noone, P. G., Knowles, M. R., Silverman, L. M., and Jowell, P. S. (1998). Relation between Mutations of the Cystic Fibrosis Gene and Idiopathic Pancreatitis. *New Engl. J. Medicine*, **339**(10), 653–658.

Collaco, J. M. and Cutting, G. R. (2008). Update on gene modifiers in cystic fibrosis. *Curr. Opin. Pulm. Medicine*, **14**(6), 559–66.

Collins, S. M. (2014). A role for the gut microbiota in IBS. *Nat. Rev. Gastroenterol. & Hepatol.*, **11**(8), 497–505.

Conrad, D., Haynes, M., Salamon, P., Rainey, P. B., Youle, M., and Rohwer, F. (2013). Cystic Fibrosis Therapy: A Community Ecology Perspective. *Am. J. Respir. Cell Mol. Biol.*, **48**(2), 150–156.

Conway, S. P., Brownlee, K. G., Denton, M., and Peckham, D. G. (2003). Antibiotic Treatment of Multidrug-Resistant Organisms in Cystic Fibrosis. *Am. J. Respir. Medicine*, **2**(4), 321–332.

Corey, M., Edwards, L., Levison, H., and Knowles, M. (1997). Longitudinal analysis of pulmonary function decline in patients with cystic fibrosis. *The J. Pediatr.*, **131**(6), 809–14.

Cox, M. J., Allgaier, M., Taylor, B., Baek, M. S., Huang, Y. J., Daly, R. A., Karaoz, U., Andersen, G. L., Brown, R., Fujimura, K. E., Wu, B., Tran, D., Koff, J., Kleinhenz, M. E., Nielson, D., Brodie, E. L., and Lynch, S. V. (2010). Airway

microbiota and pathogen abundance in age-stratified cystic fibrosis patients. *PLOS ONE*, **5**(6), e11044.

Crighton, E. J., Elliott, S. J., Moineddin, R., Kanaroglou, P., and Upshur, R. E. G. (2007). An exploratory spatial analysis of pneumonia and influenza hospitalizations in Ontario by age and gender. *Epidemiol. Infect.*, **135**(02), 253.

Crossley, K. B. and Peterson, P. K. (1996). Infections in the Elderly. *Clin. Infect. Dis.*, **22**, 209–215.

Cuthbertson, L., Rogers, G. B., Walker, A. W., Oliver, A., Green, L. E., Daniels, T. W. V., Carroll, M. P., Parkhill, J., Bruce, K. D., and van der Gast, C. J. (2015). Respiratory microbiota resistance and resilience to pulmonary exacerbation and subsequent antimicrobial intervention. *The ISME J.*, **10**(5), 1081–1091.

Cutting, G. R. (2014). Cystic fibrosis genetics: from molecular understanding to clinical application. *Nat. Rev. Genet.*.

Cystic Fibrosis Canada (2013). Canadian Cystic Fibrosis Registry 2013 Annual Report. Technical report, Toronto, Ontario.

Cystic Fibrosis Canada (2016). The Canadian Cystic Fibrosis Registry 2014 Annual Report. Technical report, Toronto, Ontario.

Cystic Fibrosis Foundation (2013). Patient Registry Annual Data Report 20. Technical report, Toronto, Ontario.

De Lisle, R. C. and Borowitz, D. (2013). The Cystic Fibrosis Intestine. *Cold Spring Harb. Perspectives Medicine*, **3**(9), a009753–a009753.

De Palma, G., Lynch, M. D. J., Lu, J., Dang, V. T., Deng, Y., Jury, J., Umeh, G., Miranda, P. M., Pigrau Pastor, M., Sidani, S., Pinto-Sanchez, M. I., Philip, V., McLean, P. G., Hagelsieb, M.-G., Surette, M. G., Bergonzelli, G. E., Verdu, E. F., Britz-McKibbin, P., Neufeld, J. D., Collins, S. M., and Bercik, P. (2017). Transplantation of fecal microbiota from patients with irritable bowel syndrome alters gut function and behavior in recipient mice. *Sci. Transl. Medicine*, **9**(379), eaaf6397.

de Steenhuijsen Piters, W. A. A., Huijskens, E. G. W., Wyllie, A. L., Biesbroek, G., van den Bergh, M. R., Veenhoven, R. H., Wang, X., Trzciński, K., Bonten, M. J., Rossen, J. W. A., Sanders, E. A. M., and Bogaert, D. (2015). Dysbiosis of upper respiratory tract microbiota in elderly pneumonia patients. *The ISME J.*, **10**(1), 97–108.

Denton, M., Hall, M., Todd, N., Kerr, K., and Littlewood, J. (2000). Improved isolation of *Stenotrophomonas maltophilia* from the sputa of patients with cystic fibrosis using a selective medium. *Clin. Microbiol. Infect.*, **6**(7), 395–396.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, **72**(7), 5069–5072.

Dickson, R. P., Erb-Downward, J. R., Freeman, C. M., McCloskey, L., Beck, J. M., Huffnagle, G. B., and Curtis, J. L. (2015). Spatial Variation in the Healthy Human Lung Microbiome and the Adapted Island Model of Lung Biogeography. *Annals Am. Thorac. Soc.*.

Dickson, R. P., Erb-Downward, J. R., Freeman, C. M., McCloskey, L., Falkowski, N. R., Huffnagle, G. B., and Curtis, J. L. (2017). Bacterial Topography of the Healthy Human Lower Respiratory Tract. *mBio*, **8**(1), e02287–16.

Dogaru, C. M., Nyffenegger, D., Pescatore, A. M., Spycher, B. D., and Kuehni, C. E. (2014). Breastfeeding and Childhood Asthma: Systematic Review and Meta-Analysis. *Am. J. Epidemiol.*, **179**(10), 1153–1167.

Dominiani, C., Wu, J., Hayes, R. B., and Ahn, J. (2014). Comparison of methods for fecal microbiome biospecimen collection. *BMC Microbiol.*, **14**(103).

Döring, G., Flume, P., Heijerman, H., Elborn, J. S., and Consensus Study Group (2012). Treatment of lung infection in patients with cystic fibrosis: current and future strategies. *J. Cyst. Fibros.*, **11**(6), 461–79.

Drumm, M. L., Konstan, M. W., Schluchter, M. D., Handler, A., Pace, R., Zou, F., Zariwala, M., Fargo, D., Xu, A., Dunn, J. M., Darrah, R. J., Dorfman, R., Sandford, A. J., Corey, M., Zielenski, J., Durie, P., Goddard, K., Yankaskas, J. R., Wright, F. A., and Knowles, M. R. (2005). Genetic Modifiers of Lung Disease in Cystic Fibrosis. *New Engl. J. Medicine*, **353**(14), 1443–1453.

Duong, J., Booth, S. C., McCartney, N. K., Rabin, H. R., Parkins, M. D., and Storey, D. G. (2015). Phenotypic and Genotypic Comparison of Epidemic and Non-Epidemic Strains of Pseudomonas aeruginosa from Individuals with Cystic Fibrosis. *PLOS ONE*, **10**(11), e0143466.

Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent,

M., Gill, S. R., Nelson, K. E., and Relman, D. A. (2005). Diversity of the Human Intestinal Microbial Flora. *Sci.*, **308**(5728), 1635–1638.

Eddy, S. R. (2004). What is dynamic programming? *Nat. Biotechnol.*, **22**(7), 909–910.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**(5), 1792–7.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics.*, **26**(19), 2460–1.

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods*, **10**(10), 996–8.

Elborn, J. S. (2016). Cystic fibrosis. *The Lancet*.

Facklam, R. (2002). What happened to the streptococci: overview of taxonomic and nomenclature changes. *Clin. Microbiol. Rev.*, **15**(4), 613–30.

Fazeli, H., Akbari, R., Moghim, S., and Esfahani, B. N. (2013). Phenotypic characterization and PCR-Ribotypic profile of *Pseudomonas aeruginosa* isolated from cystic fibrosis patients in Iran. *Adv. Biomed. Res.*, **2**, 18.

Fegan, M., Francis, P., Hayward, A. C., Davis, G. H., and Fuerst, J. A. (1990). Phenotypic conversion of *Pseudomonas aeruginosa* in cystic fibrosis. *J. Clin. Microbiol.*, **28**(6), 1143–6.

Feigelman, R., Kahlert, C. R., Baty, F., Rassouli, F., Kleiner, R. L., Kohler, P.,

Brutsche, M. H., and von Mering, C. (2017). Sputum DNA sequencing in cystic fibrosis: non-invasive access to the lung microbiome and to pathogen details. *Microbiome*, **5**(1), 20.

Feinberg, M. J., Knebl, J., Tully, J., and Segall, L. (1990). Aspiration and the elderly. *Dysphagia*, **5**(2), 61–71.

Ferkol, T., Rosenfeld, M., and Milla, C. E. (2006). Cystic fibrosis pulmonary exacerbations. *The J. Pediatr.*, **148**(2), 259–64.

Filkins, L. M., Hampton, T. H., Gifford, A. H., Gross, M. J., Hogan, D. A., Sogin, M. L., Morrison, H. G., Paster, B. J., and O'Toole, G. A. (2012). Prevalence of streptococci and increased polymicrobial diversity associated with cystic fibrosis patient stability. *J. Bacteriol.*, **194**(17), 4709–17.

Finegold, S. M., Attebery, H. R., and Sutter, V. L. (1974). Effect of diet on human fecal flora: comparison of Japanese and American diets. *The Am. J. Clin. Nutr.*, **27**(12), 1456–69.

Flamaing, J., Peetermans, W. E., Vandeven, J., and Verhaegen, J. (2010). Pneumococcal colonization in older persons in a nonoutbreak setting. *J. Am. Geriatr. Soc.*, **58**(2), 396–398.

Flass, T., Tong, S., Frank, D. N., Wagner, B. D., Robertson, C. E., Kotter, C. V., Sokol, R. J., Zemanick, E., Accurso, F., Hoffenberg, E. J., and Narkewicz, M. R. (2015). Intestinal Lesions Are Associated with Altered Intestinal Microbiome and Are More Frequent in Children and Young Adults with Cystic Fibrosis and Cirrhosis. *PLOS ONE*, **10**(2), e0116967.

Flume, P. A., Mogayzel, P. J., Robinson, K. A., Goss, C. H., Rosenblatt, R. L., Kuhn, R. J., Marshall, B. C., and Clinical Practice Guidelines for Pulmonary Therapies Committee (2009). Cystic fibrosis pulmonary guidelines: treatment of pulmonary exacerbations. *Am. J. Respir. Critical Care Medicine*, **180**(9), 802–8.

Fodor, A. A., Klem, E. R., Gilpin, D. F., Elborn, J. S., Boucher, R. C., Tunney, M. M., and Wolfgang, M. C. (2012). The Adult Cystic Fibrosis Airway Microbiota Is Stable over Time and Infection Type, and Highly Resilient to Antibiotic Treatment of Exacerbations. *PLOS ONE*, **7**(9), e45001.

Fox, G. E., Pechman, K. R., and Woese, C. R. (1977). Comparative Cataloging of 16S Ribosomal Ribonucleic Acid: Molecular Approach to Procaryotic Systematics. *Int. J. Syst. Evol. Microbiol.*, **27**(1), 44–57.

Fox, G. E., Stackebrandt, E., Hespell, R. B., Gibson, J., Maniloff, J., Dyer, T. A., Wolfe, R. S., Balch, W. E., Tanner, R. S., Magrum, L. J., Zablen, L. B., Blakemore, R., Gupta, R., Bonen, L., Lewis, B. J., Stahl, D. A., Luehrsen, K. R., Chen, K. N., and Woese, C. R. (1980). The phylogeny of prokaryotes. *Sci.*, **209**(4455), 457–63.

Franceschi, C., Bonafè, M., Valensin, S., Olivieri, F., De Luca, M., Ottaviani, E., and De Benedictis, G. (2000). Inflamm-aging. An evolutionary perspective on immunosenescence. *Annals New York Acad. Sci.*, **908**, 244–54.

Frank, D. N., St Amand, A. L., Feldman, R. A., Boedeker, E. C., Harpaz, N., and Pace, N. R. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl. Acad. Sci.*, **104**(34), 13780–5.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics.*, **28**(23), 3150–2.

Fuchs, H. J., Borowitz, D. S., Christiansen, D. H., Morris, E. M., Nash, M. L., Ramsey, B. W., Rosenstein, B. J., Smith, A. L., and Wohl, M. E. (1994). Effect of aerosolized recombinant human DNase on exacerbations of respiratory symptoms and on pulmonary function in patients with cystic fibrosis. The Pulmozyme Study Group. *The New Engl. J. Medicine*, **331**(10), 637–42.

Geller, D. E., Konstan, M. W., Smith, J., Noonberg, S. B., and Conrad, C. (2007). Novel tobramycin inhalation powder in cystic fibrosis subjects: Pharmacokinetics and safety. *Pediatr. Pulmonol.*, **42**(4), 307–313.

Georg, L. K. and Brown, J. M. (1967). Rothia, gen. nov. an aerobic genus of the family Actinomycetaceae. *Int. J. Syst. Bacteriol.*, **17**(1), 79–88.

Ghodsi, M., Liu, B., and Pop, M. (2011). DNACLUST: accurate and efficient clustering of phylogenetic marker genes. *BMC bioinformatics*, **12**(1), 271.

Gilbert, J. A. and Dupont, C. L. (2011). Microbial Metagenomics: Beyond the Genome. *Annu. Rev. Mar. Sci.*, **3**(1), 347–371.

Gilbert, J. A., Quinn, R. A., Debelius, J., Xu, Z. Z., Morton, J., Garg, N., Jansson, J. K., Dorrestein, P. C., and Knight, R. (2016). Microbiome-wide association studies link dynamic microbial consortia to disease. *Nat.*, **535**(7610), 94–103.

Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M., and Nelson, K. E. (2006).

Metagenomic Analysis of the Human Distal Gut Microbiome. *Sci.*, **312**(5778), 1355–1359.

Goddard, A. F., Staudinger, B. J., Dowd, S. E., Joshi-Datar, A., Wolcott, R. D., Aitken, M. L., Fligner, C. L., and Singh, P. K. (2012). Direct sampling of cystic fibrosis lungs indicates that DNA-based analyses of upper-airway specimens can misrepresent lung microbiota. *Proc. Natl. Acad. Sci.*, **109**(34), 13769–13774.

Goldman, M. J., Anderson, G. M., Stolzenberg, E. D., Kari, U. P., Zasloff, M., and Wilson, J. M. (1997). Human beta-defensin-1 is a salt-sensitive antibiotic in lung that is inactivated in cystic fibrosis. *Cell*, **88**(4), 553–60.

Goodman, A. L., Kallstrom, G., Faith, J. J., Reyes, A., Moore, A., Dantas, G., and Gordon, J. I. (2011). Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *Proc. Natl. Acad. Sci.*, **108**(15), 6252–7.

Goss, C. H. and Burns, J. L. (2007). Exacerbations in cystic fibrosis. 1: Epidemiology and pathogenesis. *Thorax*, **62**(4), 360–7.

Goss, C. H., Newsom, S. A., Schildcrout, J. S., Sheppard, L., and Kaufman, J. D. (2004). Effect of Ambient Air Pollution on Pulmonary Exacerbations and Lung Function in Cystic Fibrosis. *Am. J. Respir. Critical Care Medicine*, **169**(7), 816–821.

Govan, J. R. and Deretic, V. (1996). Microbial pathogenesis in cystic fibrosis: mucoid *Pseudomonas aeruginosa* and *Burkholderia cepacia*. *Microbiol. Rev.*, **60**(3), 539–74.

Gronow, S., Welnitz, S., Lapidus, A., Nolan, M., Ivanova, N., Glavina Del Rio, T., Copeland, A., Chen, F., Tice, H., Pitluck, S., Cheng, J.-F., Saunders, E., Brettin, T., Han, C., Detter, J. C., Bruce, D., Goodwin, L., Land, M., Hauser, L., Chang, Y.-J., Jeffries, C. D., Pati, A., Mavromatis, K., Mikhailova, N., Chen, A., Palaniappan, K., Chain, P., Rohde, M., Göker, M., Bristow, J., Eisen, J. A., Markowitz, V., Hugenholtz, P., Kyrpides, N. C., Klenk, H.-P., and Lucas, S. (2010). Complete genome sequence of *Veillonella parvula* type strain (Te3T). *Standards Genomic Sci.*, **2**(1), 57–65.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. & Biol.*, **5**(10), R245–9.

Hapfelmeier, S., Lawson, M. A. E., Slack, E., Kirundi, J. K., Stoel, M., Heikenwalder, M., Cahenzli, J., Velykoredko, Y., Balmer, M. L., Endt, K., Geuking, M. B., Curtiss, R., McCoy, K. D., and Macpherson, A. J. (2010). Reversible Microbial Colonization of Germ-Free Mice Reveals the Dynamics of IgA Immune Responses. *Sci.*, **328**(5986), 1705–1709.

Harrison, F. (2007). Microbial ecology of the cystic fibrosis lung. *Microbiol.*, **153**(4), 917–923.

Hauser, P. M., Bernard, T., Greub, G., Jaton, K., Pagni, M., and Hafen, G. M. (2014). Microbiota Present in Cystic Fibrosis Lungs as Revealed by Whole Genome Sequencing. *PLOS ONE*, **9**(3), e90934.

He, Y., Caporaso, J. G., Jiang, X.-T., Sheng, H.-F., Huse, S. M., Rideout, J. R., Edgar, R. C., Kopylova, E., Walters, W. A., Knight, R., and Zhou, H.-W. (2015).

Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. *Microbiome*, **3**(1), 20.

Hehemann, J.-H., Correc, G., Barbeyron, T., Helbert, W., Czjzek, M., and Michel, G. (2010). Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nat.*, **464**(7290), 908–912.

Heijerman, H. (2005). Infection and inflammation in cystic fibrosis: A short review. *J. Cyst. Fibros.*, **4**, 3–5.

Hiatt, P. W., Grace, S. C., Kozinetz, C. A., Raboudi, S. H., Treece, D. G., Taber, L. H., and Piedra, P. A. (1999). Effects of viral lower respiratory tract infection on lung function in infants with cystic fibrosis. *Pediatr.*, **103**(3), 619–26.

Hieken, T. J., Chen, J., Hoskin, T. L., Walther-Antonio, M., Johnson, S., Ramaker, S., Xiao, J., Radisky, D. C., Knutson, K. L., Kalari, K. R., Yao, J. Z., Baddour, L. M., Chia, N., and Degnim, A. C. (2016). The Microbiome of Aseptically Collected Human Breast Tissue in Benign and Malignant Disease. *Sci. Reports*, **6**, 30751.

Hilt, E. E., McKinley, K., Pearce, M. M., Rosenfeld, A. B., Zilliox, M. J., Mueller, E. R., Brubaker, L., Gai, X., Wolfe, A. J., and Schreckenberger, P. C. (2014). Urine is not sterile: use of enhanced urine culture techniques to detect resident bacterial flora in the adult female bladder. *J. Clin. Microbiol.*, **52**(3), 871–6.

Hugerth, L. W., Larsson, J., Alneberg, J., Lindh, M. V., Legrand, C., Pinhassi, J., and Andersson, A. F. (2015). Metagenome-assembled genomes uncover a global brackish microbiome. *Genome Biol.*, **16**(1), 279.

Hugon, P., Dufour, J.-C., Colson, P., Fournier, P.-E., Sallah, K., and Raoult, D. (2015). A comprehensive repertoire of prokaryotic species identified in human beings. *The Lancet Infect. Dis.*.

Huse, S. M., Welch, D. M., Morrison, H. G., and Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.*, **12**(7), 1889–98.

Isles, A., Maclusky, I., Corey, M., Gold, R., Prober, C., Fleming, P., and Levison, H. (1984). *Pseudomonas cepacia* infection in cystic fibrosis: an emerging problem. *The J. Pediatr.*, **104**(2), 206–10.

Jackson, M. A., Bell, J. T., Spector, T. D., and Steves, C. J. (2016). A heritability-based comparison of methods used to cluster 16S rRNA gene sequences into operational taxonomic units. *PeerJ*, **4**, e2341.

Jakobsson, H. E., Abrahamsson, T. R., Jenmalm, M. C., Harris, K., Quince, C., Jernberg, C., Björkstén, B., Engstrand, L., and Andersson, A. F. (2014). Decreased gut microbiota diversity, delayed Bacteroidetes colonisation and reduced Th1 responses in infants delivered by Caesarean section. *Gut*, **63**(4), 559–566.

Jandhyala, S. M., Talukdar, R., Subramanyam, C., Vuyyuru, H., Sasikala, M., and Nageshwar Reddy, D. (2015). Role of the normal gut microbiota. *World J. Gastroenterol.*, **21**(29), 8787–803.

Jarad, N. A. and Sequeiros, I. M. (2012). A novel respiratory symptom scoring system for CF pulmonary exacerbations. *Q J Med*, **105**(2), 137–43.

Jeffery, I. B., Lynch, D. B., and O'Toole, P. W. (2015). Composition and temporal stability of the gut microbiota in older persons. *The ISME J.*, **10**(1), 170–182.

Johnson, C. L. and Versalovic, J. (2012). The human microbiome and its potential importance to pediatrics. *Pediatr.*, **129**(5), 950–60.

Johnstone, J., Millar, J., Lelic, A., Verschoor, C. P., Walter, S. D., Devereaux, P. J., Bramson, J., and Loeb, M. (2014). Immunosenescence in the nursing home elderly. *BMC Geriatr.*, **14**(1), 50.

Jokinen, C., Heiskanen, L., Juvonen, H., Kallinen, S., Karkola, K., Korppi, M., Kurki, S., Rönnberg, P. R., Seppä, A., and Soimakallio, S. (1993). Incidence of community-acquired pneumonia in the population of four municipalities in eastern Finland. *Am. J. Epidemiol.*, **137**(9), 977–88.

Jorth, P., Staudinger, B. J., Wu, X., Hisert, K. B., Hayden, H., Garudathri, J., Harding, C. L., Radey, M. C., Rezayat, A., Bautista, G., Berrington, W. R., Goddard, A. F., Zheng, C., Angermeyer, A., Brittnacher, M. J., Kitzman, J., Shendure, J., Fligner, C. L., Mittler, J., Aitken, M. L., Manoil, C., Bruce, J. E., Yahr, T. L., and Singh, P. K. (2015). Regional Isolation Drives Bacterial Diversification within Cystic Fibrosis Lungs. *Cell Host & Microbe*.

Kang, D.-W., Park, J. G., Ilhan, Z. E., Wallstrom, G., LaBaer, J., Adams, J. B., and Krajmalnik-Brown, R. (2013). Reduced Incidence of Prevotella and Other Fermenters in Intestinal Microflora of Autistic Children. *PLOS ONE*, **8**(7), e68322.

Kaplan, V., Angus, D. C., Griffin, M. F., CLERMONT, G., SCOTT WATSON, R.,

and LINDE-ZWIRBLE, W. T. (2002). Hospitalized Community-acquired Pneumonia in the Elderly. *Am. J. Respir. Critical Care Medicine*, **165**(6), 766–772.

Kerem, B., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., Buchwald, M., and Tsui, L. C. (1989). Identification of the cystic fibrosis gene: genetic analysis. *Sci.*, **245**(4922), 1073–80.

Kerem, E., Corey, M., Gold, R., and Levison, H. (1990a). Pulmonary function and clinical course in patients with cystic fibrosis after pulmonary colonization with *Pseudomonas aeruginosa*. *The J. Pediatr.*, **116**(5), 714–9.

Kerem, E., Corey, M., Kerem, B. S., Rommens, J., Markiewicz, D., Levison, H., Tsui, L. C., and Durie, P. (1990b). The Relation between Genotype and Phenotype in Cystic Fibrosis Analysis of the Most Common Mutation (ΔF508). *New Engl. J. Medicine*, **323**(22), 1517–1522.

Khan, T. Z., Wagener, J. S., Bost, T., Martinez, J., Accurso, F. J., and Riches, D. W. (1995). Early pulmonary inflammation in infants with cystic fibrosis. *Am. J. Respir. Critical Care Medicine*, **151**(4), 1075–1082.

Knudsen, B. E., Bergmark, L., Munk, P., Lukjancenko, O., Priemé, A., Aarestrup, F. M., and Pamp, S. J. (2016). Impact of Sample Type and DNA Isolation Procedure on Genomic Inference of Microbiome Composition. *mSystems*, **1**(5), e00095–16.

Kobelska-Dubiel, N., Klincewicz, B., and Cichy, W. (2014). Liver disease in cystic fibrosis. *Prz Gastroenterol*, **9**(3), 136–41.

Kolde, R. (2012). Pheatmap: pretty heatmaps.

Konstantinidis, K. T. and Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci.*, **102**(7), 2567–2572.

Kopp, B. T., Nicholson, L., Paul, G., Tobias, J., Ramanathan, C., and Hayes, D. (2015). Geographic variations in cystic fibrosis: An analysis of the U.S. CF Foundation Registry. *Pediatr. Pulmonol.*.

Kopylova, E., Navas-Molina, J. A., Mercier, C., Xu, Z. Z., Mahé, F., He, Y., Zhou, H.-W., Rognes, T., Caporaso, J. G., and Knight, R. (2016). Open-Source Sequence Clustering Methods Improve the State Of the Art. *mSystems*, **1**(1), e00003–15.

Kosorok, M. R., Zeng, L., West, S. E., Rock, M. J., Splaingard, M. L., Laxova, A., Green, C. G., Collins, J., and Farrell, P. M. (2001). Acceleration of lung disease in children with cystic fibrosis after *Pseudomonas aeruginosa* acquisition. *Pediatr. Pulmonol.*, **32**(4), 277–87.

Kramer, R., Sauer-Heilborn, A., Welte, T., Jauregui, R., Brettar, I., Guzman, C. A., and Höfle, M. G. (2015). High Individuality of Respiratory Bacterial Communities in a Large Cohort of Adult Cystic Fibrosis Patients under Continuous Antibiotic Treatment. *PLOS ONE*, **10**(2), e0117436.

Krone, C. L., van de Groep, K., Trzciński, K., Sanders, E. A. M., and Bogaert, D. (2014). Immunosenescence and pneumococcal disease: an imbalance in host-pathogen interactions. *The Lancet Respir. Medicine*, **2**(2), 141–153.

Kwambana, B. A., Barer, M. R., Bottomley, C., Adegbola, R. A., and Antonio, M. (2011). Early acquisition and high nasopharyngeal co-colonisation by *Streptococcus*

pneumoniae and three respiratory pathogens amongst Gambian new-borns and infants. *BMC Infect. Dis.*, **11**(1), 175.

Lagier, J.-C., Armougom, F., Million, M., Hugon, P., Pagnier, I., Robert, C., Bittar, F., Fournous, G., Gimenez, G., Maraninchi, M., Trape, J.-F., Koonin, E. V., La Scola, B., and Raoult, D. (2012). Microbial culturomics: paradigm shift in the human gut microbiome study. *Clin Microbiol Infect*, **18**(12), 1185–93.

Lagier, J.-C., Khelaifia, S., Alou, M. T., Ndongo, S., Dione, N., Hugon, P., Caputo, A., Cadoret, F., Traore, S. I., Seck, E. H., Dubourg, G., Durand, G., Mourembou, G., Guilhot, E., Togo, A., Bellali, S., Bachar, D., Cassir, N., Bittar, F., Delerce, J., Mailhe, M., Ricaboni, D., Bilen, M., Dangui Nieko, N. P. M., Dia Badiane, N. M., Valles, C., Mouelhi, D., Diop, K., Million, M., Musso, D., Abrahão, J., Azhar, E. I., Bibi, F., Yasir, M., Diallo, A., Sokhna, C., Djossou, F., Vitton, V., Robert, C., Rolain, J. M., La Scola, B., Fournier, P.-E., Levasseur, A., and Raoult, D. (2016). Culture of previously uncultured members of the human gut microbiota by culturomics. *Nat. Microbiol.*, **1**, 16203.

Lam, J. C., Somayaji, R., Surette, M. G., Rabin, H. R., and Parkins, M. D. (2015). Reduction in *Pseudomonas aeruginosa* sputum density during a cystic fibrosis pulmonary exacerbation does not predict clinical response. *BMC Infect. Dis.*, **15**, 145.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**(4), 357–359.

Lau, J. T., Whelan, F. J., Herath, I., Lee, C. H., Collins, S. M., Bercik, P., and

Surette, M. G. (2016). Capturing the diversity of the human gut microbiota through culture-enriched molecular profiling. *Genome Medicine*, **8**(1), 72.

Laurenzi, G. A., Potter, R. T., and Kass, E. H. (1961). Bacteriologic Flora of the Lower Respiratory Tract. *New Engl. J. Medicine*, **265**(26), 1273–1278.

Laurila, J., Kostamovaara, P., and Alahuhta, S. (1998). *Streptococcus salivarius* Meningitis after Spinal Anesthesia. *The J. Am. Soc. Anesthesiol.*, **89**(6), 1579–1580.

Leibovitz, E., Satran, R., Piglansky, L., Raiz, S., Press, J., Leiberman, A., and Dagan, R. (2003). Can acute otitis media caused by *Haemophilus influenzae* be distinguished from that caused by *Streptococcus pneumoniae*? *The Pediatr. Infect. Dis. J.*, **22**(6), 509–514.

Lemon, K. P., Klepac-Ceraj, V., Schiffer, H. K., Brodie, E. L., Lynch, S. V., and Kolter, R. (2010). Comparative analyses of the bacterial microbiota of the human nostril and oropharynx. *mBio*, **1**(3), e00129–10.

Levesque, C., Lamothe, J., and Frenette, M. (2003). Coaggregation of *Streptococcus salivarius* with periodontopathogens: evidence for involvement of fimbriae in the interaction with *Prevotella intermedia*. *Oral Microbiol. Immunol.*, **18**(5), 333–337.

Ley, R. E., Turnbaugh, P. J., Klein, S., and Gordon, J. I. (2006). Microbial ecology: Human gut microbes associated with obesity. *Nat.*, **444**(7122), 1022–1023.

Ley, R. E., Lozupone, C. A., Hamady, M., Knight, R., and Gordon, J. I. (2008). Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat. Rev. Microbiol.*, **6**(10), 776–88.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics.*, **25**(16), 2078–2079.

Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics.*, **22**(13), 1658–9.

Li, Z., Kosorok, M. R., Farrell, P. M., Laxova, A., West, S. E. H., Green, C. G., Collins, J., Rock, M. J., and Splaingard, M. L. (2005). Longitudinal Development of Mucoid *Pseudomonas aeruginosa* Infection and Lung Disease Progression in Children With Cystic Fibrosis. *JAMA*, **293**(5), 581.

Lim, Y. W., Schmieder, R., Haynes, M., Willner, D., Furlan, M., Youle, M., Abbott, K., Edwards, R., Evangelista, J., Conrad, D., and Rohwer, F. (2013). Metagenomics and metatranscriptomics: Windows on CF-associated viral and microbial communities. *J. Cyst. Fibros.*, **12**(2), 154–164.

Lim, Y. W., Evangelista, J. S., Schmieder, R., Bailey, B., Haynes, M., Furlan, M., Maughan, H., Edwards, R., Rohwer, F., Conrad, D., and Conrad, D. (2014). Clinical insights from metagenomic analysis of sputum samples from patients with cystic fibrosis. *J. Clin. Microbiol.*, **52**(2), 425–37.

Limoli, D. H., Yang, J., Khansaheb, M. K., Helfman, B., Peng, L., Stecenko, A. A., and Goldberg, J. B. (2016). *Staphylococcus aureus* and *Pseudomonas aeruginosa* co-infection is associated with cystic fibrosis-related diabetes and poor clinical outcomes. *Eur. J. Clin. Microbiol. & Infect. Dis.*, pages DOI 10.1007/s10096–016–2621–0.

Lindemann, J., Sannwald, D., and Wiesmiller, K. (2008). Age-Related Changes in Intranasal Air Conditioning in the Elderly. *The Laryngoscope*, **118**(8), 1472–1475.

LiPuma, J. J. (2010). The changing microbial epidemiology in cystic fibrosis. *Clin. Microbiol. Rev.*, **23**(2), 299–323.

Littman, D. and Pamer, E. (2011). Role of the Commensal Microbiota in Normal and Pathogenic Host Immune Responses. *Cell Host & Microbe*, **10**(4), 311–323.

Lozupone, C., Cota-Gomez, A., Palmer, B. E., Linderman, D. J., Charlson, E. S., Sodergren, E., Mitreva, M., Abubucker, S., Martin, J., Yao, G., Campbell, T. B., Flores, S. C., Ackerman, G., Stombaugh, J., Ursell, L., Beck, J. M., Curtis, J. L., Young, V. B., Lynch, S. V., Huang, L., Weinstock, G. M., Knox, K. S., Twigg, H., Morris, A., Ghedin, E., Bushman, F. D., Collman, R. G., Knight, R., Fontenot, A. P., and The Lung HIV Microbiome Project (2013). Widespread Colonization of the Lung by *Tropheryma whipplei* in HIV Infection. *Am. J. Respir. Critical Care Medicine*, **187**(10), 1110–1117.

Lupp, C., Robertson, M. L., Wickham, M. E., Sekirov, I., Champion, O. L., Gaynor, E. C., and Finlay, B. B. (2007). Host-Mediated Inflammation Disrupts the Intestinal Microbiota and Promotes the Overgrowth of Enterobacteriaceae. *Cell Host & Microbe*, **2**(2), 119–129.

Lyczak, J. B., Cannon, C. L., and Pier, G. B. (2002). Lung infections associated with cystic fibrosis. *Clin. Microbiol. Rev.*, **15**(2), 194–222.

Lynch, D. B., Jeffery, I. B., and O'Toole, P. W. (2015). The role of the microbiota in ageing: current state and perspectives. *Wiley Interdiscip Rev Syst Biol Med*.

Lynch, S. V. and Bruce, K. D. (2013). The Cystic Fibrosis Airway Microbiome. *Cold Spring Harb. Perspectives Medicine*, **3**(3), a009738–a009738.

MacConkey, A. (1905). Lactose-Fermenting Bacteria in Faeces. *Epidemiol. & Infect.*, **5**(3), 333–379.

Mahe, F., Rognes, T., Quince, C., Vargas, C. D., and Dunthorn, M. (2014). Swarm : robust and fast clustering method for amplicon-based studies. *PeerJ*, **2**(593), 1–13.

Malone, J. G. (2015). Role of small colony variants in persistence of *Pseudomonas aeruginosa* infections in cystic fibrosis lungs. *Infect. Drug Resist.*, **8**, 237–47.

Manor, O., Levy, R., Pope, C. E., Hayden, H. S., Brittnacher, M. J., Carr, R., Radey, M. C., Hager, K. R., Heltshe, S. L., Ramsey, B. W., Miller, S. I., Hoffman, L. R., and Borenstein, E. (2016). Metagenomic evidence for taxonomic dysbiosis and functional imbalance in the gastrointestinal tracts of children with cystic fibrosis. *Sci. Reports*, **6**, 22493.

Marchesi, J. R. and Ravel, J. (2015). The vocabulary of microbiome research: a proposal. *Microbiome*, **3**(1), 31.

Marcy, Y., Ouverney, C., Bik, E. M., Lösekann, T., Ivanova, N., Martin, H. G., Szeto, E., Platt, D., Hugenholtz, P., Relman, D. A., and Quake, S. R. (2007). Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc. Natl. Acad. Sci.*, **104**(29), 11889–94.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**(1), 10.

Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G., and Neufeld, J. D. (2012). PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics.*, **13**(1), 31.

Matamoros, S., Gras-Leguen, C., Le Vacon, F., Potel, G., and de La Cochetiere, M.-F. (2013). Development of intestinal microbiota in infants and its impact on health. *Trends Microbiol.*, **21**(4), 167–173.

Matsui, H., Grubb, B. R., Tarran, R., Randell, S. H., Gatzy, J. T., Davis, C. W., and Boucher, R. C. (1998). Evidence for periciliary liquid layer depletion, not abnormal ion composition, in the pathogenesis of cystic fibrosis airways disease. *Cell*, **95**(7), 1005–15.

May, A., Abeln, S., Crielaard, W., Heringa, J., and Brandt, B. W. (2014). Unraveling the outcome of 16S rDNA-based taxonomy analysis through mock data and simulations. *Bioinformatics.*, **30**(11), 1530–8.

McAvin, J. C., Reilly, P. A., Roudabush, R. M., Barnes, W. J., Salmen, A., Jackson, G. W., Beninga, K. K., Astorga, A., McCleskey, F. K., Huff, W. B., Niemeyer, D., and Lohman, K. L. (2001). Sensitive and specific method for rapid identification of Streptococcus pneumoniae using real-time fluorescence PCR. *J. Clin. Microbiol.*, **39**(10), 3446–51.

McCallum, S. J., Corkill, J., Gallagher, M., Ledson, M. J., Hart, C. A., and Walshaw, M. J. (2001). Superinfection with a transmissible strain of *Pseudomonas aeruginosa* in adults with cystic fibrosis chronically colonised by *P aeruginosa*. *The Lancet*, **358**(9281), 558–60.

McMurdie, P. J. and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLOS ONE*, **8**(4), e61217.

McMurdie, P. J. and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLOS Comput. Biol.*, **10**(4), e1003531.

Mendonca, M. L. (2017). *Characterizing Cooperative and competitive interactions involving Streptococcus intermedius*. Ph.D. thesis.

Meyer, K. C., Sharma, A., Rosenthal, N. S., Peterson, K., and Brennan, L. (1997). Regional Variability of Lung Inflammation in Cystic Fibrosis. *Am. J. Respir. Critical Care Medicine*, **156**(5), 1536–1540.

Miragoli, F., Federici, S., Ferrari, S., Minuti, A., Rebecchi, A., Bruzzese, E., Buccigrossi, V., Guarino, A., and Callegari, M. L. (2016). Impact of Cystic Fibrosis Disease on Archaea and Bacteria Composition of Gut Microbiota. *FEMS Microbiol. Ecol.*, page fiw230.

Mizrahi-Man, O., Davenport, E. R., and Gilad, Y. (2013). Taxonomic Classification of Bacterial 16S rRNA Genes Using Short Sequencing Reads: Evaluation of Effective Study Designs. *PLOS ONE*, **8**(1), e53608.

Moayyedi, P., Surette, M. G., Kim, P. T., Libertucci, J., Wolfe, M., Onischi, C., Armstrong, D., Marshall, J. K., Kassam, Z., Reinisch, W., and Lee, C. H. (2015). Fecal Microbiota Transplantation Induces Remission in Patients With Active Ulcerative Colitis in a Randomized Controlled Trial. *Gastroenterol.*, **149**(1), 102–109.e6.

Moran Losada, P., Chouvarine, P., Dorda, M., Hedtfeld, S., Mielke, S., Schulz, A., Wiehlmann, L., and Tümmler, B. (2016). The cystic fibrosis lower airways microbial metagenome. *ERJ Open Res*, **2**(2).

Morris, A., Beck, J. M., Schloss, P. D., Campbell, T. B., Crothers, K., Curtis, J. L., Flores, S. C., Fontenot, A. P., Ghedin, E., Huang, L., Jablonski, K., Kleerup, E., Lynch, S. V., Sodergren, E., Twigg, H., Young, V. B., Bassis, C. M., Venkataraman, A., Schmidt, T. M., and Weinstock, G. M. (2013). Comparison of the Respiratory Microbiome in Healthy Nonsmokers and Smokers. *Am. J. Respir. Critical Care Medicine*, **187**(10), 1067–1075.

Moss, R. B., Milla, C., Colombo, J., Accurso, F., Zeitlin, P. L., Clancy, J. P., Spencer, L. T., Pilewski, J., Waltz, D. A., Dorkin, H. L., Ferkol, T., Pian, M., Ramsey, B., Carter, B. J., Martin, D. B., and Heald, A. E. (2007). Repeated Aerosolized AAV-CFTR for Treatment of Cystic Fibrosis: A Randomized Placebo-Controlled Phase 2B Trial. *Hum. Gene Ther.*, **18**(8), 726–732.

Mouton, C. P., Bazaldua, O. V., Pierce, B., and Espino, D. V. (2001). Common Infections in Older Adults - American Family Physician. *Am. Fam. Physician*, **15**(63), 257–259.

Murray, S., Charbeneau, J., Marshall, B. C., and LiPuma, J. J. (2008). Impact of *Burkholderia* Infection on Lung Transplantation in Cystic Fibrosis. *Am. J. Respir. Critical Care Medicine*, **178**(4), 363–371.

Muyzer, G., de Waal, E. C., and Uitterlinden, A. G. (1993). Profiling of complex

microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl. Environ. Microbiol.*, **59**(3), 695–700.

Muzanye, G., Morgan, K., Johnson, J., and Mayanja-Kizza, H. (2009). Impact of mouth rinsing before sputum collection on culture contamination. *Afr. Heal. Sci.*, **9**(3), 200.

Myles, I. A., Reckhow, J. D., Williams, K. W., Sastalla, I., Frank, K. M., and Datta, S. K. (2016). A method for culturing Gram-negative skin microbiota. *BMC Microbiol.*, **16**(1), 60.

Naseribafrouei, A., Hestad, K., Avershina, E., Sekelja, M., Linløkken, A., Wilson, R., and Rudi, K. (2014). Correlation between the human fecal microbiota and depression. *Neurogastroenterol Motil*, **26**(8), 1155–1162.

Nicoletti, C., Yang, X., and Cerny, J. (1993). Repertoire diversity of antibody response to bacterial antigens in aged mice. III. Phosphorylcholine antibody from young and aged mice differ in structure and protective activity against infection with *Streptococcus pneumoniae*. *J. Immunol.*, **150**(2), 543–9.

Nielsen, H. B., Almeida, M., Juncker, A. S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D. R., Gautier, L., Pedersen, A. G., Le Chatelier, E., Pelletier, E., Bonde, I., Nielsen, T., Manichanh, C., Arumugam, M., Batto, J.-M., Quintanilha Dos Santos, M. B., Blom, N., Borruel, N., Burgdorf, K. S., Boumezbeur, F., Casellas, F., Doré, J., Dworzynski, P., Guarner, F., Hansen, T., Hildebrand, F., Kaas, R. S., Kennedy, S., Kristiansen, K., Kultima, J. R., Léonard, P., Levenez, F., Lund, O., Moumen, B., Le Paslier, D., Pons, N., Pedersen, O., Prifti, E., Qin, J., Raes, J.,

Sørensen, S., Tap, J., Tims, S., Ussery, D. W., Yamada, T., Renault, P., Sicheritz-Ponten, T., Bork, P., Wang, J., Brunak, S., and Ehrlich, S. D. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.*, **32**(8), 822–8.

Nielsen, S., Needham, B., Leach, S. T., Day, A. S., Jaffe, A., Thomas, T., and Ooi, C. Y. (2016). Disrupted progression of the intestinal microbiota with age in children with cystic fibrosis. *Sci. Reports*, **6**, 24857.

NIH HMP Working Group, Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. A., Bonazzi, V., McEwen, J. E., Wetterstrand, K. A., Deal, C., Baker, C. C., Di Francesco, V., Howcroft, T. K., Karp, R. W., Lunsford, R. D., Wellington, C. R., Belachew, T., Wright, M., Giblin, C., David, H., Mills, M., Salomon, R., Mullins, C., Akolkar, B., Begg, L., Davis, C., Grandison, L., Humble, M., Khalsa, J., Little, A. R., Peavy, H., Pontzer, C., Portnoy, M., Sayre, M. H., Starke-Reed, P., Zakhari, S., Read, J., Watson, B., and Guyer, M. (2009). The NIH Human Microbiome Project. *Genome Res.*, **19**(12), 2317–23.

Oh, J., Conlan, S., Polley, E. C., Segre, J. A., and Kong, H. H. (2012). Shifts in human skin and nares microbiota of healthy children and adults. *Genome Medicine*, **4**(10), 77.

Olesen, S. W. and Alm, E. J. (2016). Dysbiosis is not an answer. *Nat. Microbiol.*, **1**, 16228.

O'Sullivan, B. P. and Freedman, S. D. (2009). Cystic fibrosis. *The Lancet*, **373**(9678), 1891–904.

Pandya, S., Ravi, K., Srinivas, V., Jadhav, S., Khan, A., Arun, A., Riley, L. W., and Madhivanan, P. (2016). Comparison of Culture-dependent and Culture-Independent Molecular Methods for Characterisation of Vaginal Microflora. *J. Med. Microbiol.*.

Parada, A. E., Needham, D. M., and Fuhrman, J. A. (2016). Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ. Microbiol.*, **18**(5), 1403–1414.

Park, S.-H., Kim, K.-A., Ahn, Y.-T., Jeong, J.-J., Huh, C.-S., and Kim, D.-H. (2015). Comparative analysis of gut microbiota in elderly people of urbanized towns and longevity villages. *BMC Microbiol.*, **15**(1), 386.

Parkins, M. D. and Floto, R. A. (2015). Emerging bacterial pathogens and changing concepts of bacterial pathogenesis in cystic fibrosis. *J. Cyst. Fibrosiss.*

Parkins, M. D., Sibley, C. D., Surette, M. G., and Rabin, H. R. (2008). The *Streptococcus* milleri group - An unrecognized cause of disease in cystic fibrosis: A case series and literature review. *Pediatr. Pulmonol.*, **43**(5), 490–497.

Parkins, M. D., Glezerson, B. A., Sibley, C. D., Sibley, K. A., Duong, J., Purighalla, S., Mody, C. H., Workentine, M. L., Storey, D. G., Surette, M. G., and Rabin, H. R. (2014). Twenty-five-year outbreak of *Pseudomonas aeruginosa* infecting individuals with cystic fibrosis: identification of the prairie epidemic strain. *J. Clin. Microbiol.*, **52**(4), 1127–35.

Pickles, R. J. (2004). Physical and Biological Barriers to Viral Vector-mediated Delivery of Genes to the Airway Epithelium. *Proc. Am. Thorac. Soc.*, **1**(4), 302–308.

Planer, J. D., Peng, Y., Kau, A. L., Blanton, L. V., Ndao, I. M., Tarr, P. I., Warner, B. B., and Gordon, J. I. (2016). Development of the gut microbiota and mucosal IgA responses in twins and gnotobiotic mice. *Nat.*, **534**(7606), 263–266.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009). FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Mol. Biol. Evol.*, **26**(7), 1641–1650.

Proctor, R. A., von Eiff, C., Kahl, B. C., Becker, K., McNamara, P., Herrmann, M., and Peters, G. (2006). Small colony variants: a pathogenic form of bacteria that facilitates persistent and recurrent infections. *Nat. Rev. Microbiol.*, **4**(4), 295–305.

Pruitt, K. D., Tatusova, T., Klimke, W., and Maglott, D. R. (2009). NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, **37**(Database issue), D32–6.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, **41**(Database issue), D590–6.

Quinn, R. A., Lim, Y. W., Maughan, H., Conrad, D., Rohwer, F., and Whiteson, K. L. (2014). Biogeochemical forces shape the composition and physiology of polymicrobial communities in the cystic fibrosis lung. *mBio*, **5**(2), e00956–13.

Quinton, P. M. (1989). Defective epithelial ion transport in cystic fibrosis. *Clin. Chem.*, **35**(5), 726–30.

Rabin, H. R. and Surette, M. G. (2012). The cystic fibrosis airway microbiome. *Curr. Opin. Pulm. Medicine*, **18**(6), 622–627.

Ramsey, B. W., Pepe, M. S., Quan, J. M., Otto, K. L., Montgomery, A. B., Williams-Warren, J., Vasiljev-K, M., Borowitz, D., Bowman, C. M., Marshall, B. C., Marshall, S., and Smith, A. L. (1999). Intermittent administration of inhaled tobramycin in patients with cystic fibrosis. Cystic Fibrosis Inhaled Tobramycin Study Group. *The New Engl. J. Medicine*, **340**(1), 23–30.

Ramsey, B. W., Davies, J., McElvaney, N. G., Tullis, E., Bell, S. C., Devínek, P., Griese, M., McKone, E. F., Wainwright, C. E., Konstan, M. W., Moss, R., Ratjen, F., Sermet-Gaudelus, I., Rowe, S. M., Dong, Q., Rodriguez, S., Yen, K., Ordoñez, C., and Elborn, J. S. (2011). A CFTR Potentiator in Patients with Cystic Fibrosis and the *G551D* Mutation. *New Engl. J. Medicine*, **365**(18), 1663–1672.

Rappé, M. S. and Giovannoni, S. J. (2003). The Uncultured Microbial Majority. *Annu. Rev. Microbiol.*, **57**(1), 369–394.

Razvi, S., Quittell, L., Sewall, A., Quinton, H., Marshall, B., and Saiman, L. (2009). Respiratory Microbiology of Patients With Cystic Fibrosis in the United States, 1995 to 2005. *Chest*, **136**(6), 1554–1560.

Reeves, E. P., McCarthy, C., McElvaney, O. J., Vijayan, M. S. N., White, M. M., Dunlea, D. M., Pohl, K., Lacey, N., and McElvaney, N. G. (2015). Inhaled hypertonic saline for cystic fibrosis: Reviewing the potential evidence for modulation of neutrophil signalling and function. *World J. Critical Care Medicine*, **4**(3), 179–91.

Rettedal, E. A., Gumpert, H., and Sommer, M. O. A. (2014). Cultivation-based multiplex phenotyping of human gut microbiota allows targeted recovery of previously uncultured bacteria. *Nat. Commun.*, **5**, 4714.

Rich, D. P., Anderson, M. P., Gregory, R. J., Cheng, S. H., Paul, S., Jefferson, D. M., McCann, J. D., Klinger, K. W., Smith, A. E., and Welsh, M. J. (1990). Expression of cystic fibrosis transmembrane conductance regulator corrects defective chloride channel regulation in cystic fibrosis airway epithelial cells. *Nat.*, **347**(6291), 358–363.

Ridda, I., Macintyre, C. R., Lindley, R., McIntyre, P. B., Brown, M., Oftadeh, S., Sullivan, J., and Gilbert, G. L. (2010). Lack of pneumococcal carriage in the hospitalised elderly. *Vaccine*, **28**(23), 3902–4.

Riordan, J. R., Rommens, J. M., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., and Chou, J. L. (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Sci.*, **245**(4922), 1066–73.

Rogers, G. B., Hart, C. A., Mason, J. R., Hughes, M., Walshaw, M. J., and Bruce, K. D. (2003). Bacterial diversity in cases of lung infection in cystic fibrosis patients: 16S ribosomal DNA (rDNA) length heterogeneity PCR and 16S rDNA terminal restriction fragment length polymorphism profiling. *J. Clin. Microbiol.*, **41**(8), 3548–58.

Rogers, G. B., Carroll, M. P., Serisier, D. J., Hockey, P. M., Kehagia, V., Jones, G. R., and Bruce, K. D. (2005). Bacterial activity in cystic fibrosis lung infections. *Respir. Res.*, **6**(1), 49.

Rogers, G. B., Carroll, M. P., Serisier, D. J., Hockey, P. M., Jones, G., Kehagia, V., Connett, G. J., and Bruce, K. D. (2006). Use of 16S rRNA gene profiling by terminal restriction fragment length polymorphism analysis to compare bacterial communities in sputum and mouthwash samples from patients with cystic fibrosis. *J. Clin. Microbiol.*, **44**(7), 2601–4.

Rommens, J. M., Iannuzzi, M. C., Kerem, B., Drumm, M. L., Melmer, G., Dean, M., Rozmahel, R., Cole, J. L., Kennedy, D., and Hidaka, N. (1989). Identification of the cystic fibrosis gene: chromosome walking and jumping. *Sci.*, **245**(4922), 1059–65.

Ronquist, F. and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics.*, **19**(12), 1572–4.

Rosenfeld, M., Gibson, R. L., McNamara, S., Emerson, J., Burns, J. L., Castile, R., Hiatt, P., McCoy, K., Wilson, C. B., Inglis, A., Smith, A., Martin, T. R., and Ramsey, B. W. (2001). Early pulmonary infection, inflammation, and clinical outcomes in infants with cystic fibrosis. *Pediatr. Pulmonol.*, **32**(5), 356–66.

Rotimi, V. O., Olowe, S. A., and Ahmed, I. (1985). The development of bacterial flora of premature neonates. *The J. Hyg.*, **94**(3), 309–18.

Roumpeka, D. D., Wallace, R. J., Escalettes, F., Fotheringham, I., and Watson, M. (2017). A review of bioinformatics tools for bio-prospecting from metagenomic sequence data. *Front. Genet.*, **8**, 23.

Round, J. L. and Mazmanian, S. K. (2009). The gut microbiota shapes intestinal immune responses during health and disease. *Nat. Rev. Immunol.*, **9**(5), 313–23.

Ruoff, K. L., Miller, S. I., Garner, C. V., Ferraro, M. J., and Calderwood, S. B. (1989). Bacteremia with *Streptococcus bovis* and *Streptococcus salivarius*: clinical correlates of more accurate identification of isolates. *J. Clin. Microbiol.*, **27**(2), 305–8.

Saiman, L., Anstead, M., Mayer-Hamblett, N., Lands, L. C., Kloster, M., Hocevar-Trnka, J., Goss, C. H., Rose, L. M., Burns, J. L., Marshall, B. C., Ratjen, F., and AZ0004 Azithromycin Study Group (2010). Effect of Azithromycin on Pulmonary Function in Patients With Cystic Fibrosis Uninfected With *Pseudomonas aeruginosa*: A Randomized Controlled Trial. *JAMA*, **303**(17), 1707.

Sakamoto, H., Kato, H., Sato, T., and Sasaki, J. (1998). Semiquantitative bacteriology of closed odontogenic abscesses. *The Bull. Tokyo Dental Coll.*, **39**(2), 103–7.

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N. J., and Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.*, **12**(1), 87.

Sampson, T. R., Debelius, J. W., Thron, T., Janssen, S., Shastri, G. G., Ilhan, Z. E., Challis, C., Schretter, C. E., Rocha, S., Gradinaru, V., Chesselet, M.-F., Keshavarzian, A., Shannon, K. M., Krajmalnik-Brown, R., Wittung-Stafshede, P., Knight, R., and Mazmanian, S. K. (2016). Gut Microbiota Regulate Motor Deficits and Neuroinflammation in a Model of Parkinson's Disease. *Cell*, **167**(6), 1469–1480.e12.

Sánchez, F., Mensa, J., Martínez, J. A., Angrill, J., Marcos, M. A., Marco, F., Coll-Vinent, B., Torres, A., and Soriano, E. (1999). Pneumonia caused by *Haemophilus influenzae*. Study in a series of 58 patients. *Rev Esp Quimioter.*, **12**(4), 369–74.

Sanders, D. B., Bittner, R. C. L., Rosenfeld, M., Hoffman, L. R., Redding, G. J., and Goss, C. H. (2010). Failure to Recover to Baseline Pulmonary Function after Cystic Fibrosis Pulmonary Exacerbation. *Am. J. Respir. Critical Care Medicine*, **182**(5), 627–632.

Sanders, D. B., Bittner, R. C. L., Rosenfeld, M., Redding, G. J., and Goss, C. H. (2011). Pulmonary exacerbations are associated with subsequent FEV1 decline in both adults and children with cystic fibrosis. *Pediatr. Pulmonol.*, **46**(4), 393–400.

Sangwan, N., Xia, F., and Gilbert, J. A. (2016). Recovering complete and draft population genomes from metagenome datasets. *Microbiome*, **4**(1), 8.

Savage, D. C. (1977). Microbial Ecology of the Gastrointestinal Tract. *Annu. Rev. Microbiol.*, **31**(1), 107–133.

Sawicki, G. S., Sellers, D. E., and Robinson, W. M. (2009). High treatment burden in adults with cystic fibrosis: challenges to disease self-management. *J. Cyst. Fibros.*, **8**(2), 91–6.

Scannapieco, F. A. (1999). Role of oral bacteria in respiratory infection. *J. Periodontol.*, **70**(7), 793–802.

Schaubeck, M., Clavel, T., Calasan, J., Lagkouvardos, I., Haange, S. B., Jehmlich, N., Basic, M., Dupont, A., Hornef, M., von Bergen, M., Bleich, A., and Haller, D.

(2016). Dysbiotic gut microbiota causes transmissible Crohn's disease-like ileitis independent of failure in antimicrobial defence. *Gut*, **65**(2), 225–237.

Schenck, L. P., Surette, M. G., and Bowdish, D. M. E. (2016). Composition and immunological significance of the upper respiratory tract microbiota. *FEBS Lett.*.

Schloss, P. D. (2016). Application of a Database-Independent Approach To Assess the Quality of Operational Taxonomic Unit Picking Methods. *mSystems*, **1**(2), e00027–16.

Schloss, P. D. and Westcott, S. L. (2011). Assessing and Improving Methods Used in Operational Taxonomic Unit-Based Approaches for 16S rRNA Gene Sequence Analysis. *Appl. Environ. Microbiol.*, **77**(10), 3219–3226.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., and Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**(23), 7537–41.

Schmieder, R. and Edwards, R. (2011). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLOS ONE*, **6**(3), e17288.

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Droege, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., Bremges, A., Fritz, A., Garrido-Oter, R., Jorgensen, T. S., Shapiro, N., Blood, P. D., Gurevich, A., Bai, Y., Turaev, D., DeMaere, M. Z., Chikhi, R., Nagarajan, N., Quince, C., Hansen, L. H., Sorensen, S. J.,

Chia, B. K. H., Denis, B., Froula, J. L., Wang, Z., Egan, R., Kang, D. D., Cook, J. J., Deltel, C., Beckstette, M., Lemaitre, C., Peterlongo, P., Rizk, G., Lavenier, D., Wu, Y.-W., Singer, S. W., Jain, C., Strous, M., Klingenberg, H., Meinicke, P., Barton, M., Lingner, T., Lin, H.-H., Liao, Y.-C., Silva, G. G. Z., Cuevas, D. A., Edwards, R. A., Saha, S., Piro, V. C., Renard, B. Y., Pop, M., Klenk, H.-P., Goeker, M., Kyrpides, N., Woyke, T., Vorholt, J. A., Schulze-Lefert, P., Rubin, E. M., Darling, A. E., Rattei, T., and McHardy, A. C. (2017). Critical Assessment of Metagenome Interpretation a benchmark of computational metagenomics software. *bioRxiv*, page 099127.

Segal, L. N., Alekseyenko, A. V., Clemente, J. C., Kulkarni, R., Wu, B., Chen, H., Berger, K. I., Goldring, R. M., Rom, W. N., Blaser, M. J., and Weiden, M. D. (2013). Enrichment of lung microbiome with supraglottic taxa is associated with increased pulmonary inflammation. *Microbiome*, **1**(1), 19.

Sender, R., Fuchs, S., and Milo, R. (2016). Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. *Cell*, **164**(3), 337–340.

Shade, A. (2017). Diversity is the question, not the answer. *The ISME J.*, **11**(1), 1–6.

Shah, H. N. and Collins, D. M. (1990). *Prevotella*, a New Genus To Include *Bacteroides melaninogenicus* and Related Species Formerly Classified in the Genus *Bacteroides*. *Int. J. Syst. Bacteriol.*, **40**(2), 205–208.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell Syst. Tech. J.*, **27**(3), 379–423.

Shinzato, T. and Saito, A. (1994). A mechanism of pathogenicity of "Streptococcus milleri group" in pulmonary infection: synergy with an anaerobe. *J. Med. Microbiol.*, **40**(2), 118–23.

Shwachman, H. and Kulczycki, L. L. (1958). Long-Term Study of One Hundred Five Patients with Cystic Fibrosis. *AMA Am J Dis Child*, **96**(1), 6.

Sibley, C. D., Parkins, M. D., Rabin, H. R., Duan, K., Norgaard, J. C., and Surette, M. G. (2008). A polymicrobial perspective of pulmonary infections exposes an enigmatic pathogen in cystic fibrosis patients. *Proc. Natl. Acad. Sci.*, **105**(39), 15070–5.

Sibley, C. D., Parkins, M. D., Rabin, H. R., and Surette, M. G. (2009). The relevance of the polymicrobial nature of airway infection in the acute and chronic management of patients with cystic fibrosis. *Curr. Opin. Investig. Drugs*, **10**(8), 787–94.

Sibley, C. D., Grinwis, M. E., Rabin, H. R., and Surette, M. G. (2010a). Azithromycin paradox in the treatment of cystic fibrosis airway disease. *Futur. Microbiol.*, **5**(9), 1315–1319.

Sibley, C. D., Grinwis, M. E., Field, T. R., Parkins, M. D., Norgaard, J. C., Gregson, D. B., Rabin, H. R., and Surette, M. G. (2010b). McKay agar enables routine quantification of the 'Streptococcus milleri' group in cystic fibrosis patients. *J. Med. Microbiol.*, **59**(Pt 5), 534–40.

Sibley, C. D., Grinwis, M. E., Field, T. R., Eshaghurshan, C. S., Faria, M. M., Dowd, S. E., Parkins, M. D., Rabin, H. R., and Surette, M. G. (2011). Culture enriched

molecular profiling of the cystic fibrosis airway microbiome. *PLOS ONE*, **6**(7), e22702.

Simoes, E. A. F., Cherian, T., Chow, J., Shahid-Salles, S. A., Laxminarayan, R., and John, T. J. (2006). *Acute Respiratory Infections in Children*. The International Bank for Reconstruction and Development / The World Bank.

Simpson, E. (1949). Measurement of Diversity. *Nat.*, **162**(688).

Sinha, R., Abnet, C. C., White, O., Knight, R., and Huttenhower, C. (2015). The microbiome quality control project: baseline study design and future directions. *Genome Biol.*, **16**(1), 276.

Sly, P. D., Gangell, C. L., Chen, L., Ware, R. S., Ranganathan, S., Mott, L. S., Murray, C. P., and Stick, S. M. (2013). Risk Factors for Bronchiectasis in Children with Cystic Fibrosis. *New Engl. J. Medicine*, **368**(21), 1963–1970.

Smith, D. J., Badrick, A. C., Zakrzewski, M., Krause, L., Bell, S. C., Anderson, G. J., and Reid, D. W. (2014). Pyrosequencing reveals transient cystic fibrosis lung microbiome changes with intravenous antibiotics. *The Eur. Respir. J.*, **44**(4), 922–30.

Smith, J. J., Travis, S. M., Greenberg, E. P., and Welsh, M. J. (1996). Cystic fibrosis airway epithelia fail to kill bacteria because of abnormal airway surface fluid. *Cell*, **85**(2), 229–36.

Smyth, R. L., Croft, N. M., O'Hea, U., Marshall, T. G., and Ferguson, A. (2000). Intestinal inflammation in cystic fibrosis. *Arch Dis Child*, **82**(5), 394–9.

Speirs, J. J., van der Ent, C. K., and Beekman, J. M. (2012). Effects of *Aspergillus fumigatus* colonization on lung function in cystic fibrosis. *Curr. Opin. Pulm. Medicine*, **18**(6), 632–638.

Spor, A., Koren, O., and Ley, R. (2011). Unravelling the effects of the environment and host genotype on the gut microbiome. *Nat. Rev. Microbiol.*, **9**(4), 279–290.

Starner, T. D., Zhang, N., Kim, G., Apicella, M. A., and McCray, P. B. (2006). *Haemophilus influenzae* Forms Biofilms on Airway Epithelia. *Am. J. Respir. Critical Care Medicine*, **174**(2), 213–220.

Stearns, J. C., Davidson, C. J., McKeon, S., Whelan, F. J., Fontes, M. E., Schryvers, A. B., Bowdish, D. M. E., Kellner, J. D., and Surette, M. G. (2015). Culture and molecular-based profiles show shifts in bacterial communities of the upper respiratory tract that occur with age. *The ISME J.*.

Stewart, E. J. (2012). Growing unculturable bacteria. *J. Bacteriol.*, **194**(16), 4151–60.

Stressmann, F. A., Rogers, G. B., Klem, E. R., Lilley, A. K., Donaldson, S. H., Daniels, T. W., Carroll, M. P., Patel, N., Forbes, B., Boucher, R. C., Wolfgang, M. C., and Bruce, K. D. (2011a). Analysis of the Bacterial Communities Present in Lungs of Patients with Cystic Fibrosis from American and British Centers. *J. Clin. Microbiol.*, **49**(1), 281–291.

Stressmann, F. A., Rogers, G. B., Marsh, P., Lilley, A. K., Daniels, T. W., Carroll, M. P., Hoffman, L. R., Jones, G., Allen, C. E., Patel, N., Forbes, B., Tuck, A., and Bruce, K. D. (2011b). Does bacterial density in cystic fibrosis sputum increase prior to pulmonary exacerbation? *J. Cyst. Fibros.*, **10**(5), 357–365.

Sudo, N., Sawamura, S., Tanaka, K., Aiba, Y., Kubo, C., and Koga, Y. (1997). The requirement of intestinal bacterial flora for the development of an IgE production system fully susceptible to oral tolerance induction. *J. Immunol.*, **159**(4), 1739–45.

Sun, Y., Cai, Y., Liu, L., Yu, F., Farrell, M. L., McKendree, W., and Farmerie, W. (2009). ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Res.*, **37**(10), e76–e76.

Sun, Y., Cai, Y., Huse, S. M., Knight, R., Farmerie, W. G., Wang, X., and Mai, V. (2012). A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Briefings Bioinformatics.*, **13**(1), 107–21.

Surette, M. G. (2014). The cystic fibrosis lung microbiome. *Annals Am. Thorac. Soc.*, **11 Suppl 1**, S61–5.

Szaff, M. and Hoiby, N. (1982). Antibiotic treatment of *Staphylococcus aureus* infection in cystic fibrosis. *Acta Paediatr.*, **71**(5), 821–826.

Sze, M. and Schloss, P. D. (2016). Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome. *bioRxiv*, page 057331.

Szego, E. and Canadian Physiotherapy Group (2017). Cystic Fibrosis Canada.

Tan, K., Conway, S. P., Brownlee, K. G., Etherington, C., and Peckham, D. G. (2002). Alcaligenes infection in cystic fibrosis. *Pediatr. Pulmonol.*, **34**(2), 101–104.

Temperton, B. and Giovannoni, S. J. (2012). Metagenomics: microbial diversity through a scratched lens. *Curr. Opin. Microbiol.*, **15**(5), 605–612.

The Human Microbiome Project Consortium (2012a). A framework for human microbiome research. *Nat.*, **486**(7402), 215–21.

The Human Microbiome Project Consortium (2012b). Structure, function and diversity of the healthy human microbiome. *Nat.*, **486**(7402), 207–14.

Thevaranjan, N., Whelan, F. J., Puchta, A., Ashu, E., Rossi, L., Surette, M. G., and Bowdish, D. M. E. (2016). *Streptococcus pneumoniae* colonization disrupts the microbial community within the upper respiratory tract of aging mice. *Infect. Immun.*, **84**(4), 906–16.

Thompson, H., Rybalka, A., Moazzez, R., Dewhirst, F. E., and Wade, W. G. (2015). In-vitro culture of previously uncultured oral bacterial phylotypes. *Appl. Environ. Microbiol.*, **81**(24), 8307–8314.

Tirouvanziam, R., de Bentzmann, S., Hubeau, C., Hinnrasky, J., Jacquot, J., Péault, B., and Puchelle, E. (2000). Inflammation and Infection in Naive Human Cystic Fibrosis Airway Grafts. *Am. J. Respir. Cell Mol. Biol.*, **23**(2), 121–127.

Tunney, M. M., Field, T. R., Moriarty, T. F., Patrick, S., Doering, G., Muhlebach, M. S., Wolfgang, M. C., Boucher, R., Gilpin, D. F., McDowell, A., and Elborn, J. S. (2008). Detection of anaerobic bacteria in high numbers in sputum from patients with cystic fibrosis. *Am. J. Respir. Critical Care Medicine*, **177**(9), 995–1001.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nat.*, **449**(7164), 804–10.

Turnbaugh, P. J., Ridaura, V. K., Faith, J. J., Rey, F. E., Knight, R., and Gordon,

J. I. (2009). The Effect of Diet on the Human Gut Microbiome: A Metagenomic Analysis in Humanized Gnotobiotic Mice. *Sci. Transl. Medicine*, **1**(6), 6ra14–6ra14.

Ulrich, M., Herbert, S., Berger, J., Bellon, G., Louis, D., Münker, G., and Döring, G. (1998). Localization of *Staphylococcus aureus* in Infected Airways of Patients with Cystic Fibrosis and in a Cell Culture Model of *S. aureus* Adherence. *Am. J. Respir. Cell Mol. Biol.*, **19**(1), 83–91.

van Benten, I. J., van Drunen, C. M., Koopman, L. P., van Middelkoop, B. C., Hop, W. C. J., Osterhaus, A. D. M. E., Neijens, H. J., and Fokkens, W. J. (2005). Age- and infection-related maturation of the nasal immune response in 0-2-year-old children. *Allergy*, **60**(2), 226–32.

Van Leeuwenhoek, A. (1683). Microscopical observations about animals in the scurf of the teeth. *Philos Trans R Soc Lond B Biol Sci*, **14**, 568–574.

van Nimwegen, F. A., Penders, J., Stobberingh, E. E., Postma, D. S., Koppelman, G. H., Kerkhof, M., Reijmerink, N. E., Dompeling, E., van den Brandt, P. A., Ferreira, I., Mommers, M., and Thijs, C. (2011). Mode and place of delivery, gastrointestinal microbiota, and their influence on asthma and atopy. *J. Allergy Clin. Immunol.*, **128**(5), 948–955.e3.

Venkataraman, A., Bassis, C. M., Beck, J. M., Young, V. B., Curtis, J. L., Huffnagle, G. B., and Schmidt, T. M. (2015). Application of a neutral community model to assess structuring of the human lung microbiome. *mBio*, **6**(1), e002284–14.

Vogtmann, E., Hua, X., Zeller, G., Sunagawa, S., Voigt, A. Y., Hercog, R., Goedert, J. J., Shi, J., Bork, P., and Sinha, R. (2016). Colorectal Cancer and the Human

Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing. *PLOS ONE*, **11**(5), e0155362.

Wainwright, C. E., Elborn, J. S., Ramsey, B. W., Marigowda, G., Huang, X., Cipolli, M., Colombo, C., Davies, J. C., De Boeck, K., Flume, P. A., Konstan, M. W., McColley, S. A., McCoy, K., McKone, E. F., Munck, A., Ratjen, F., Rowe, S. M., Waltz, D., and Boyle, M. P. (2015). Lumacaftor-Ivacaftor in Patients with Cystic Fibrosis Homozygous for Phe508del CFTR. *New Engl. J. Medicine*, **373**(3), 220–231.

Walter, N. D., Taylor, T. H., Dowell, S. F., Mathis, S., and Moore, M. R. (2009). Holiday Spikes in Pneumococcal Disease among Older Adults. *New Engl. J. Medicine*, **361**(26), 2584–2585.

Walters, W., Hyde, E. R., Berg-Lyons, D., Ackermann, G., Humphrey, G., Parada, A., Gilbert, J. A., Jansson, J. K., Caporaso, J. G., Fuhrman, J. A., Apprill, A., and Knight, R. (2016). Improved Bacterial 16S rRNA Gene (V4 and V4-5) and Fungal Internal Transcribed Spacer Marker Gene Primers for Microbial Community Surveys. *mSystems*, **1**(1), e00009–15.

Walters, W. A., Xu, Z., and Knight, R. (2014). Meta-analyses of human gut microbes associated with obesity and IBD. *FEBS Lett.*, **588**(22), 4223–33.

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, **73**(16), 5261–7.

Wang, Y., Zhang, Z. S., Ruan, J. S., Wang, Y. M., and Ali, S. M. (1999). Investigation

of actinomycete diversity in the tropical rainforests of Singapore. *J. Ind. Microbiol. Biotechnol.*, **23**(3), 178–187.

Werlin, S. L., Benuri-Silbiger, I., Kerem, E., Adler, S. N., Goldin, E., Zimmerman, J., Malka, N., Cohen, L., Armoni, S., Yatzkan-Israelit, Y., Bergwerk, A., Aviram, M., Bentur, L., Mussaffi, H., Bjarnasson, I., and Wilschanski, M. (2010). Evidence of Intestinal Inflammation in Patients With Cystic Fibrosis. *J. Pediatr. Gastroenterol. Nutr.*, **51**(3), 1.

Westcott, S. L. and Schloss, P. D. (2015). De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ*, **3**, e1487.

Whelan, F. J. and Surette, M. G. (2015). Clinical insights into pulmonary exacerbations in cystic fibrosis from the microbiome what are we missing? *Annals Am. Thorac. Soc.*, **12**(6), S207–S211.

Whelan, F. J., Verschoor, C. P., Stearns, J. C., Rossi, L., Luinstra, K., Loeb, M., Smieja, M., Johnstone, J., Surette, M. G., and Bowdish, D. M. E. (2014). The loss of topography in the microbial communities of the upper respiratory tract in the elderly. *Annals Am. Thorac. Soc.*, **11**(4), 513–21.

Whiteson, K. L., Meinardi, S., Lim, Y. W., Schmieder, R., Maughan, H., Quinn, R., Blake, D. R., Conrad, D., and Rohwer, F. (2014). Breath gas metabolites and bacterial metagenomes from cystic fibrosis airways indicate active pH neutral 2,3-butanedione fermentation. *The ISME J.*, **8**(6), 1247–58.

Wickham, H. (2007). Reshaping Data with the reshape Package. *J. Stat. Softw.*, **21**(12), 1–20.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, New York.

Williams, B. G., Gouws, E., Boschi-Pinto, C., Bryce, J., and Dye, C. (2002). Estimates of world-wide distribution of child deaths from acute respiratory infections. *The Lancet Infect. Dis.*, **2**(1), 25–32.

Willner, D., Furlan, M., Haynes, M., Schmieder, R., Angly, F. E., Silva, J., Tammadoni, S., Nosrat, B., Conrad, D., and Rohwer, F. (2009). Metagenomic Analysis of Respiratory Tract DNA Viral Communities in Cystic Fibrosis and Non-Cystic Fibrosis Individuals. *PLOS ONE*, **4**(10), e7370.

Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A., and Singer, S. W. (2014). MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome*, **2**(1), 26.

Wu, Y. W., Simmons, B. A., and Singer, S. W. (2015). MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics.*, **32**(4).

Xia, L. C., Steele, J. A., Cram, J. A., Cardon, Z. G., Simmons, S. L., Vallino, J. J., Fuhrman, J. A., and Sun, F. (2011). Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. *BMC Syst. Biol.*, **5 Suppl 2**(Suppl 2), S15.

Ye, Y. (2011). Identification and Quantification of Abundant Species from Pyrosequences of 16S rRNA by Consensus Alignment. *Proc. (IEEE Int Conf Bioinformatics. Biomed)*, pages 153–157.

Zapata, H. J. and Quagliarello, V. J. (2015). The Microbiota and Microbiome in Aging: Potential Implications in Health and Age-Related Diseases. *J. Am. Geriatr. Soc.*.

Zhao, J., Li, J., Schloss, P. D., Kalikin, L. M., Raymond, T. A., Petrosino, J. F., Young, V. B., and LiPuma, J. J. (2011). Effect of sample storage conditions on culture-independent bacterial community measures in cystic fibrosis sputum specimens. *J. Clin. Microbiol.*, **49**(10), 3717–8.

Zhao, J., Schloss, P. D., Kalikin, L. M., Carmody, L. A., Foster, B. K., Petrosino, J. F., Cavalcoli, J. D., VanDevanter, D. R., Murray, S., Li, J. Z., Young, V. B., and LiPuma, J. J. (2012). Decade-long bacterial community dynamics in cystic fibrosis airways. *Proc. Natl. Acad. Sci.*, **109**(15), 5809–14.

Appendix A

Appendix to Chapter 2

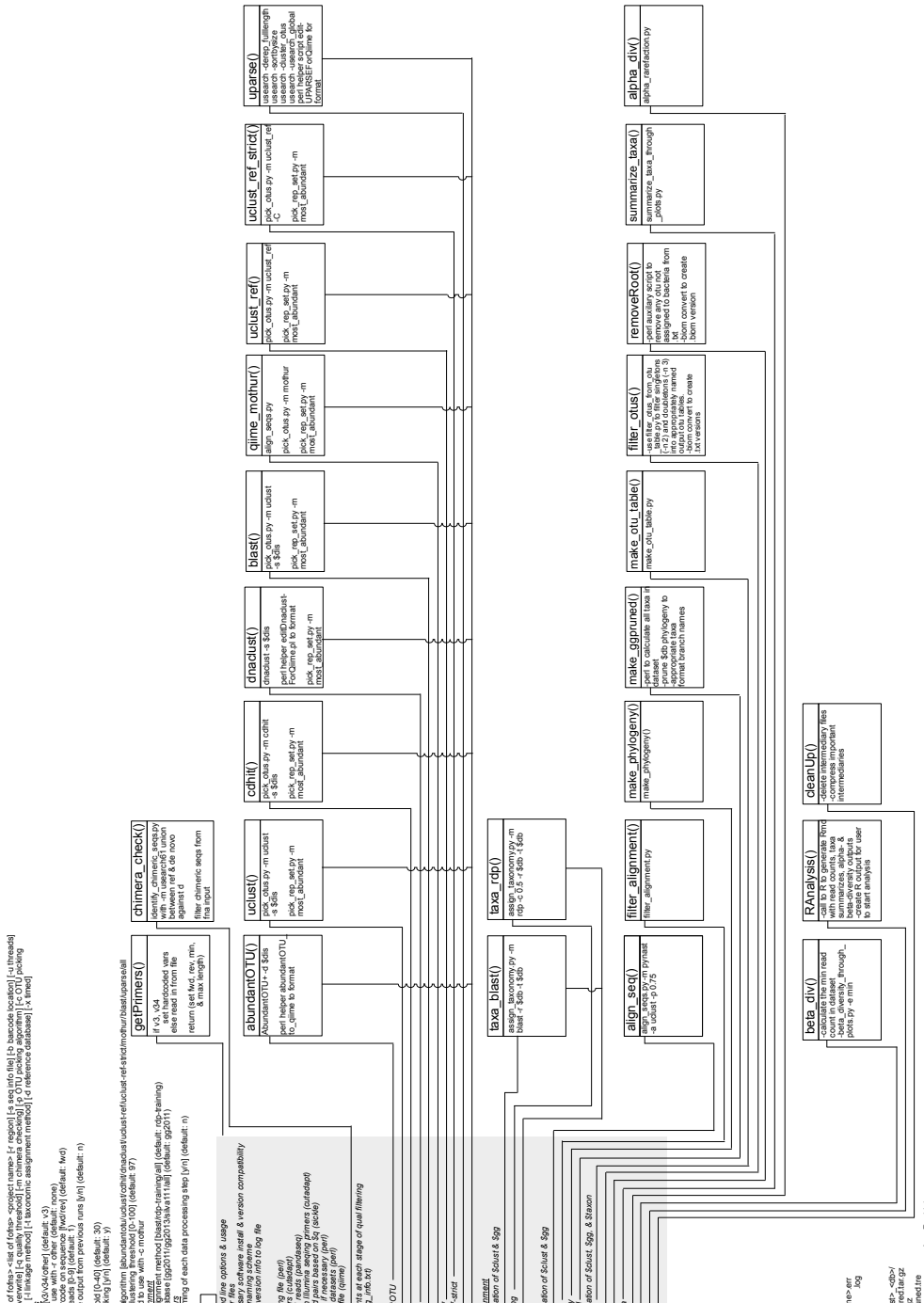


Figure A.1: **Full schematic of the sl1p software.** sl1p is a pipeline script which calls a series of sub-routines based on user-defined parameters given as a series of runtime flags. This Figure accompanies **Figure 2.1**.

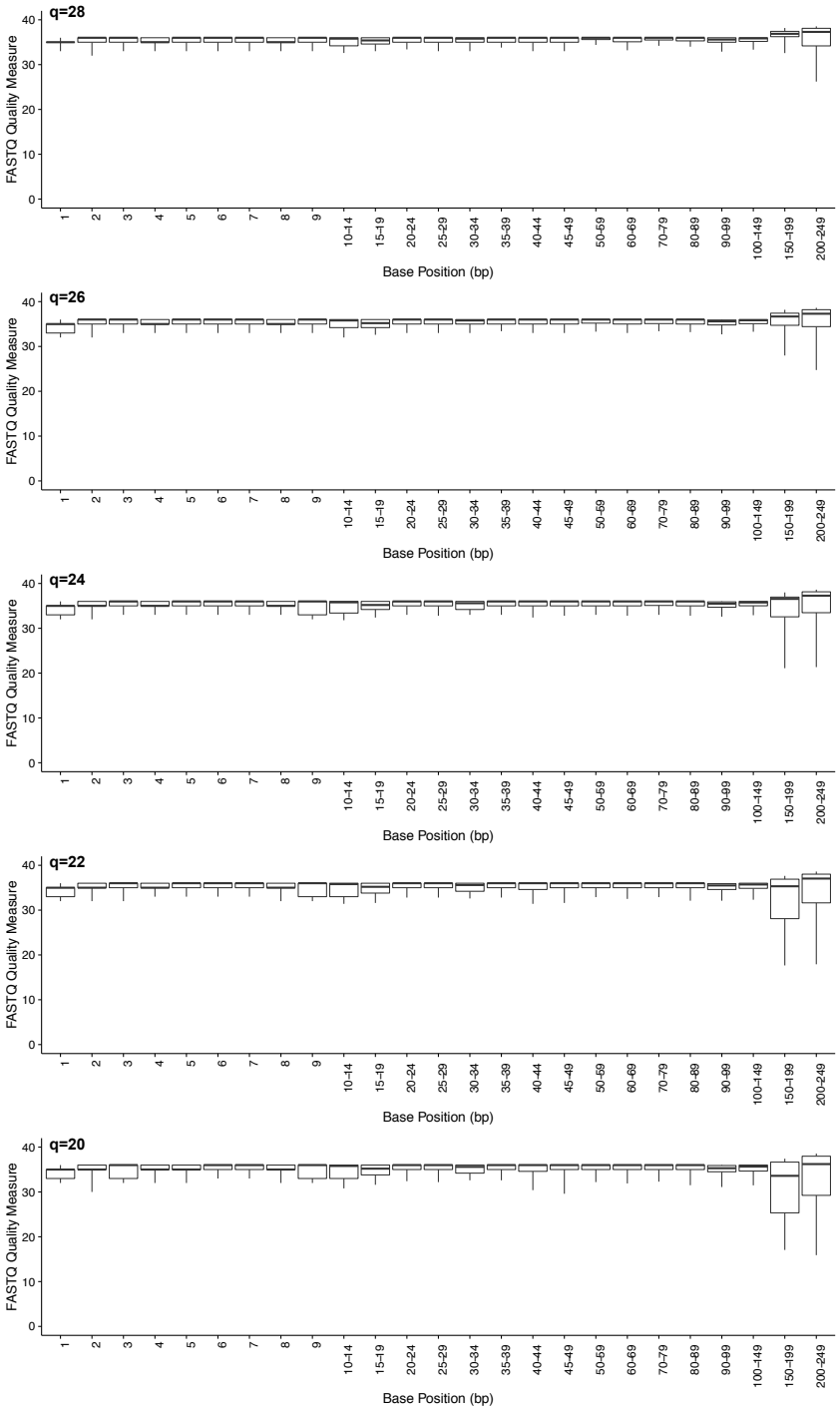


Figure A.2: **Comparisons of various thresholds for quality trimming.** Sickle takes as input a quality threshold with which it determines its quality trimming parameters. Here, we compare the results with a threshold of 30 (**Fig 2.3**) with sequentially lower quality threshold inputs into sickle.

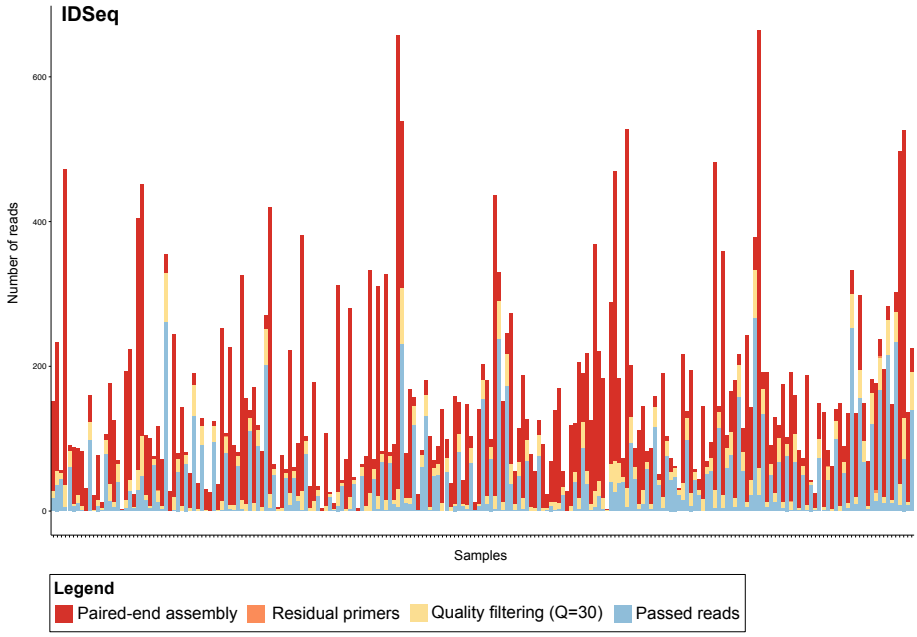


Figure A.3: **Outline of reads lost in the URTCul dataset during sl1p's quality control pipeline.** More input reads were culled during the PANDAseq alignment step in this dataset compared to HMP-mock (**Fig 2.3**), possibly due to a difference in target variable region length between the two datasets.

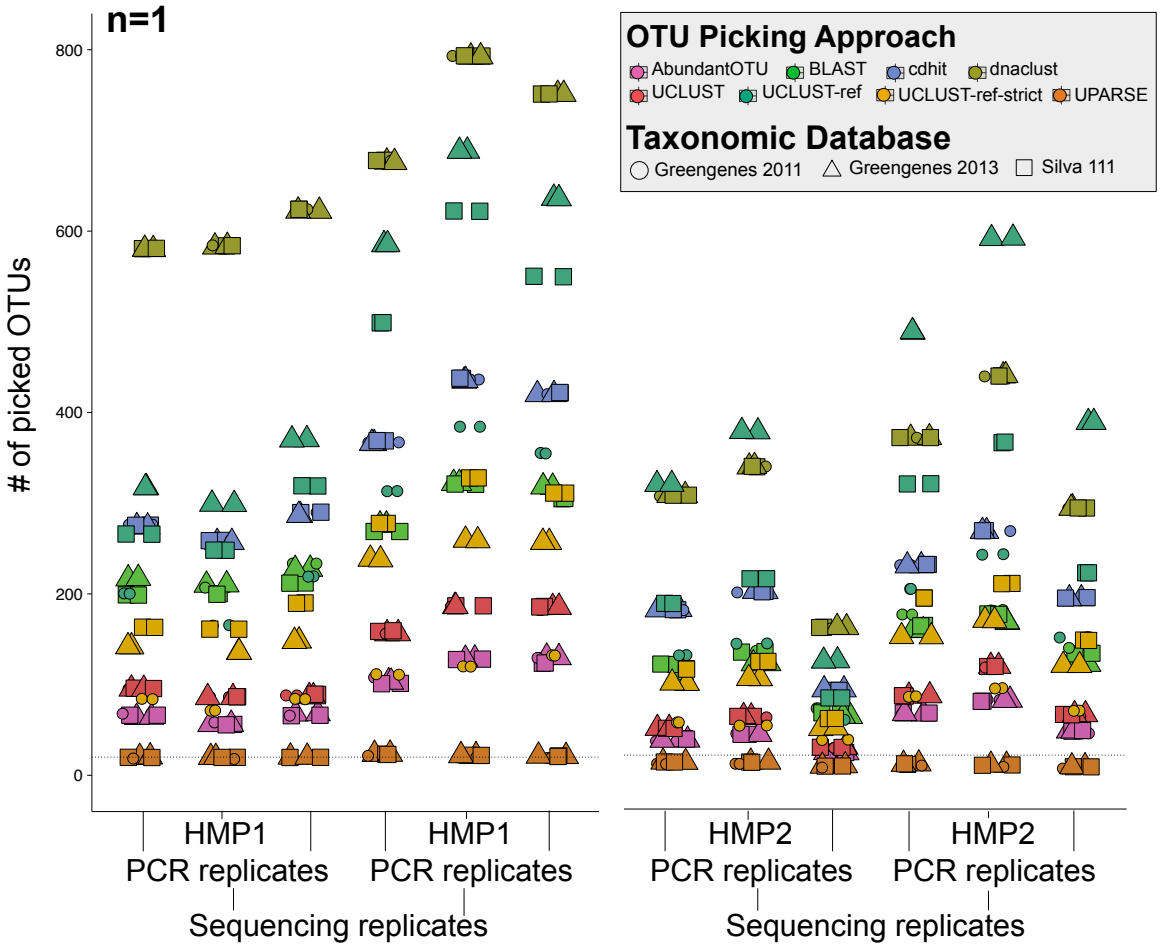


Figure A.4: **OTU clustering methods perform variably when OTUs $\leq$ 1 read are culled.** When 8 methods were used on a control community of known composition, many reported vastly increased OTU counts compared to known sample diversity (n=20, dotted line). Singletons and non-bacterial sequences were removed as part of sequence processing. The dotted line indicates the expected number of OTUs.

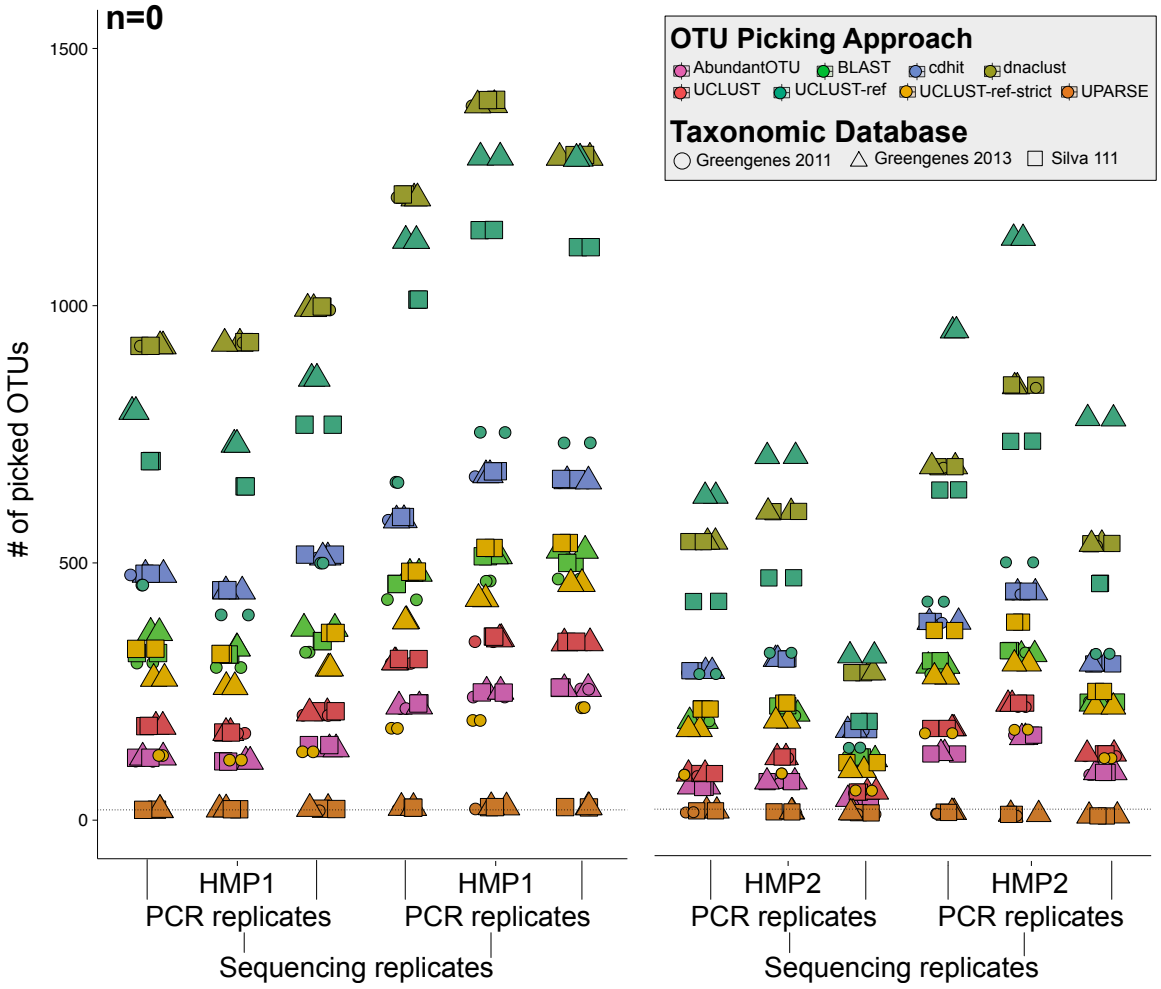


Figure A.5: **OTU clustering methods perform variably when all OTUs are included.** As visualized in **Figure 2.4**, the number of observed OTUs varies depending on clustering approach. Variability is also observed between sequencing and PCR replicates. OTUs not recognized as Bacteria were removed prior to analysis.

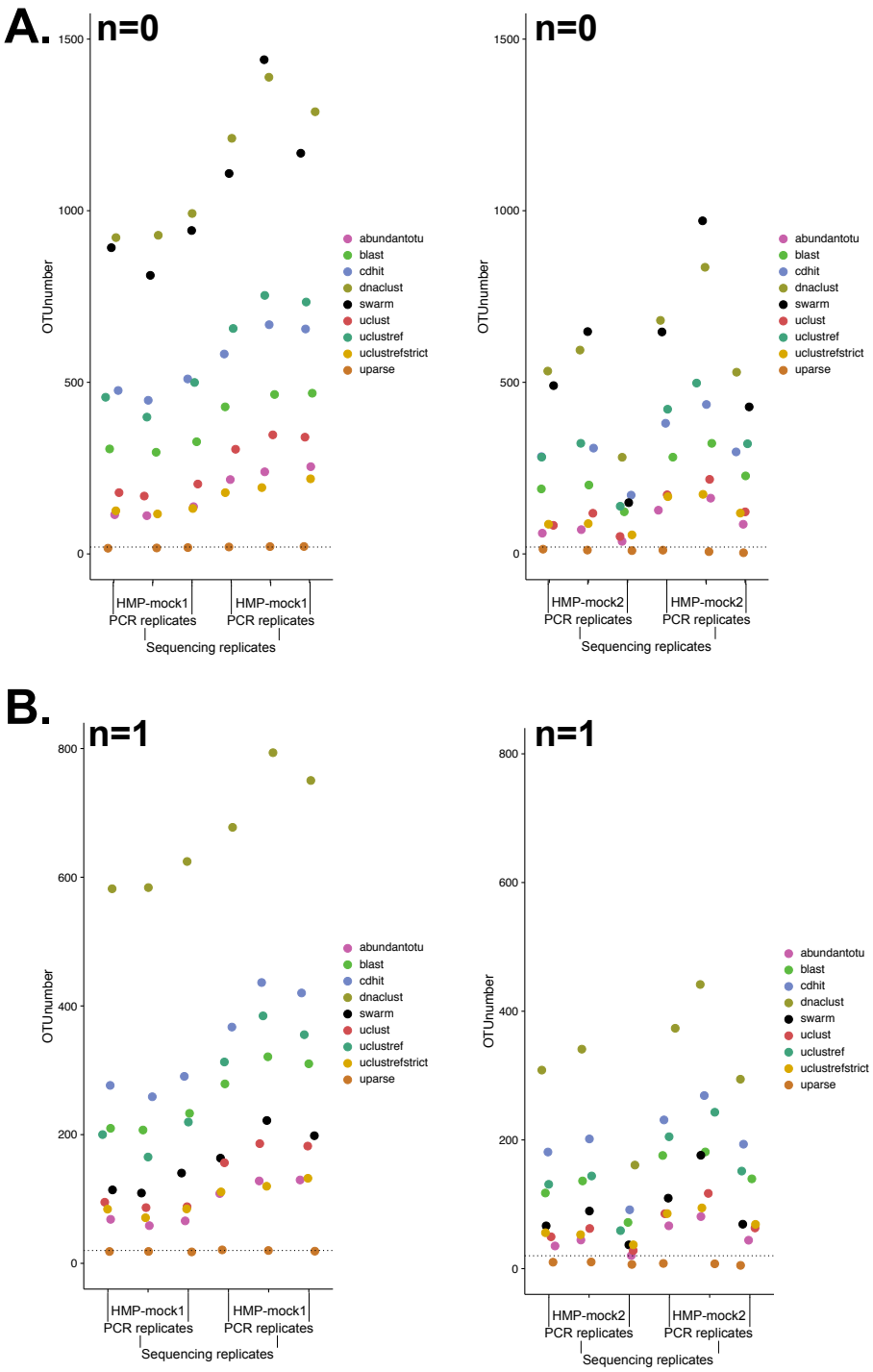


Figure A.6: **Swarm also over-estimates sample diversity. A.** When sl1p-generated quality filtered reads were used to pick OTUs with the Swarm algorithm, it also over-estimated within-sample diversity. **B.** However, maybe of these spurious OTUs are singletons, indicated by the decrease in the number of OTUs per sample after singletons are removed.

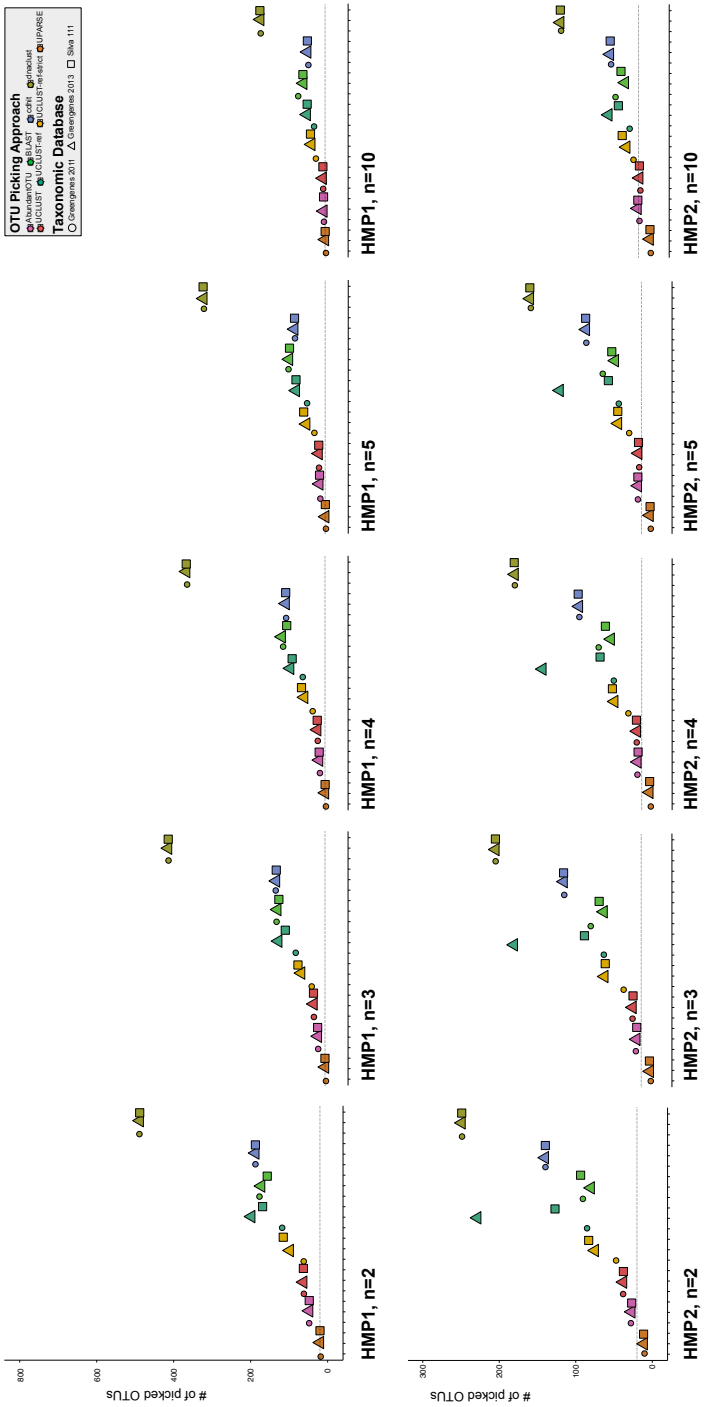


Figure A.7: **The number of observed OTUs converges on the expected community composition as low-abundance OTUs are removed.** OTUs with less than n reads were removed (n=2 to n=10); as n increases, the number of observed OTUs decreases towards the known sample diversity (n=20, dotted lined).

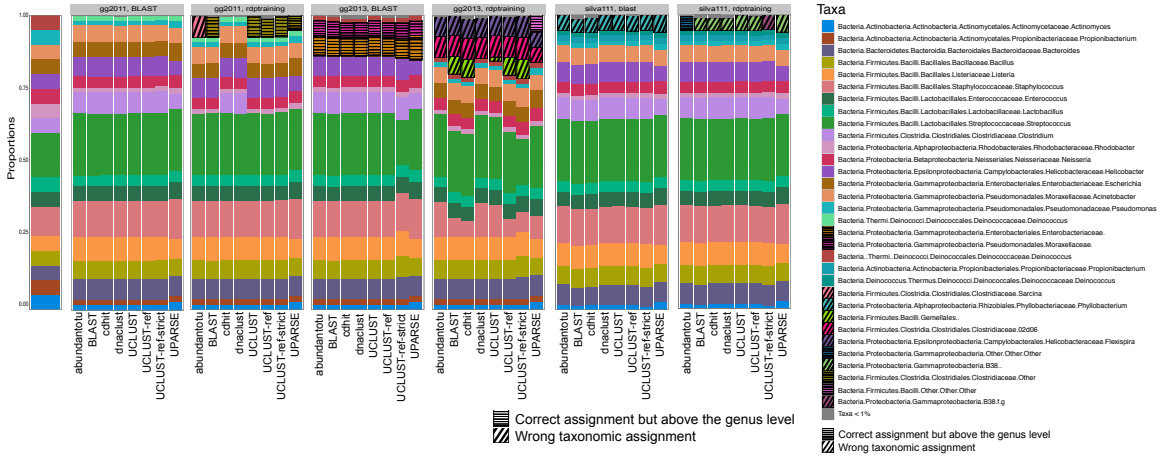


Figure A.8: **Taxa present in the taxonomic assignment of HMP-mock1.** For the first HMP mock community, the genus-level taxonomic assignments are compared to the known mock community in terms of taxonomic assignment and estimated proportions. Mis-assigned taxa are identified with overlaid patterns.

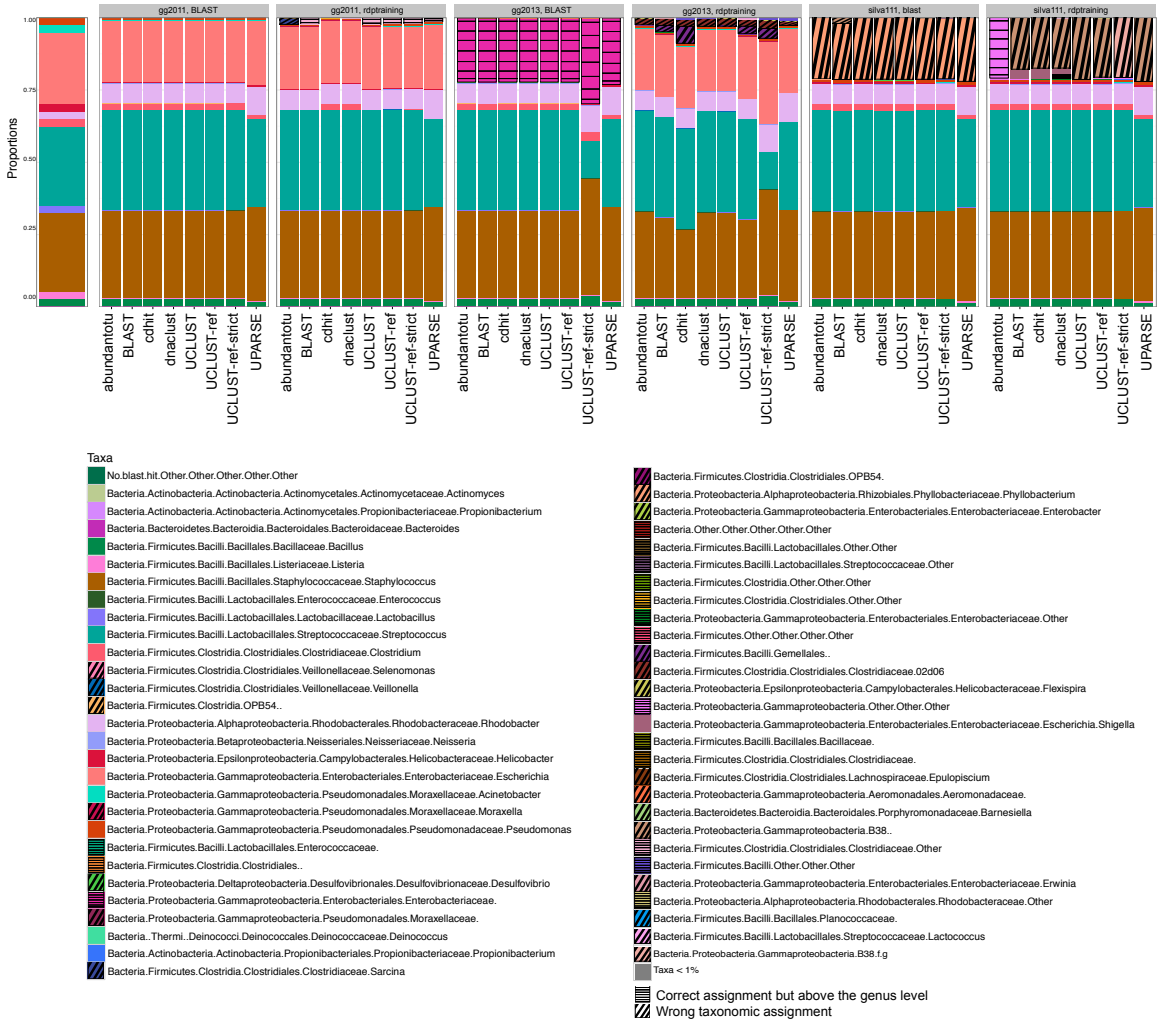


Figure A.9: **Taxon assignment of HMP-mock2.** Taxa were assigned to OTUs resulting from sl1p's options for OTU clustering, taxon assignment, and choice of reference database. Resulting taxa was compared to the known composition of the community The Human Microbiome Project Consortium (2012a) to determine correct taxa assignment.

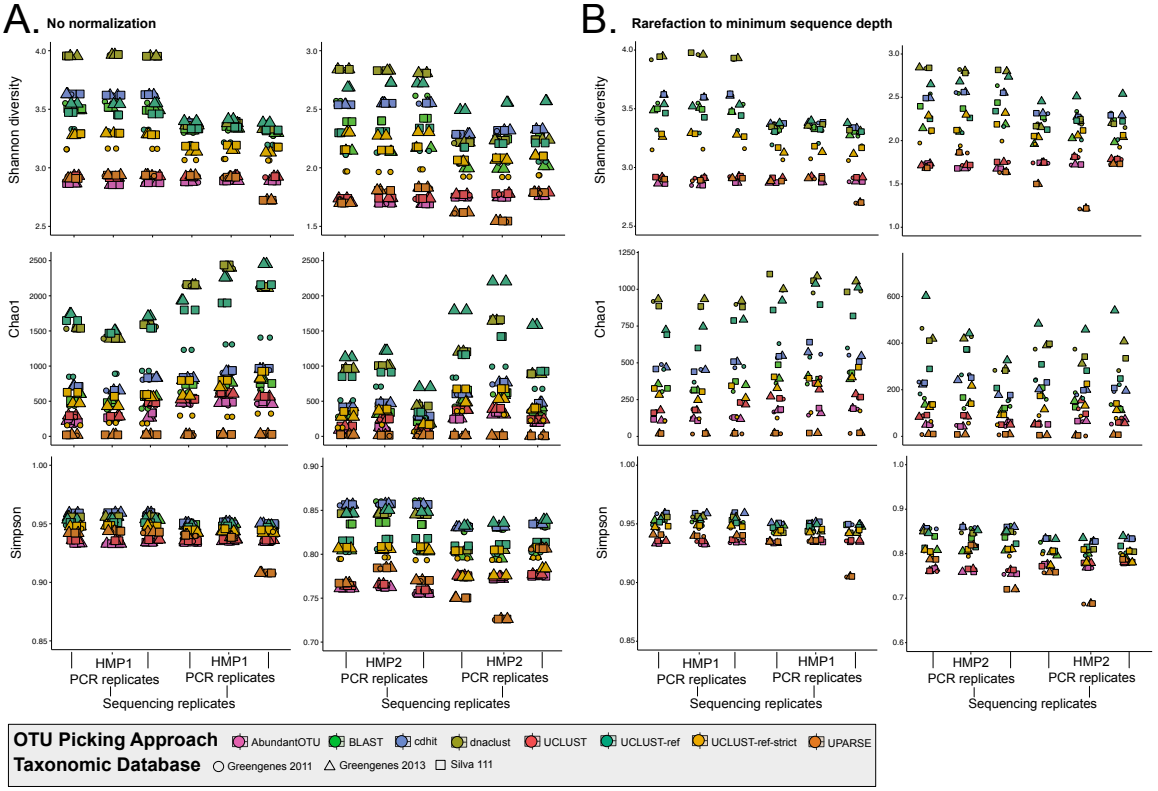


Figure A.10: **Effects of data processing and PCR/sequencing replicates on α diversity metrics.** Samples from the HMP-mock community were used to calculate Shannon, Chao1, and Simpson measures of α diversity. Together with **Figure 2.6**, these results indicate that choice of OTU clustering algorithm creates large variability in the resulting diversity output. Further, variation is observed across PCR and sequencing replicates, which is only partially mitigated by use of rarefaction (**B**).

Appendix B

Appendix to Chapter 3

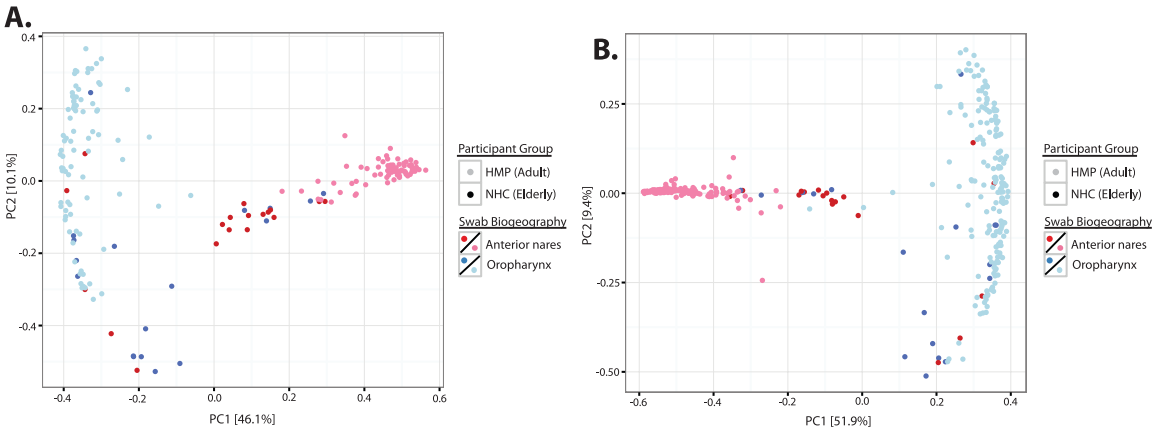


Figure B.1: **β-diversity calculated with the full-length v1-3 (A) and v3-5 (B) 16S rRNA sequences obtained from NIHs Human Microbiome Project produce similar clustering patterns as those trimmed to just the v3 region.** Because the Human Microbiome Project samples were processed with a technology that allowed for longer read-length then we obtained for the nursing home cohort, we trimmed the Human Microbiome Project sequences for OTU comparisons. However, taxonomic comparisons were completed on the full-length v1-3 (**A**) and v3-5 (**B**) Human Microbiome Project sequences with the nursing home cohort to ensure that trimming did not affect the dataset in any way.

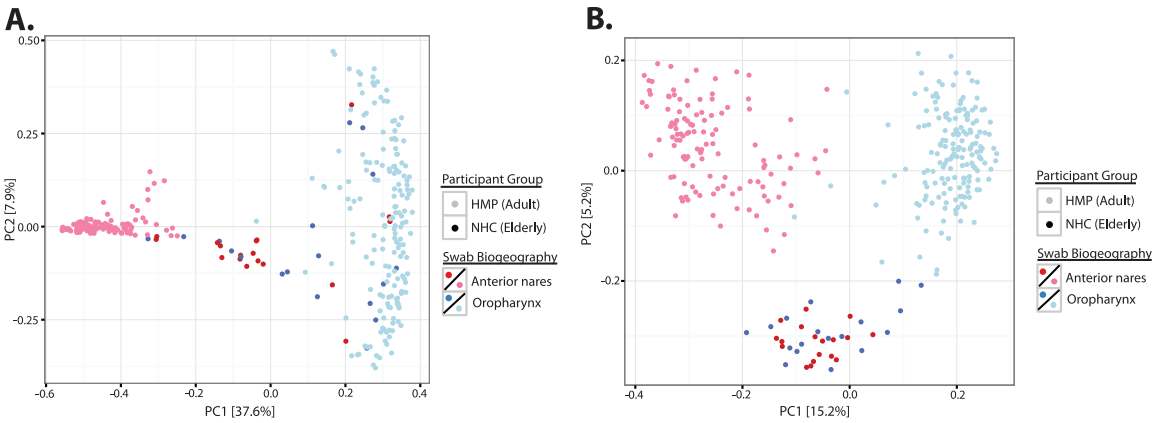


Figure B.2: **β-diversity calculated with measures other then weighted UniFrac produce similar clustering patterns.** In addition to the weighted UniFrac metrics used in Figure 3.2, Bray Curtis (**A**) and unweighted UniFrac (**B**) metrics were additionally used. The elderly anterior nares and oropharynx samples cluster together, as was seen using weighted UniFrac.

Table B.1: **Collected metadata from the nursing home cohort.**All metadata was examined in conjunction with β-diversity measures. No associations between the microbiota of the elderly and these data where discovered.

General	**Co-morbidities**	**Vaccination Information**
Age	COPD	Influenza (seasonal, 2009)
Gender	CHF	Influenza (H1N1, 2009)
Nursing room residence	Arrhythmia	Whether Influenza vaccine ever received
Swab biogeography	CAD	Whether Pneumonia vaccine ever received
# of medications	Asthma	Pneumovax (within last 5 years)
Cognition	Anemia	Whether Influenza prophylaxis received
Behaviour	Dementia	
Mood	CVA	
Education	DM	
Contact with children	Hypothyroid	
In shared room	Smoker	
Barthel score	Influenza within last year	
Frailty score		
Death during study		
Hospitalization during study		
Development of pneumonia during study		

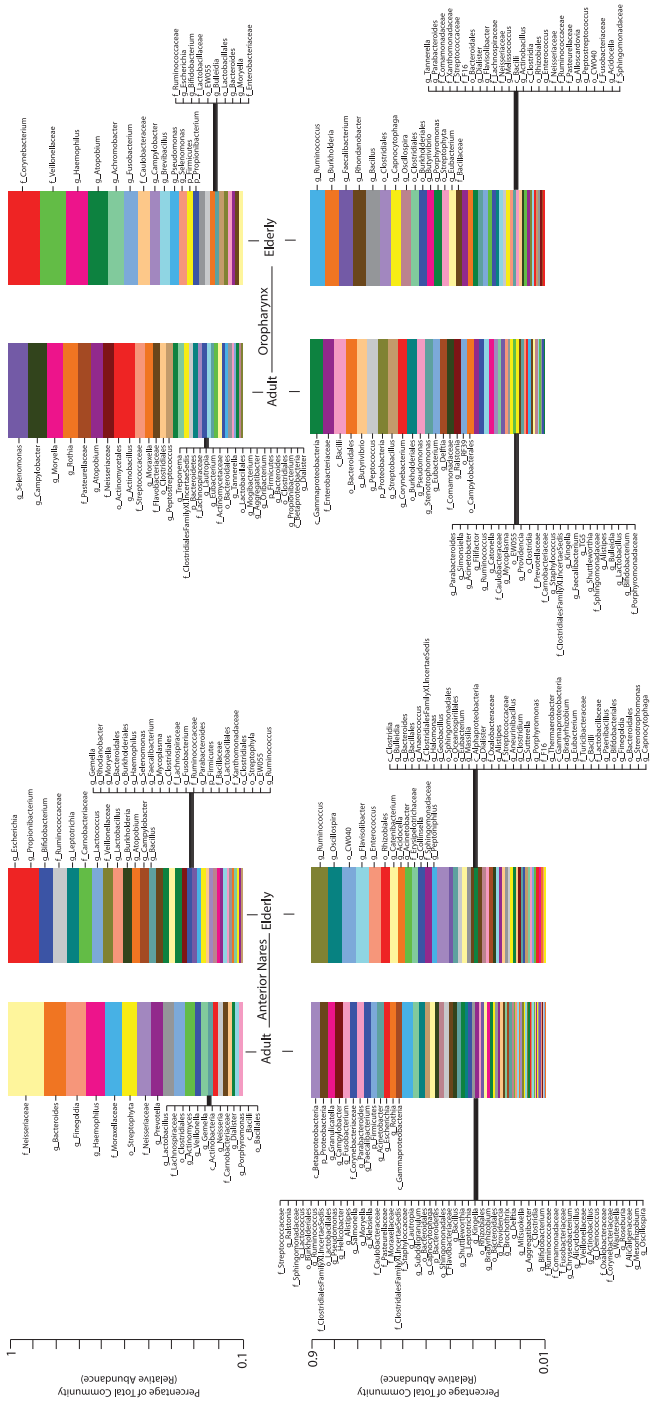


Figure B.3: **The taxonomic distributions of rare taxa (<1% relative abundance) of the mid-aged adult oropharynx, elderly adult oropharynx, mid-aged adult anterior nares, and elderly adult anterior nares.** Percentages indicate relative abundance of each taxon in the corresponding category. Abundances have been split into rare (between 0.1-1%) and very rare (0.01-0.1%) for visualization purposes.

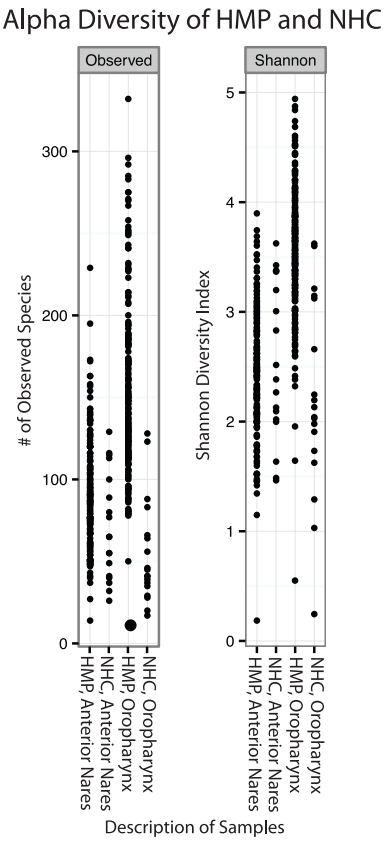


Figure B.4: **α-diversity measures calculated on the mid-aged adult and elderly anterior nares and oropharynx samples displays a loss of species diversity in the elderly oropharynx.** Observed species and Shannon α-diversity measures were calculated on the Human Microbiome Project and nursing home cohort samples grouped by biogeography. A decrease in Shannon diversity in the elderly oropharynx indicates a loss of species richness at this locale with age, which is seen as an increase in the Observed Species.

Table B.2: **Statistically significant differences between the adult (HMP) and elderly (NHC) oropharynx.**

Taxon	**Probability**	**Bonferroni corrected**	**FDR corrected**	**NHC Oropharynx mean**	**HMP Oropharynx mean**
Root;p_Firmicutes;c_Bacilli;o_Bacillales;f_Paenibacillaceae;g_Brevibacillus	1.52263329728E-018	2.4666659416E-016	2.4666659416E-016	0.0074643915	0
Root;p_Firmicutes;c_Bacilli;o_Bacillales;f_Bacillaceae;g_Bacillus	8.18259825899E-018	1.32558091796E-015	6.62790458978E-016	0.0021100069	0.000011039
Root;p_Proteobacteria;c_Betaproteobacteria;o_Burkholderiales;f_Alcaligenaceae;g_Achromobacter	1.4578798275E-017	2.36176532054E-015	7.87255106847E-016	0.0119308975	2.72692381142E-006
Root;p_Firmicutes;c_Bacilli;o_Bacillales;f_Staphylococcaceae;g_Staphylococcus	6.62540488638E-017	1.07331559159E-014	2.68328897898E-015	0.0392269878	0.0002486403
Root;p_Firmicutes;c_Bacilli;o_Bacillales;Other;Other	1.59209810065E-016	2.57919892305E-014	5.15839784609E-015	0.0003778416	2.28333961252E-006
Root;p_Proteobacteria;c_Gammaproteobacteria;o_Oceanospirillales;Other;Other	7.69545957278E-016	1.24666445079E-013	2.07777408465E-014	0.0002074548	0
Root;p_Proteobacteria;c_Alphaproteobacteria;o_Rhodospirillales;f_Acetobacteraceae;g_Acidocella	7.74766598889E-016	1.25512189102E-013	1.79303127172E-014	0.0003904411	0
Root;p_Firmicutes;c_Bacilli;o_Bacillales;f_Bacillaceae;Other	1.34427927787E-015	2.17773243015E-013	2.72216553769E-014	0.0012247119	4.78391647284E-006
Root;p_Proteobacteria;c_Gammaproteobacteria;o_Enterobacteriales;f_Enterobacteriaceae;g_Escherichia	1.54506721784E-015	2.5030088929E-013	2.78112099211E-014	0.0042282771	9.03186570558E-005
Root;p_Firmicutes;c_Bacilli;o_Bacillales;f_Paenibacillaceae;g_Aneurinibacillus	5.75977112418E-015	9.33082922118E-013	9.33082922118E-014	0.0003183812	0
Root;p_Proteobacteria;c_Gammaproteobacteria;o_Xanthomonadales;f_Xanthomonadaceae;g_Rhodanobacter	1.94787444701E-014	3.15555660415E-012	2.86868782196E-013	0.0044540281	0
Root;p_Actinobacteria;c_Actinobacteria;o_Bifidobacteriales;f_Bifidobacteriaceae;g_Bifidobacterium	8.73280108817E-014	1.41471377628E-011	1.1789281469E-012	0.0020635087	9.05535192035E-005

Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_	3.0193838692E-013	4.89140186811E-011	3.76261682162E-012	0.0003977036	8.30552247132E-006
Root;p_Cyanobacteria;c_Chloroplast;o_Streptophyta;f_;g_	6.86825893975E-013	1.11265794824E-010	7.94755677314E-012	0.0015093954	5.38674280623E-005
Root;p_Bacteroidetes;c_Sphingobacteria;o_Sphingobacteriales;f_;g_Flavisolibacter	7.66008000215E-013	1.24093296035E-010	8.27288640232E-012	0.0008602496	1.55038759 69E-006
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Eubacteriaceae;Other	3.4111258143E-012	5.52602381917E-010	3.45376488698E-011	0.0001191108	0
Root;p_Proteobacteria;c_Alphaproteobacteria;o_Caulobacterales;f_Caulobacteraceae;Other	8.82329715475E-011	1.42937413907E-008	8.40808317 1E-010	0.0195828405	0.0004251017
Root;p_Proteobacteria;c_Alphaproteobacteria;Other;Other;Other	9.72754256177E-011	1.57586189501E-008	8.75478830559E-010	0.0003281847	5.85278390069E-006
Root;p_Tenericutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_Clostridium	2.55604684863E-010	4.14079589478E-008	2.17936626041E-009	0.0001738291	2.68884232491E-006
Root;p_Proteobacteria;c_Alphaproteobacteria;o_Sphingomonadales;Other;Other	1.26858333312E-009	2.05510499965E-007	1.02755249983E-008	0.0004986291	8.31654359698E-006
Root;p_Proteobacteria;c_Betaproteobacteria;o_Burkholderiales;Other;Other	1.38659701862E-009	2.24628717016E-007	1.06966055722E-008	0.0026853943	4.17984696 06E-005
Root;p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;Other;Other	4.28656809936E-009	6.94424032096E-007	3.15647287316E-008	0.05660498	0.005458826
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;Other	5.39511137084E-009	0.000000874	0.000000038	0.0021767337	3.82788334623E-005
Root;p_Proteobacteria;c_Gammaproteobacteria;o_Pseudomonadales;f_Pseudomonadaceae;g_Pseudomonas	7.27394585089E-009	1.17837922784E-006	4.90991344935E-008	0.0074832963	0.0007038249
Root;p_Tenericutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_	1.12501088391E-008	1.82251763193E-006	7.29007052771E-008	0.0001581385	0.000008308
Root;p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Lactobacillaceae;Other	2.13650467575E-008	3.46113757471E-006	1.33120675 95E-007	0.0008409203	2.16477055375E-006

Root;p_Tenericutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_Bulleidia	2.27445683508E-008	3.68462007283E-006	1.36467410105E-007	0.0006235247	8.86727297071E-005
Root;p_Proteobacteria;c_Alphaproteobacteria;o_Rhizobiales;Other;Other	7.10662729184E-008	1.15127362128E-005	4.11169150457E-007	0.0009993374	6.90206433972E-005
Root;p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Carnobacteriaceae;g_	1.00136460467E-007	1.62221065957E-005	5.59382986059E-007	0.0179715613	0.000307497
Root;p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Enterococcaceae;g_Enterococcus	1.04809857996E-007	1.69791969953E-005	0.000000566	0.0005427867	0.000019976
Root;p_Firmicutes;c_Bacilli;o_Bacillales;f_Planococcaceae;Other	1.71233042751E-007	2.77397529256E-005	8.94830739536E-007	8.26849316929E-005	7.08155114298E-007
Root;p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Corynebacteriaceae;g_Corynebacterium	1.7550515247E-007	2.84318347001E-005	8.88494834379E-007	0.0224266348	0.0008113172
Root;p_Actinobacteria;c_Actinobacteria;o_Coriobacteriales;f_Coriobacteriaceae;g_Collinsella	2.24660972556E-007	3.63950775541E-005	1.102881138E-006	0.0001706675	1.17063461608E-005
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Ruminococcus	7.25207632916E-007	0.0001174836	3.45540107448E-006	0.0002761977	1.66089390363E-005
Root;p_Proteobacteria;c_Gammaproteobacteria;o_Pseudomonadales;f_Moraxellaceae;g_	1.01364993102E-006	0.0001642113	4.6917511093E-006	0.0001257665	8.06290285792E-006
Root;p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Lactobacillaceae;g_Lactobacillus	1.1745230986E-006	0.0001902727	5.28535394371E-006	0.0361669508	0.0001397069
Root;p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Streptococcaceae;g_Lactococcus	0.000001176	0.0001905103	5.1489269928E-006	0.0195275065	8.91417762353E-006
Root;p_Proteobacteria;c_Betaproteobacteria;o_Burkholderiales;f_Burkholderiaceae;g_Burkholderia	1.33045998056E-005	0.0021553452	5.67196096974E-005	0.0039041193	3.52315258837E-005
Root;p_TM7;c_TM7-3;o_EW055;f_;g_	3.88608274887E-005	0.0062954541	0.0001614219	0.0011363563	0.0001677786

Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Dorea	5.85537897122E-005	0.0094857139	0.0002371428	6.37281013629E-005	6.15172175705E-006
Root;p_Proteobacteria;c_Gammaproteobacteria;o_Xanthomonadales;f_Xanthomonadaceae;Other	7.01418497495E-005	0.0113629797	0.0002771458	0.0016420555	0.0001268279
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_ClostridialesFamilyXI.IncertaeSedis;g_Anaerococcus	7.0899744997E-005	0.0114857587	0.0002734704	0.0005556686	6.95709276259E-006
Root;p_Bacteroidetes;Other;Other;Other;Other	0.0002264051	0.0366776257	0.000852968	2.26315473148E-005	0.0026534265
Root;p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Microbacteriaceae;g_Microbacterium	0.0002466908	0.0399639068	0.0009082706	6.32794646736E-005	6.69758463853E-006
Root;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;Other;Other	0.0003055278	0.0494955003	0.0010999	0.000068862	0.0020174219
Root;p_Proteobacteria;c_Betaproteobacteria;o_Burkholderiales;f_Alcaligenaceae;g_Alcaligenes	0.0004052272	0.0656468054	0.0014271045	5.14797512125E-005	4.42394110141E-006
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_ClostridialesFamilyXI.IncertaeSedis;g_Peptoniphilus	0.0004344504	0.0703809682	0.0014974674	0.0003083005	7.16882846644E-006
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Faecalibacterium	0.0007253053	0.1174994586	0.0024479054	0.0009604389	0.0001578148
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Veillonellaceae;g_Veillonella	0.000818025	0.1325200451	0.0027044907	0.0366590338	0.0823913909
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_ClostridialesFamilyXIII.IncertaeSedis;g_Mogibacterium	0.0008668334	0.1404270187	0.0028085404	5.56586698987E-005	0.0014489424
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Oscillospira	0.0015071631	0.2441604193	0.0047874592	0.0008106552	0.0001019948
Root;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Porphyromonadaceae;g_Porphyromonas	0.0020620785	0.334056712	0.0064241675	0.0004674616	0.0280018297

Root;p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Propionibacteriaceae;g_Propionibacterium	0.0038113016	0.6174308584	0.0116496388	0.0092584512	0.002086088
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Peptostreptococcaceae;g_Peptostreptococcus	0.005030085	0.8148737674	0.015090255	7.18761588871E-005	0.0028549189
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;Other	0.0056546662	0.9160559288	0.0166555623	0.0180204741	0.0411892007
Root;p_Bacteroidetes;c_Flavobacteria;o_Flavobacteriales;f_Flavobacteriaceae;Other	0.0067001207	1.0854195604	0.0193824921	0.0001163892	0.0033968682
Root;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Prevotellaceae;g_Prevotella	0.008220732	1.3317585823	0.0233641857	0.0641497769	0.1229637604
Root;p_Proteobacteria;c_Epsilonproteobacteria;o_Campylobacterales;Other;Other	0.0104535369	1.6934729769	0.0291978099	1.63837636576E-006	0.000381302
Root;p_Proteobacteria;c_Gammaproteobacteria;o_Pasteurellales;f_Pasteurellaceae;g_Haemophilus	0.0113719105	1.8422495061	0.0312245679	0.0065153546	0.0609253254
Root;p_Proteobacteria;c_Gammaproteobacteria;o_Pasteurellales;f_Pasteurellaceae;Other	0.0114902534	1.8614210519	0.0310236842	0.0001629355	0.0068855716
Root;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Prevotellaceae;Other	0.0141021118	2.2845421101	0.03745151	0	0.0001978204
Root;p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Streptococcaceae;Other	0.0159101115	2.5774380623	0.0415715817	0.0002156509	0.0059292257
Root;p_Proteobacteria;c_Epsilonproteobacteria;o_Campylobacterales;f_Campylobacteraceae;g_Campylobacter	0.0173279666	2.807130591	0.0445576284	0.0018108607	0.0071035577

Table B.3: **Statistically significant differences between the adult (HMP) and elderly (NHC) anterior nares.**

Taxon	**Probability**	**Bonferroni corrected**	**FDR corrected**	**NHC Anterior nares mean**	**HMP Anterior nares mean**
Root;p_Firmicutes;c_Bacilli;o_Bacillales;f_Paenibacillaceae;g_Brevibacillus	4.48378911153E-036	8.16049618298E-034	8.16049618298E-034	0.0170342491	0
Root;p_Proteobacteria;c_Betaproteobacteria;o_Burkholderiales;f_Alcaligenaceae;g_Achromobacter	1.51063860472E-035	2.74936226059E-033	1.3746811303E-033	0.0288724689	1.95877560854E-006
Root;p_Proteobacteria;c_Betaproteobacteria;o_Neisseriales;f_Neisseriaceae;g_Neisseria	1.45169737983E-034	2.64208923129E-032	8.80696410429E-033	0.0460974087	0.0016442119
Root;p_Proteobacteria;c_Gammaproteobacteria;o_Xanthomonadales;f_Xanthomonadaceae;g_Rhodanobacter	8.9562560471E-029	1.6300386057E-026	4.0750965143E-027	0.0047119504	0
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Veillonellaceae;g_Selenomonas	1.1025143644E-027	2.0065761432E-025	4.0131522864E-026	0.0030062084	0.0001326603
Root;p_Proteobacteria;c_Gammaproteobacteria;o_Pseudomonadales;f_Pseudomonadaceae;g_Pseudomonas	2.66728937425E-026	4.85446666113E-024	8.09077776855E-025	0.0159669873	0.0007149752
Root;p_Proteobacteria;c_Gammaproteobacteria;o_Xanthomonadales;f_Xanthomonadaceae;Other	6.56072857E-022	1.19405259974E-019	1.7057894282E-020	0.0018777615	2.34431061967E-005
Root;p_Proteobacteria;c_Gammaproteobacteria;o_Oceanospirillales;Other;Other	3.34191857051E-021	6.08229179833E-019	7.6028647479 2E-020	0.0003028255	4.75219670291E-006
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_	5.8539727401E-021	1.0654230387E-018	1.1838033763 3E-019	0.0020890667	0.000072552
Root;p_Proteobacteria;c_Alphaproteobacteria;o_Rhodospirillales;f_Acetobacteraceae;g_Acidocella	1.31990485437E-020	2.40222683495E-018	2.4022268349 5E-019	0.0006250025	0

Root;p_Firmicutes;c_Bacilli;o_Bacillales;f_Bacillaceae;g_Bacillus	4.08552405449E-017	7.43565377917E-015	6.75968525379E-016	0.00496801 97	0.00018341 96
Root;p_Proteobacteria;c_Betaproteobacteria;o_Burkholderiales;Other;Other	1.12337554667E-016	2.04454349494E-014	1.70378624579E-015	0.00514540 04	0.00011911 85
Root;p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Enterococcaceae;g_Enterococcus	1.06287018288E-015	1.93442373285E-013	1.48801825604E-014	0.00096820 31	1.61556111471E-005
Root;p_Proteobacteria;c_Alphaproteobacteria;o_Caulobacterales;f_Caulobacteraceae;Other	1.12300367886E-015	2.04386669553E-013	1.45990478252E-014	0.03194443 28	0.00116494 83
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Ruminococcus	4.08363170576E-015	7.43220970448E-013	4.95480646965E-014	0.00105845	2.08294575281E-005
Root;p_Firmicutes;c_Bacilli;o_Bacillales;f_Paenibacillaceae;g_Aneurinibacillus	5.41086315688E-015	9.84777094551E-013	6.15485684095E-014	0.00022028 15	0
Root;p_Proteobacteria;c_Betaproteobacteria;o_Burkholderiales;f_Burkholderiaceae;g_Burkholderia	1.38159344799E-014	2.51450007534E-012	1.47911769138E-013	0.01007861 11	3.16235333714E-005
Root;p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Micrococcaceae;g_Rothia	2.66788442154E-014	4.85554964721E-012	2.69752758178E-013	0.03087497 26	0.00060577 28
Root;p_Firmicutes;c_Bacilli;o_Bacillales;f_Paenibacillaceae;g_Paenibacillus	3.6933665544E-014	6.721927129E-012	3.53785638368E-013	0.00014304 85	4.09854068608E-006
Root;p_Proteobacteria;c_Gammaproteobacteria;o_Enterobacteriales;f_Enterobacteriaceae;g_Escherichia	1.33389796825E-013	2.42769430222E-011	1.21384715111E-012	0.00938600 56	0.00081863 67
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Veillonellaceae;g_Veillonella	1.4907128254E-013	2.71309734224E-011	1.29195111535E-012	0.03919151 66	0.00191673 35
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_	2.30601421191E-013	4.19694586568E-011	1.90770266622E-012	0.00171810 89	5.80015657475E-005
Root;p_Bacteroidetes;c_Sphingobacteria;o_Sphingobacteriales;f_;g_Flavisolibacter	5.52306407613E-013	1.00519766186E-010	4.37042461676E-012	0.00182005 01	1.06058368729E-005
Root;p_Tenericutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_Bulleidia	1.61368189711E-012	2.93690105273E-010	1.22370877197E-011	0.00020879 01	0

Root;p_Proteobacteria;c_Betaproteobacteria;o_Burkholderiales;f_Oxalobacteraceae;g_Massilia	3.05309486 78E-012	5.55663265 94E-010	2.22265306 376E-011	0.00021189 08	7.23593523 212E-006
Root;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Prevotellaceae;g_Prevotella	3.14259677 91E-012	5.71952613 796E-010	2.19981774 537E-011	0.05904315 8	0.00406824 68
Root;p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Leuconostocaceae;g_Leuconostoc	1.09353438 681E-011	0.00000000 2	7.37123179 26E-011	9.39533394 962E-005	2.56765548 302E-006
Root;p_Actinobacteria;c_Actinobacteria;o_Coriobacteriales;f_Coriobacteriaceae;g_Atopobium	1.41162313 984E-011	2.56915411 451E-009	9.17555040 897E-011	0.00300264	3.41632181 954E-005
Root;p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Streptococcaceae;g_Lactococcus	2.79041746 082E-011	5.07855977 87E-009	1.75122750 99E-010	0.00732596 23	0.00055172 24
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_;g_	5.38215379 602E-011	9.79551990 876E-009	3.26517330 292E-010	0.00134456 98	0.00011164 73
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Bacteroides	6.57161883 902E-011	0.00000001 2	3.85817622 162E-010	0.00030319 39	2.64026751 19E-006
Root;p_Tenericutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_	1.09402433 416E-009	1.99112428 816E-007	6.22226340 051E-009	0.00046366 91	2.69202172 148E-005
Root;p_Tenericutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae;g_Catenibacterium	1.38281957 284E-009	2.51673162 257E-007	7.62645946 235E-009	0.00076317 43	8.47947922 09E-006
Root;p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Streptococcaceae;g_Streptococcus	1.59533857 577E-009	2.90351620 79E-007	8.53975355 265E-009	0.18463147 37	0.02557075 22
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Veillonellaceae;Other	1.72629054 446E-009	3.14184879 092E-007	0.00000000 9	0.00388974 78	0.00013352 22
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;Other	2.20449119 047E-009	4.01217396 665E-007	1.11449276 851E-008	0.00685697 07	0.00020684 67
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Oscillospira	8.76547650 659E-009	1.59531672 42E-006	4.31166682 216E-008	0.00112737 88	0.00010322 44

Root;p_Firmicutes;c_Bacilli;o_Bacillales;f_Bacillaceae;g_Geobacillus	3.04162766547E-008	5.53576235115E-006	1.45677956609E-007	0.0001996897	4.32225729182E-006
Root;p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Propionibacteriaceae;g_Propionibacterium	5.05982768963E-008	9.20888639512E-006	2.36125292183E-007	0.0115111161	0.2890399035
Root;p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Carnobacteriaceae;g_Granulicatella	6.75884756396E-008	1.23011025664E-005	3.07527564 16E-007	0.00837769 2	0.0011811557
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_ClostridialesFamilyXIII.IncertaeSedis;g_Eubacterium	8.91780211216E-008	1.62303998441E-005	3.95863410832E-007	0.0001094518	4.50131855846E-006
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;Other	1.22091236121E-007	2.22206049 74E-005	5.29062023 19E-007	0.0347780467	0.0031428076
Root;p_Actinobacteria;c_Actinobacteria;o_Bifidobacteriales;f_Bifidobacteriaceae;g_Bifidobacterium	5.99390813283E-007	0.0001090891	0.00000253 7	0.0083517251	0.0004526352
Root;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_;g_	2.83588418731E-006	0.0005161309	1.17302482293E-005	0.00301027 5	0.0001910973
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_ClostridialesFamilyXIII.IncertaeSedis;g_Mogibacterium	5.63071652575E-006	0.0010247904	2.27731201708E-005	5.7291524054E-005	3.65111346711E-006
Root;p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Corynebacteriaceae;g_Corynebacterium	8.07088653289E-006	0.0014689013	3.19326380214E-005	0.0354249249	0.2497395669
Root;p_Firmicutes;c_Bacilli;o_Bacillales;f_Staphylococcaceae;g_Jeotgalicoccus	1.18335521945E-005	0.0021537065	4.58235425406E-005	0.0001351 74	0.0000062 13
Root;p_Fusobacteria;c_Fusobacteria;o_Fusobacteriales;f_Fusobacteriaceae;g_Leptotrichia	1.45470545009E-005	0.0026475639	5.51575816492E-005	0.0023224846	0.0002751597
Root;p_TM7;c_TM7-3;o_EW055;f_;g_	1.94094247343E-005	0.0035325153	7.20921490133E-005	0.0005059568	0
Root;p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Promicromonosporaceae;g_Cellulosimicrobium	2.82717467738E-005	0.0051454579	0.0001029092	0.00018424	0.0000128 58

Root;p_Actinobacteria;c_Actinobacteria;o_Actinomycetales;f_Actinomycetaceae;g_Actinomyces	6.01504976839E-005	0.0109473906	0.0002146547	0.0069950025	0.0017130726
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae;g_Eubacterium	6.31109287325E-005	0.011486189	0.0002208883	0.0003094365	2.24051246002E-005
Root;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Prevotellaceae;g_	9.55246455398E-005	0.0173854855	0.000328028	8.33722386373E-005	9.48577989385E-006
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae;g_Moryella	0.0001268347	0.0230839075	0.0004274798	0.0034396639	0.0007140945
Root;p_Proteobacteria;c_Gammaproteobacteria;o_Enterobacteriales;f_Enterobacteriaceae;Other	0.0003415155	0.062155826	0.0011301059	0.0704520111	0.0129498684
Root;p_Actinobacteria;c_Actinobacteria;o_Coriobacteriales;f_Coriobacteriaceae;g_Collinsella	0.0009994507	0.1819000262	0.0032482148	0.0003911761	6.2005468232E-005
Root;p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Lactobacillaceae;Other	0.0012713645	0.2313883414	0.0040594446	0.0001345444	0.000012811
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_ClostridialesFamilyXI.IncertaeSedis;g_Anaerococcus	0.0036962424	0.6727161137	0.0115985537	0.0005566854	0.0119765631
Root;p_Firmicutes;c_Clostridia;o_Clostridiales;f_ClostridialesFamilyXI.IncertaeSedis;g_Peptoniphilus	0.0043144196	0.7852243674	0.0133088876	0.0005577223	0.0124391531
Root;p_Proteobacteria;c_Alphaproteobacteria;Other;Other;Other	0.0070531231	1.2836684	0.0213944733	0.000432218	7.96693108683E-005
Root;p_Proteobacteria;c_Alphaproteobacteria;o_Rhodospirillales;f_Rhodospirillaceae;Other	0.0126507082	2.3024288986	0.037744736	0.0001309075	2.04844645786E-005
Root;p_Proteobacteria;c_Gammaproteobacteria;o_Xanthomonadales;f_Xanthomonadaceae;g_Xanthomonas	0.0127807942	2.3261045433	0.0375178152	4.23478866879E-005	6.17914097449E-006

Table B.4: **The top five BLAST hits of the *Streptococcus* OTUs against NCBIs rRNA Reference database.**

O T U #	Description	Query cover	E-value	Identity
1	Streptococcus salivarius CCHSS3 strain CCHSS3 16S ribosomal RNA, complete sequence	100.00%	2E-078	100.00%
	Streptococcus salivarius strain ATCC 7073 16S ribosomal RNA, complete sequence	00.00%	2E-078	100.00%
	Streptococcus thermophilus MN-ZLW-002 strain MN-ZLW-002 16S ribosomal RNA, complete sequence	100.00%	8E-077	99.00%
	Streptococcus thermophilus strain ATCC 19258 16S ribosomal RNA, complete sequence	100.00%	8E-077	99.00%
	Streptococcus vestibularis ATCC 49124 strain ATCC 49124 16S ribosomal RNA, complete sequence	100.00%	3E-076	99.00%
2	Streptococcus oralis Uo5 strain Uo5 16S ribosomal RNA, complete sequence	100.00%	2E-078	100.00%
	Streptococcus mitis B6 strain B6 16S ribosomal RNA, complete sequence	100.00%	2E-078	100.00%
	Streptococcus pseudopneumoniae IS7493 strain IS7493 16S ribosomal RNA, complete sequence	100.00%	2E-078	100.00%
	Streptococcus pneumoniae R6 strain R6 16S ribosomal RNA, complete sequence	100.00%	2E-078	100.00%
	Streptococcus australis strain AI-1 16S ribosomal RNA, partial sequence	100.00%	2E-078	100.00%
54	Streptococcus sinensis strain HKU4 16S ribosomal RNA, partial sequence	98.00%	7E-067	99.00%
	Streptococcus pasteurianus ATCC 43144 strain ATCC 43144 16S ribosomal RNA, complete sequence	98.00%	2E-063	98.00%
	Streptococcus infantarius subsp. infantarius CJ18 strain CJ18 16S ribosomal RNA, complete sequence	98.00%	2E-063	98.00%
	Streptococcus sanguinis SK36 strain SK36 16S ribosomal RNA, complete sequence	98.00%	2E-063	98.00%
	Streptococcus gallolyticus UCN34 strain UCN34 16S ribosomal RNA, complete sequence	98.00%	2E-063	98.00%

55	Streptococcus oralis Uo5 strain Uo5 16S ribosomal RNA, complete sequence	99.00%	0	99.00%
	Streptococcus mitis B6 strain B6 16S ribosomal RNA, complete sequence	99.00%	0	99.00%
	Streptococcus pseudopneumoniae IS7493 strain IS7493 16S ribosomal RNA, complete sequence	99.00%	0	99.00%
	Streptococcus infantis ATCC 700779 strain ATCC 700779 16S ribosomal RNA, partial sequence	99.00%	0	99.00%
	Streptococcus oralis ATCC 35037 strain ATCC 35037 16S ribosomal RNA, partial sequence	99.00%	0	99.00%
64	Streptococcus salivarius CCHSS3 strain CCHSS3 16S ribosomal RNA, complete sequence	100.00%	0	99.00%
	Streptococcus salivarius strain ATCC 7073 16S ribosomal RNA, complete sequence	100.00%	0	99.00%
	Streptococcus thermophilus strain ATCC 19258 16S ribosomal RNA, complete sequence	100.00%	0	99.00%
	Streptococcus vestibularis ATCC 49124 strain ATCC 49124 16S ribosomal RNA, complete sequence	100.00%	0	99.00%
	Streptococcus thermophilus MN-ZLW-002 strain MN-ZLW-002 16S ribosomal RNA, complete sequence	100.00%	0	99.00%
91	Streptococcus oralis Uo5 strain Uo5 16S ribosomal RNA, complete sequence	100.00%	0	99.00%
	Streptococcus mitis B6 strain B6 16S ribosomal RNA, complete sequence	100.00%	0	99.00%
	Streptococcus oralis ATCC 35037 strain ATCC 35037 16S ribosomal RNA, partial sequence	100.00%	0	99.00%
	Streptococcus mitis strain NS51 16S ribosomal RNA, complete sequence	100.00%	0	99.00%
	Streptococcus infantis ATCC 700779 strain ATCC 700779 16S ribosomal RNA, partial sequence	100.00%	0	99.00%
159	Streptococcus sanguinis SK36 strain SK36 16S ribosomal RNA, complete sequence	98.00%	3E-070	100.00%
	Streptococcus sanguinis strain ATCC 10556 16S ribosomal RNA, partial sequence	98.00%	3E-070	100.00%
	Streptococcus sinensis strain HKU4 16S ribosomal RNA, partial sequence	98.00%	7E-067	99.00%

	Streptococcus peroris strain GTC848 16S ribosomal RNA, partial sequence	98.00%	3E-065	98.00%
	Streptococcus pasteurianus ATCC 43144 strain ATCC 43144 16S ribosomal RNA, complete sequence	98.00%	2E-063	97.00%
318	Streptococcus oralis Uo5 strain Uo5 16S ribosomal RNA, complete sequence	96.00%	3E-055	97.00%
	Streptococcus mitis B6 strain B6 16S ribosomal RNA, complete sequence	96.00%	3E-055	97.00%
	Streptococcus pseudopneumoniae IS7493 strain IS7493 16S ribosomal RNA, complete sequence	96.00%	3E-055	97.00%
	Streptococcus pneumoniae R6 strain R6 16S ribosomal RNA, complete sequence	96.00%	3E-055	97.00%
	Streptococcus australis strain AI-1 16S ribosomal RNA, partial sequence	96.00%	3E-055	97.00%

Table B.5: **Results of the *S. pneumoniae* specific PCR conducted on 123 nursing home residents.** Results show that only 7 of the 123 were positive for *S. pneumoniae* carriage, of which 4 tested positive with a secondary PCR reaction using another Streptococcus-specific gene and follow-up sequencing.

Sample #	**McAvin SPN Real-time PCR**	**Crossing Point**	**Comments**	**Vaccinated in last 5 years**
1	negative			Yes
2	negative			Yes
3	Positive	33.90	Positive by 2nd PCR and by sequencing	Yes
4	negative			Yes
5	negative			Yes
6	negative			Yes
7	negative			Yes
8	negative			Yes
9	negative			Yes
10	negative			Yes
11	negative			Yes
12	negative			Unknown

13	negative			Yes
14	negative			Yes
15	negative			Yes
16	negative			Yes
17	negative			Yes
18	negative			Yes
19	negative			Yes
20	negative			Yes
21	negative			Unknown
22	negative			Yes
23	negative			Yes
24	Positive	38.99	Negative by SPN nested-PCR	Yes
25	negative			No
26	negative			Yes
27	negative			No
28	negative			Unknown
29	negative			Yes
30	negative			Yes
31	negative			No
32	negative			No
33	negative			Yes
34	negative			Yes
35	Positive	37.56	Positive by 2nd PCR and by sequencing	Yes
36	negative			Yes
37	negative			No
38	negative			No
39	negative			No
40	negative			Yes
41	negative			No
42	negative			Yes
43	negative			No
44	negative			No
45	negative			No
46	negative			No

47	negative			Yes
48	negative			Yes
49	negative			No
50	negative			Yes
51	negative			Yes
52	negative			No
53	negative			No
54	negative			No
55	negative			Yes
56	negative			Yes
57	negative			No
58	negative			No
59	negative			No
60	negative			No
61	negative			Yes
62	negative			Unknown
63	Positive	34.57	Positive by 2nd PCR and by sequencing	No
64	Positive	37.54	Negative by SPN nested-PCR	No
65	negative			Unknown
66	negative			Unknown
67	negative			No
68	negative			Yes
69	negative			No
70	negative			Yes
71	negative			Yes
72	negative			No
73	negative			No
74	negative			Yes
75	negative			No
76	negative			Yes
77	negative			No
78	negative			Yes
79	negative			Yes
80	Positive	39.20	Negative by SPN nested-PCR	No

81	negative			No
82	negative			No
83	negative			Unknown
84	negative			Unknown
85	negative			Yes
86	negative			No
87	negative			Yes
88	negative			Yes
89	negative			Yes
90	negative			No
91	negative			Yes
92	negative			Yes
93	negative			No
94	negative			Unknown
95	Positive	29.23	Positive by 2nd PCR and by sequencing	Yes
96	negative			No
97	negative			No
98	negative			Yes
99	negative			Unknown
100	negative			Unknown
101	negative			Yes
102	negative			Yes
103	negative			Yes
104	negative			Yes
105	negative			Yes
106	negative			Yes
107	negative			Yes
108	negative			Yes
109	negative			Yes
110	negative			Yes
111	negative			Yes
112	negative			No
113	negative			Yes
114	negative			Yes

115	negative			Yes
116	negative			Yes
117	negative			Yes
118	negative			Yes
119	negative			Yes
120	negative			Yes
121	negative			Unknown
122	negative			Yes
123	negative			No

Appendix C

Appendix to Chapter 4

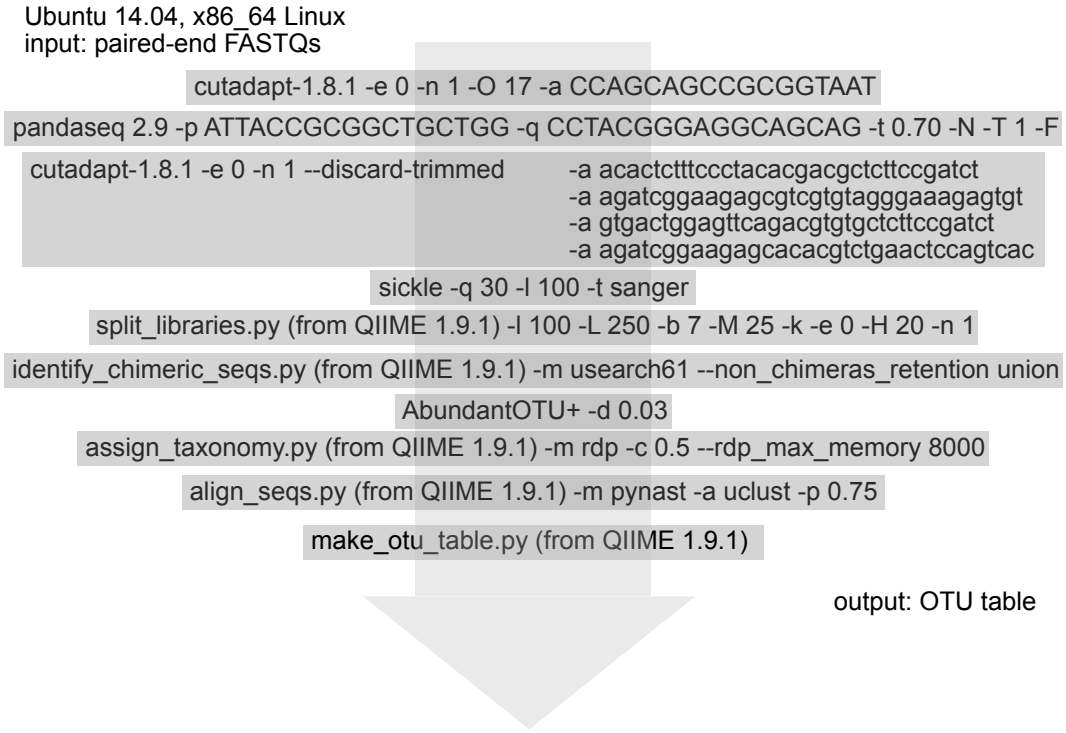


Figure C.1: **Flowchart of the 16S rRNA gene sequencing data processing approach.** Paired-end 16S rRNA gene sequencing data was processed using custom perl scripts which tied together existing processing software. These software, including their versions, options used, and order are presented here for the purposes of reproducibility.

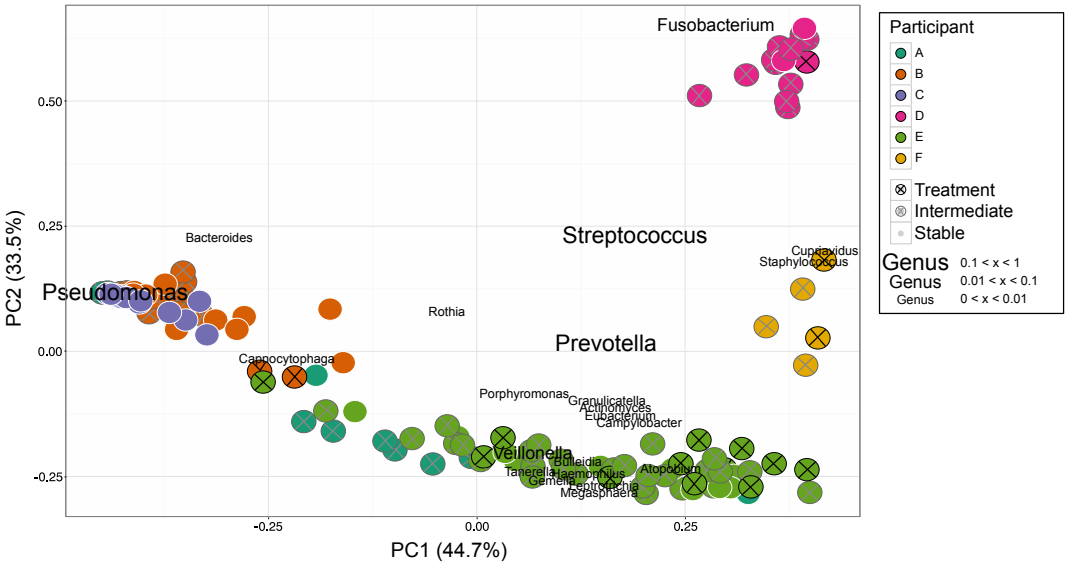


Figure C.2: **Genus-level biplot of the Participant-dependent CF lung microbiome.** A biplot of PC1 vs. PC2 of the PCoA plot displayed in Figure 2 reveals specific genera which contribute to Participant-specific separation. The mean relative abundance (0-1.0) across the dataset is displayed below genus labels.

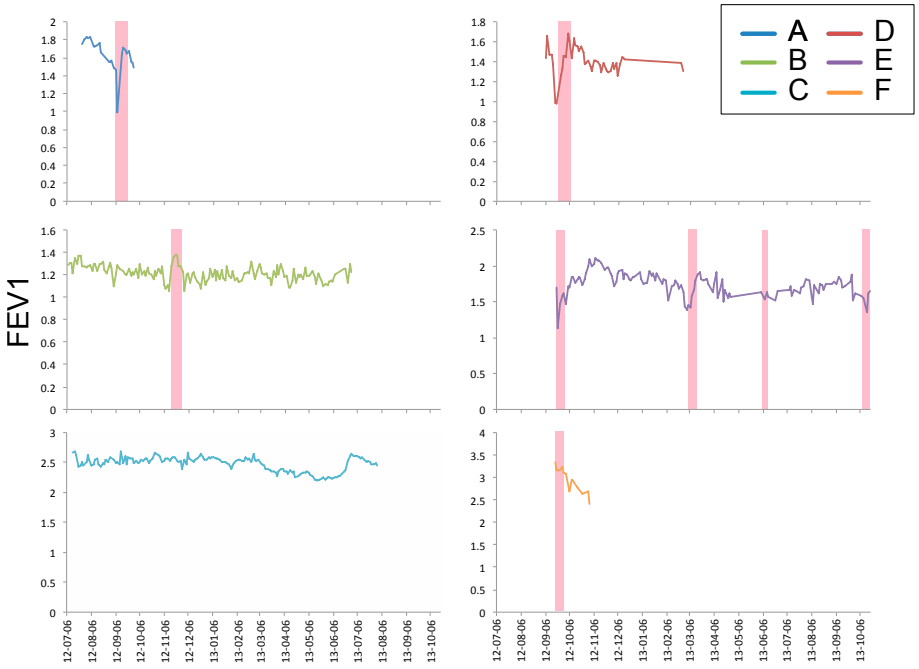


Figure C.3: **Longitudinal FEV1 values for each participant over the study period.** FEV1 data were collected 3x a week over the study period. Red vertical bars indicate Treatment time points.

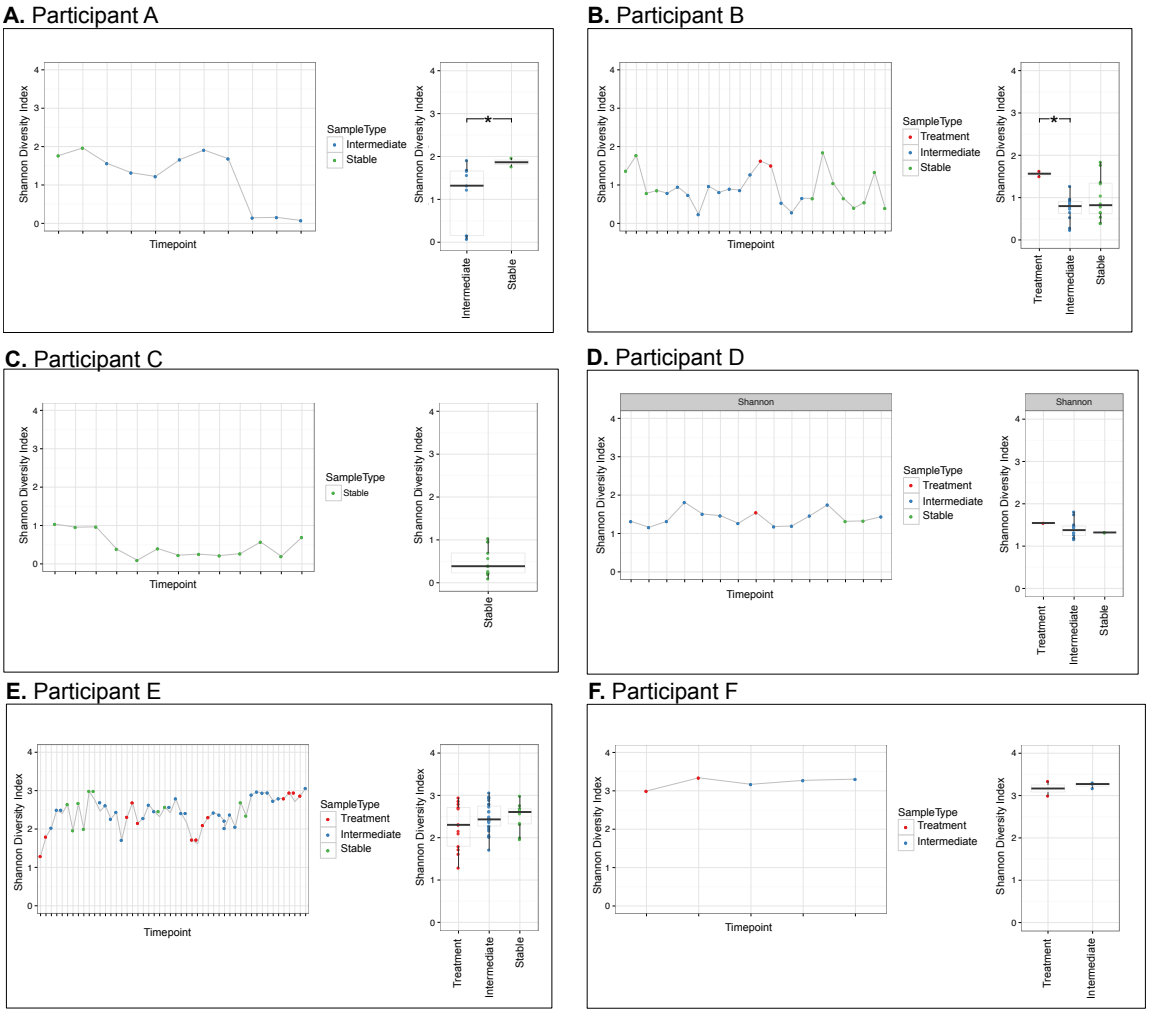


Figure C.4: **Alpha diversity measures of each participant over the study period.** Shannon diversity index was calculated for each microbiota sample collected over the study period. Statistical analyses between sample types indicated a significant difference between Intermediate and Stable samples from Participant A and Treatment and Intermediate time points in Participant B. All other comparisons were not statistically significant.

Table C.1: **Study duration and sample information.**

Participant	Study Duration: Dates	Length (in days)	# of Collected Samples	# of Sequenced Samples
A*	25/07/2012-28/09/2012	66	21	11
B	04/07/2012-28/06/2013	360	143	26
C	13/07/2012-31/07/2013	384	152	13
D*	05/07/2012-25/02/2013	236	26	15
E	19/09/2012-21/10/2013	398	154	51
F*	17/09/2012-31/10/2012	45	12	5

Table C.2: **p-values of statistical comparisons of Bray-Curtis dissimilarity scores between groups.**

Participant	**Stable vs. Intermediate**	**Stable vs. Treatment**	**Intermediate vs. Treatment**
A	0.137	-	-
B	0.662	0.17	0.045*
C	-	-	-
D	0.451	0.333	0.218
E	0.765	0.022*	0.009*
E1	0.8	0.5	0.4
E2	0.76	0.832	0.601
E3	0.643	0.067	0.002*
E4	0.745	0.4	0.499
F	-	-	0.9

Table C.3: **p-values of statistical comparisons of FEV1 changes between groups.**

Participant	**Stable vs. Intermediate**	**Stable vs. Treatment**	**Intermediate vs. Treatment**
A	0.036*	-	-
B	0.035*	0.67	0.102
C	-	-	-
D	0.806	-	-
E	0.089	0.009	0.067
E1	0.029*	0.232	0.304
E2	0.334	0.633	0.925
E3	-	-	0.436
E4	-	-	0.289
F	-	-	0.183

Table C.4: **Significantly correlating OTUs and select metadata for Participant C.** LS = local similarity score; PCC = pearson correlation coefficient.

X	Y	LS	PCC	Length	p-value	q-value
OTU1;g_Pseudomonas	OTU2;g_Prevotella	-0.887973	-0.950445	13	0.005497	0.043444
OTU1;g_Pseudomonas	OTU4;g_Streptococcus	-0.920999	-0.897723	13	0.003601	0.043444
OTU1;g_Pseudomonas	OTU8;g_Veillonella	-0.906857	-0.913167	13	0.004302	0.043444
OTU1;g_Pseudomonas	OTU9;g_Prevotella	-0.873737	-0.867662	13	0.006531	0.043444
OTU1;g_Pseudomonas	OTU11;g_Streptococcus	-0.935891	-0.907245	13	0.003007	0.043444
OTU1;g_Pseudomonas	Alpha	-0.99053	-0.981703	13	0.001428	0.043444
OTU2;g_Prevotella	OTU8;g_Veillonella	0.941744	0.941562	13	0.002695	0.043444
OTU2;g_Prevotella	OTU9;g_Prevotella	0.95917	0.937007	13	0.002161	0.043444
OTU2;g_Prevotella	OTU11;g_Streptococcus	0.873311	0.863102	13	0.006531	0.043444
OTU2;g_Prevotella	OTU12;g_Prevotella	0.951426	0.927755	13	0.002414	0.043444
OTU2;g_Prevotella	Alpha	0.909055	0.928795	13	0.004152	0.043444
OTU4;g_Streptococcus	OTU11;g_Streptococcus	0.945851	0.936433	13	0.002599	0.043444
OTU4;g_Streptococcus	Alpha	0.892802	0.840943	13	0.005128	0.043444
OTU5;g_Fusobacterium	OTU6;g_Prevotella	0.973557	0.978646	13	0.001792	0.043444
OTU5;g_Fusobacterium	OTU28;g_Prevotella	0.967562	0.995455	13	0.001932	0.043444
OTU5;g_Fusobacterium	Alpha	0.857732	0.578402	13	0.008006	0.047114
OTU6;g_Prevotella	OTU28;g_Prevotella	0.961842	0.984822	13	0.002082	0.043444
OTU8;g_Veillonella	OTU9;g_Prevotella	0.903566	0.932725	13	0.004456	0.043444
OTU8;g_Veillonella	OTU11;g_Streptococcus	0.872449	0.883309	13	0.006531	0.043444
OTU8;g_Veillonella	OTU12;g_Prevotella	0.886775	0.951942	13	0.005497	0.043444
OTU8;g_Veillonella	Alpha	0.929193	0.914985	13	0.003232	0.043444
OTU9;g_Prevotella	OTU11;g_Streptococcus	0.901649	0.896837	13	0.004616	0.043444
OTU9;g_Prevotella	OTU12;g_Prevotella	0.880315	0.933035	13	0.006098	0.043444
OTU9;g_Prevotella	Alpha	0.875753	0.841578	13	0.006311	0.043444
OTU11;g_Streptococcus	Alpha	0.918932	0.845682	13	0.003732	0.043444
OTU12;g_Prevotella	Alpha	0.872279	0.80161	13	0.006531	0.043444

Table C.5: **Significantly correlating OTUs and select metadata for Participant E.** LS = local similarity score; PCC = pearson correlation coefficient.

X	Y	LS	PCC	Length	p-value	q-value
OTU1;g_Pseudomonas	OTU13;g_Prevotella	-0.533465	-0.497165	50	0.000653	0.01071
OTU4;g_Streptococcus	OTU7;g_Streptococcus	0.48208	0.408235	50	0.002599	0.027483
OTU4;g_Streptococcus	OTU67;g_Rothia	0.691045	0.666126	50	0.000004	0.00035
OTU4;g_Streptococcus	OTU103;g_Streptococcus	0.799579	0.795728	50	0	0.000017
OTU5;g_Fusobacterium	OTU24;g_Prevotella	0.534316	0.621561	50	0.000627	0.01071
OTU7;g_Streptococcus	Alpha	0.575569	0.477824	50	0.000188	0.004889
OTU8;g_Veillonella	Alpha	0.521608	0.499919	50	0.000897	0.012957
OTU9;g_Prevotella	Alpha	0.446739	0.353607	50	0.006311	0.047559
OTU13;g_Prevotella	OTU25;g_Bulleidia	0.559686	0.560181	50	0.0003	0.006502
OTU19;g_Prevotella	OTU74;f_Streptococcaceae	-0.459854	-0.245866	50	0.004616	0.041437
OTU23;g_Prevotella	OTU33;g_Prevotella	0.561707	0.583158	50	0.000287	0.006502
OTU25;g_Bulleidia	OTU55;g_Peptostreptococcus	0.480291	0.228019	50	0.002695	0.027483
OTU27;g_Prevotella	Alpha	0.672771	0.532487	50	0.000008	0.00047
OTU74;f_Streptococcaceae	OTU101;g_Azorhizophilus	0.887692	0.900804	50	0	0.000001
OTU74;f_Streptococcaceae	OTU103;g_Streptococcus	0.485105	0.331119	50	0.002414	0.026155

Appendix D

Appendix to Chapter 5

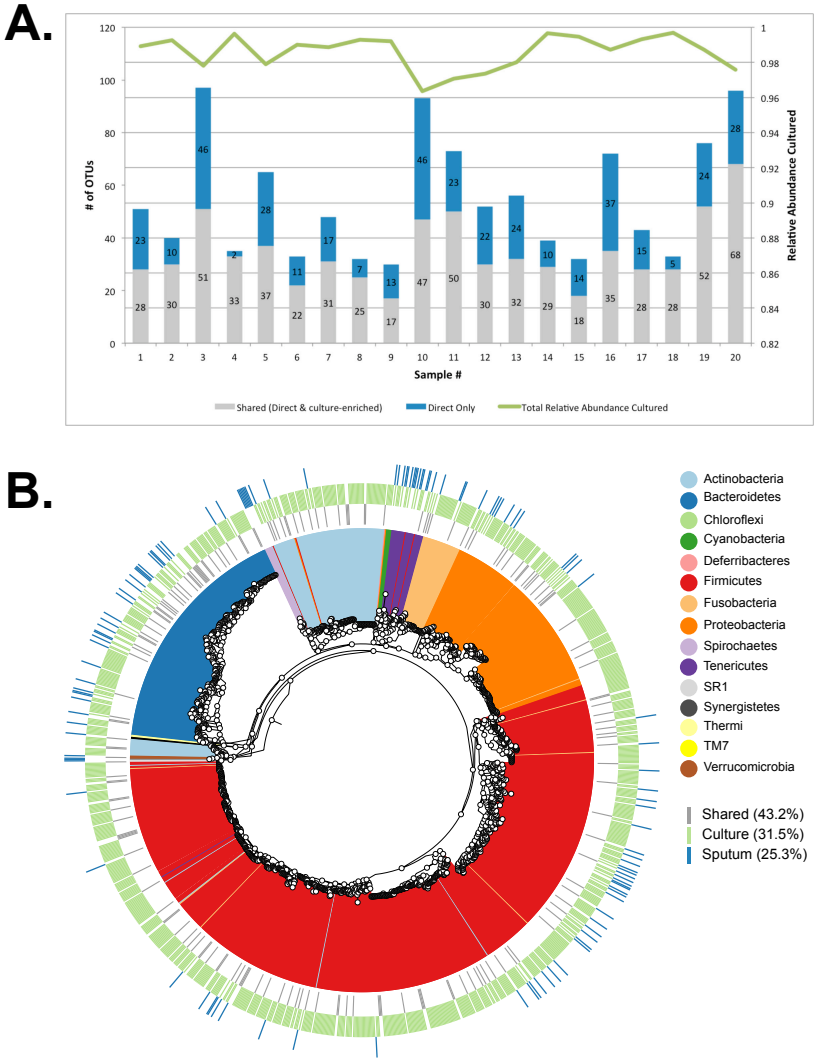


Figure D.1: **At more stringent abundance thresholds, the vast majority of the CF lung microbiota is still captured by culture-enriched methods. A.** Of the OTUs identified via sequencing of the sputum, 65.1% representing 98.5% relative abundance in the sputum samples are cultured. **B.** Using a cutoff of ≥0.1% relative abundance, 74.7% of the lung microbiota was cultured across samples. Further, only 25.3% of OTUs across the dataset were unculturable at this abundance cutoff.

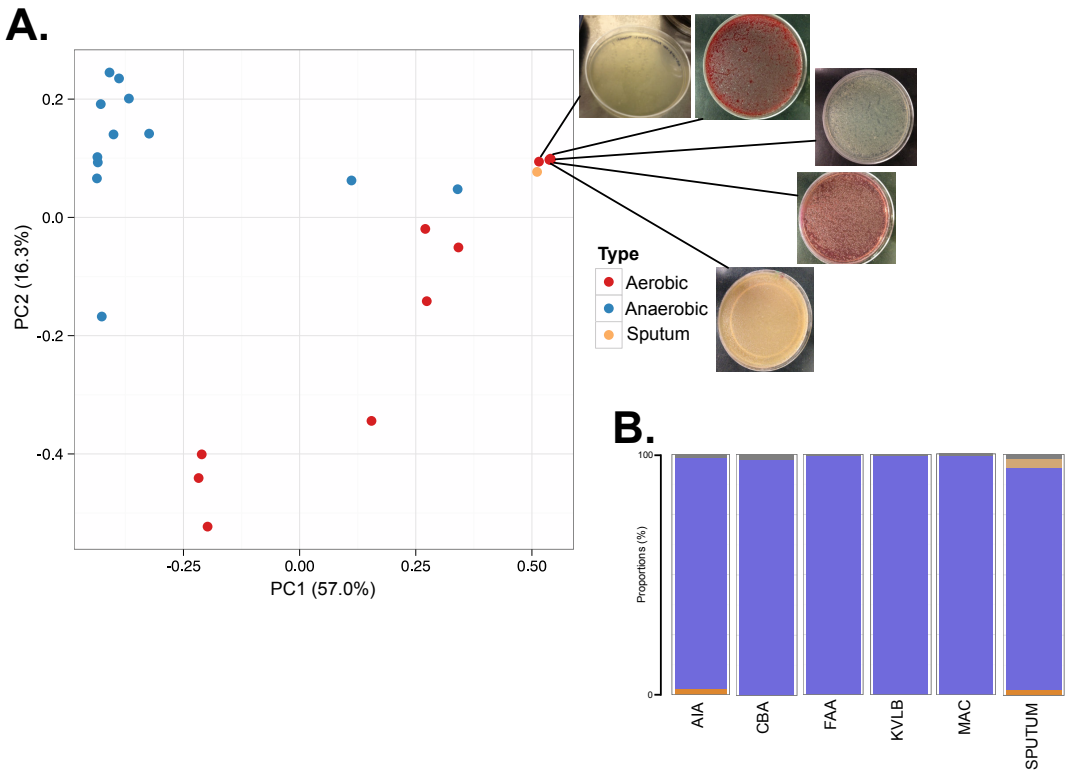


Figure D.2: **Clustering of plates used for culture-enrichment with the corresponding sputum samples is not due to non-viable DNA. A.** All plates which have similar communities to the original sputum sample as measured by the Bray-Curtis β-diversity metric have visible microbial growth as indicated by images captured immediately following culture-enrichment. **B.** Instead, these samples have similar taxonomic profiles due to the growth of *Pseudomonas* species as visualized via taxonomic summaries.

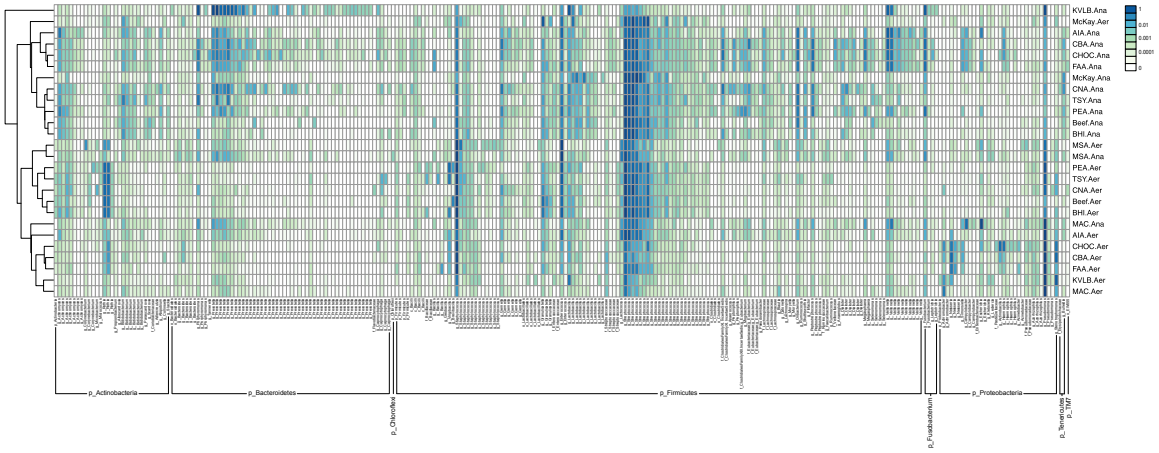


Figure D.3: **The variety in selective and non-selective media types, and aerobic and anaerobic environments is important in capturing the OTU-level diversity of the CF microbiota.** A heatmap indicates the breadth of media and conditions necessary to culture such a vast community at the OTU level. Heatmap indicates the maximum relative abundance, between 0-1, of each OTU on each environment/media pairing.

pre:
sputumAbund *= abundance cutoff in the originating sample*
abund *= abundance cutoff (corresponding to that in Figure 7)*
plateList *= ()*
allOTUs *= all cultured OTUs*

abundOTUs = a subset of allOTUs where abundance > **sputumAbund**
for each OTU **a** in abundOTUs
  count the # of plates with OTU **a** at abundance > **abund**
  if **count** == 1
    plate **p** = plate w/ OTU **a** on it
    plateList += plate **p**
    allOTUs -= any OTU on plate **p**
while abundOTUs is not empty
  find plate **q** with the greatest # of OTUs remaining in all OTUs
  plateList += **q**
  abundOTUs -= any OTU on plate **q**

post:
plateList *= list of plates for metagen seqing*
abundOTUs *= ()*

Figure D.4: **Pseudocode from a modified version of the plate coverage algorithm, the adjusted PLCA, which takes into account the abundance of the culture-independent sample as well as the OTUs recovered by culture-enrichment.**

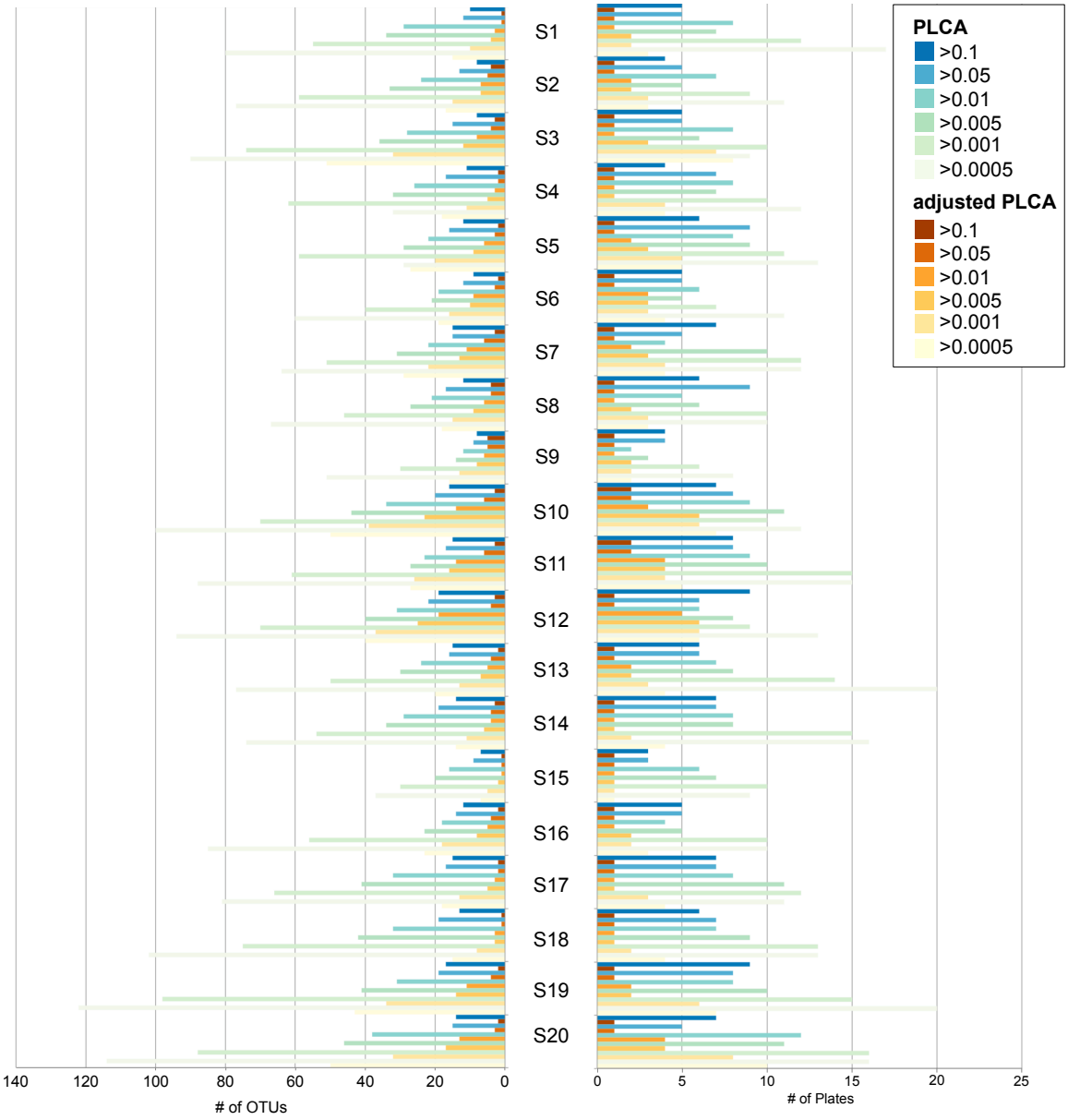


Figure D.5: **The minimal plate sets needed to sequence all sputum samples within this dataset, and the number of OTUs which would be obtained.**

Table D.1: **Full-length 16S rRNA gene sequencing results for colonies isolated from *Stenotrophomonas* re-growth.** Of the 10 isolates picked from the *Stenotrophomonas* isolation procedure, 8 were identified as *Stenotrophomonas maltophilia* strains by comparison to the Human Oral Microbiome Database (HOMD) and to the 16S ribosomal RNA sequences (Bacteria and Archaea) NCBI BLAST Database.

ID	Sequence Name	Query length (nt)	HOMD Top Hit	% Identity	BLAST Top Hit	Max Score	Query cover	E Value	Identity
SS1	Seq_Plate_56_H01_8f_25JUL14AB_F_seq_Un	855	S. maltophilia—HOT_663—Strain_LMG958—X95923—Named	98.7	NR_041577.1—S. maltophilia strain IAM 12423 16S r RNA gene	1487	97%	0	99%
SS2	Seq_Plate_56_H02_8f_25JUL14AB_F_seq_Un	867	S. maltophilia—HOT_663—Strain_LMG958—X95923—Named	98.5	NR_041577.1—S. maltophilia strain IAM 12423 16S r RNA gene	1483	97%	0	99%
SS3	Seq_Plate_56_H03_8f_25JUL14AB_F_seq_Un	919	S. maltophilia—HOT_663—Strain_LMG958—X95923—Named	98.5	NR_041577.1—S. maltophilia strain IAM 12423 16S r RNA gene	1485	89%	0	99%
SS4	Seq_Plate_56_H04_8f_25JUL14AB_F_seq_Un	924	S. maltophilia—HOT_663—Strain_LMG958—X95923—Named	98.7	NR_041577.1—S. maltophilia strain IAM 12423 16S r RNA gene	1496	90%	0	99%
SS5	Seq_Plate_56_H05_8f_25JUL14AB_F_seq_Un	926	S. maltophilia—HOT_663—Strain_LMG958—X95923—Named	98.6	NR_041577.1—S. maltophilia strain IAM 12423 16S r RNA gene	1478	89%	0	99%
SS6	Seq_Plate_56_H06_8f_25JUL14AB_F_seq_Un	919	Pseudomonas aeruginosa — H O T_536—Strain_L MG1242—Z7665 1—Named	99.2	NR_074828.1—Pseudomonas aeruginosa PAO1 strain PAO1 16S rRNA	1506	90%	0	99%
SS7	Seq_Plate_56_H07_8f_25JUL14AB_F_seq_Un	910	S. maltophilia—HOT_663—Strain_LMG958—X95923—Named	98.3	NR_041577.1—S. maltophilia strain IAM 12423 16S r RNA gene	1500	93%	0	98%
SS8	Seq_Plate_56_H08_8f_25JUL14AB_F_seq_Un	928	S. maltophilia—HOT_663—Strain_LMG958—X95923—Named	98.5	NR_041577.1—S. maltophilia strain IAM 12423 16S r RNA gene	1513	91%	0	99%
SS9	Seq_Plate_56_H09_8f_25JUL14AB_F_seq_Un	868	Pseudomonas aeruginosa — H O T_536—Strain_L MG1242—Z7665 1—Named	98.8	NR_074828.1—Pseudomonas aeruginosa PAO1 strain PAO1 16S rRNA	1347	85%	0	99%
SS10	Seq_Plate_56_H10_8f_25JUL14AB_F_seq_Un	914	S. maltophilia—HOT_663—Strain_LMG958—X95923—Named	98.1	NR_041577.1—S. maltophilia strain IAM 12423 16S r RNA gene	1504	93%	0	98%

Table D.2: **Culture-enrichment greatly mitigates host contamination.**

Sample	**Number of raw reads**	**Number of processed reads**	Read Loss (%)
Sputum	29,229,781	2,207,904	92.45
PLCA, AIA-Ana	5,708,137	238,081	95.8
PLCA, Beef-Aer	23,669,048	21,717,755	8.24
PLCA, KVLB-Ana	18,866,927	16,636,966	11.82
PLCA, MAC-Ana	21,628,034	18,290,293	15.43
PLCA, McKay-Ana	415,513	332,612	19.96
adjPLCA, CHOC-Ana	21,878,343	21,192,362	3.14
adjPLCA, McKay-Aer	12,561,265	7,943,779	36.8
adjPLCA, TSY-Aer	12,783,676	11,509,050	10.0

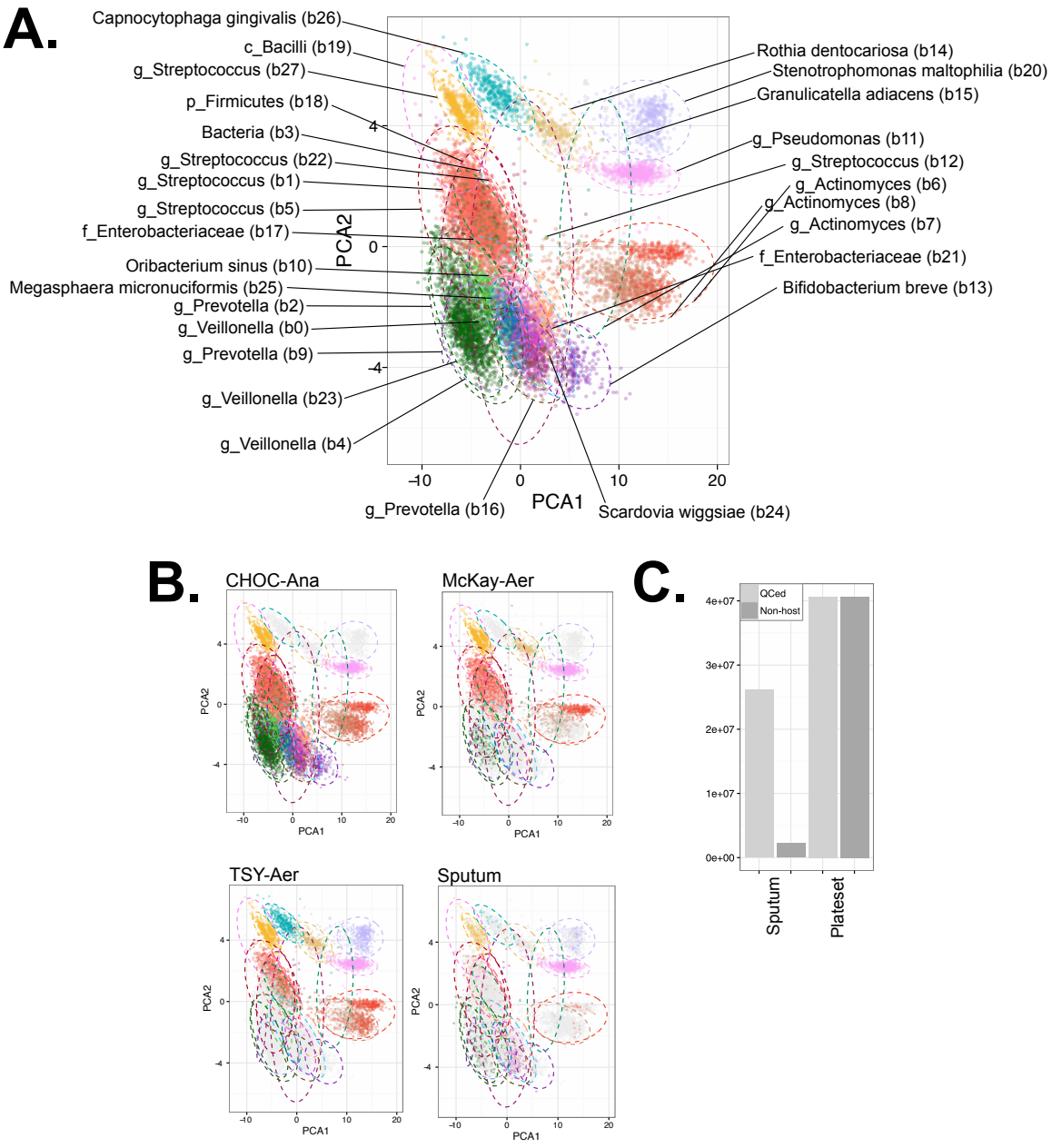


Figure D.6: **Binning of culture-enriched metagenomic contigs reveals the diversity of this approach when compared to sputum metagenomics. A.** Using CONCOCT, culture-enriched metagenomic reads from 3 plates amplified according to the adjusted PLCA were grouped into 28 bins. These bins are displayed here as a PCoA in which each dot represents a contig coloured according to its bin. A greater diversity of organisms were obtained via culture-enriched approaches (**B**); culture-enrichment contributes to the biological binning of such organisms. Further, a greater number of bacterial reads were obtained from culture-enriched sequencing (**C**) when compared to direct sputum sequencing.