THE AIRWAY MICROBIOME AFTER BURN AND INHALATION INJURY

Dana M. Walsh

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum in Toxicology in the School of Medicine.

Chapel Hill
2016

Approved by:

Scott H. Randell

Marianne S. Muhlebach

Peggy A. Cotter

Ilona Jaspers

David Diaz-Sanchez

ABSTRACT

Dana M. Walsh: The Airway Microbiome After Burn and Inhalation Injury
(Under the direction of David Diaz-Sanchez and Ilona Jaspers)

The human microbiome is composed of the entirety of microorganisms living on and in the human body along with their genetic material. Recent work has demonstrated the importance of these bacterial communities, or microbiota, in health and disease, including in the airways. Though the airways contain mechanisms to clear bacteria, disruption of homeostasis by illness and injury can induce conditions favorable to bacterial colonization and growth. Inhalation injury endured by burn victims disrupts homeostasis by damaging the airway epithelium and inhibiting innate immune responses, increasing the risk of acute respiratory distress syndrome (ARDS), infection, and pneumonia. Inhalation injury is a known cause of ARDS, which is partly diagnosed by hypoxia in the airways as indicated by a PaO_2/FiO_2 ratio ≤ 300. There is a known link between ARDS and bacterial infection in the airways, but the relationship is complex and poorly understood. Diagnosis of airway bacterial infection in this patient population can be challenging due to limitations in detecting and identifying the colonizing organism. The goal of this dissertation research was to identify differences in the airway microbiota among patients with PaO_2/FiO_2 ratios ≤ 300 and > 300 after experiencing burn and inhalation injury. Bacterial DNA was extracted from therapeutic bronchial washings of patients hospitalized for burn and inhalation injury at the North Carolina Jaycee Burn Center and sequenced. Patients with PaO_2/FiO_2 ratio ≤ 300 demonstrated increases in

low-abundance bacteria as well as significant enrichment of *Prevotella melaninogenica* that was not altered by antibiotic treatment. Bacterial taxa among patients with PaO_2/FiO_2 ratio ≤ 300 were grouped into correlation networks that were distinct both in composition and predicted function from patients with PaO_2/FiO_2 ratio > 300. Further, predicted functions important in characterizing the communities were unique for each disease state, identifying changes in bacterial interactions and functional roles that may be important in progression of hypoxia and ARDS. This combination of metagenomics with advanced computational analyses allows identification of specific changes relevant to the entire community, providing focused hypotheses for further validation and investigation that may lead to new therapeutic targets for preventing bacterial infection after burn and inhalation injury.

ACKNOWLEDGEMENTS

I owe many thanks to many people for completion of this dissertation. First, to the University of North Carolina at Chapel Hill for accepting me into an excellent biomedical graduate program. To my mentors, Drs. David Diaz-Sanchez and Ilona Jaspers, for their continuous support and guidance. To postdoctoral researchers who have helped me with the details of this project, including Drs. Janelle Arthur, Jaime Mirowksy, Shaun McCullough, Juliette Kahle, Radhika Dingra, and Shannon Grabich. To my peers in the Curriculum in Toxicology who pushed me to do better, including Phillip Wages, Natalie Holman, Desinia Miller and Nicole Kurhanewicz. To my collaborators, without whom this project would not be possible, including Drs. Bruce Cairns and Samuel Jones at the North Carolina Jaycee Burn Center, Drs. Jeff Dangl and Corbin Jones at the High Throughput Sequencing Facility, and Scott Yourstone and Natalie Stanley in the Bioinformatics and Computational Biology program. To technicians who assisted in primary cell culture work, including Lisa Dailey and Joleen Soukup. Finally, to friends and family who provided moral support along the way, particularly my husband, who moved halfway across the country and put up with my weird and stressful graduate school schedule. This wouldn't have been possible without all of you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABX	Antibiotics
ALI	Acute lung injury
ALIF	Air-liquid interface
ANOVA	Analysis of variance
ARDS	Acute respiratory distress syndrome
ATP	Adenosine triphosphate
AUC	Area under the curve
BAL	Bronchoalveolar lavage
BALF	Bronchoalveolar lavage fluid
BMI	Body mass index
BV	Bacterial vaginosis
CARS	Compensatory anti-inflammatory response syndrome
cDNA	Complementary deoxyribonucleic acid
CF	Cystic fibrosis
COHb	Carboxyhaemoglobin
CONV	Conventionalized
COPD	Chronic obstructive pulmonary disease
DA	Discriminant analysis
DAPC	Discriminant analysis of principle components
DNA	Deoxyribonucleic acid
DTT	Dithiothreitol

dNTP	Deoxynucleotide
FBS	Fetal bovine serum
FiO_2	Fraction of inspired oxygen
FMT	Fecal microbiota transplantation
GF	Germ-free
GSTM1	Glutathione S-transferase Mu 1
HBEC	Human bronchial epithelial cells
HCl	Hydrochloric acid
HIV-1	Human immunodeficiency virus-1
HMGB-1	High mobility group box 1 protein
HMP	Human Microbiome Project
HO-1	Heme oxygenase-1
HSP-70	Heat shock protein 70
HuMiX	Human-microbial crosstalk
IBD	Irritable bowel disease
IgG	Immunoglobulin G
II	Inhalation injury
IL-1Ra	Interleukin-1 receptor antagonist
IL-8	Interleukin-8
IPF	Idiopathic pulmonary fibrosis
IRB	Institutional Review Board
OTU	Operational taxonomic unit
Keap1	Kelch ECH associated protein 1

KEGG	Kyoto Encyclopedia of Genes and Genomes
LEfSe	Linear discriminant analysis effect size
mL	Milliliter
mRNA	Messenger ribonucleic acid
MT	Molecule tag
MTFS	Molecule tagging frame shifting
NGS	Next-generation sequencing
NIH	National Institutes of Health
NMI	Normalized mutual information
Nrf2	Nuclear factor erythroid 2-related factor 2
NSPCA	Non-negative, sparse principle components analysis
PAH	Polycyclic aromatic hydrocarbon
PaO_2	Partial pressure of arterial oxygen
PCA	Principle components analysis
PC	Principle component
PCI	Phenol:chloroform:isoamyl alcohol
PCR	Polymerase chain reaction
PICRUSt	Phylogenetic investigation of communities by reconstruction of unobserved states
PRR	Pattern recognition receptor
QIIME	Quantitative Insights into Microbial Ecology
qPCR	Quantitative polymerase chain reaction
rDNA	Ribosomal deoxyribonucleic acid

RF	Random forest
ROS	Reactive oxygen species
SCFA	Short chain fatty acid
SFB	Segmented filamentous bacteria
SHIME	Simulator of the Human Intestinal Microbial Ecosystem
SIRS	Systemic inflammatory response syndrome
SparCC	Sparse Correlation for Compositional Data
SPIEC-EASI	Sparse Inverse Covariance Estimation for Ecological Association Inference
SVM	Support vector machine
TBSA	Total body surface area
TLR	Toll-like receptor
TNF-α	Tumor necrosis factor alpha
UniProt	Universal Protein Resource
VAMPS	Visualization of Microbial Population Structures
VEC	Vaginal epithelial cells
WGS	Whole genome sequencing
WSP	Wood smoke particles

CHAPTER 1: THE HUMAN MICROBIOME AND ANALYSIS STRATEGIES

1.1 Introduction to The Human Microbiome

Advances in sequencing technology have revolutionized the study of microorganisms living on and within the human body. Populations of bacteria can now be identified *en masse* without first knowing each individual's unique metabolic requirements for growth, enabling rapid detection and identification of multiple species as well as discovery of new ones [1–3]. These communities, along with their genetic material, are collectively known as the microbiome [4]. Bacteria within the human microbiome, or the microbiota, outnumber human cells at an estimated ratio of three to one in healthy states [5]. Despite the abundance and diversity of bacteria present at homeostasis, traditional biomedical microbiology focuses on the ability of individual bacteria to cause disease and the elucidation of treatment strategies targeting these individuals. The emerging microbiome field, which allows study of multiple bacteria simultaneously, has emphasized the role of bacteria as members of a community that interact with each other as well as the host in a symbiotic way [6]. Study of these relationships has been advanced by initiation of The Human Microbiome Project [7]. This multi-center effort has employed next-generation sequencing (NGS) methods to characterize the symbiotic bacteria present at various body locations in healthy people and their roles in normal physiology as well as disease [7]. As a whole, the healthy human microbiome displays much variation in bacterial diversity and abundance both

between individuals and within body locations, but metabolic functions remain relatively stable, indicating similarity in function among healthy people [8]. In the gut, which maintains the largest population of bacteria within the human body, the microbiota play key roles in education of the developing immune system, extraction of nutrients from dietary fiber, metabolism of xenobiotic compounds, and production of small molecules that influence distant organ systems [9–14]. A growing body of work on the microbiota in other body locations implies parallels as well as distinctions in their composition and function in comparison to the gut. Whereas increased diversity is generally associated with improved health outcomes in the gut, this is associated with poor outcomes in the vagina [15]. Conversely, in the nasal and oral microbiome, loss of diversity is seen in conjuncture with increased severity of diseases such as chronic rhinosinusitis and periodontal disease [16,17]. Regardless of body location, development of disease alters the microbiota, their functions, and their relationship with the host. Understanding this relationship requires advancement in experimental, computational, and statistical techniques, which together contain exciting potential for advances in preventative and individualized medicine.

1.1.1 The Human Microbiome Project

The initiation of the Human Microbiome Project (HMP) in 2008 was a massive undertaking, involving multiple institutions in an effort to understand the genetic and physiologic diversity contributed by symbiotic bacteria in healthy states and during development of disease [7]. Though the Human Genome Project elucidated the genes encoded in the human genome, it did not touch on the additional microbial genomes

present and their functions [18]. The HMP is effectively an extension of the Human Genome Project in its mission to identify bacterial genes and functions that contribute to human health outcomes. Unlike the Human Genome Project, the HMP faced the challenge of analyzing multiple genomes from multiple bacterial species, many of which cannot be cultured [19]. This required deep, whole-genome sequencing of selected individual bacteria to develop reference genomes as well as other sequencing methods, specifically 16S rRNA gene sequencing, that are capable of detecting and identifying

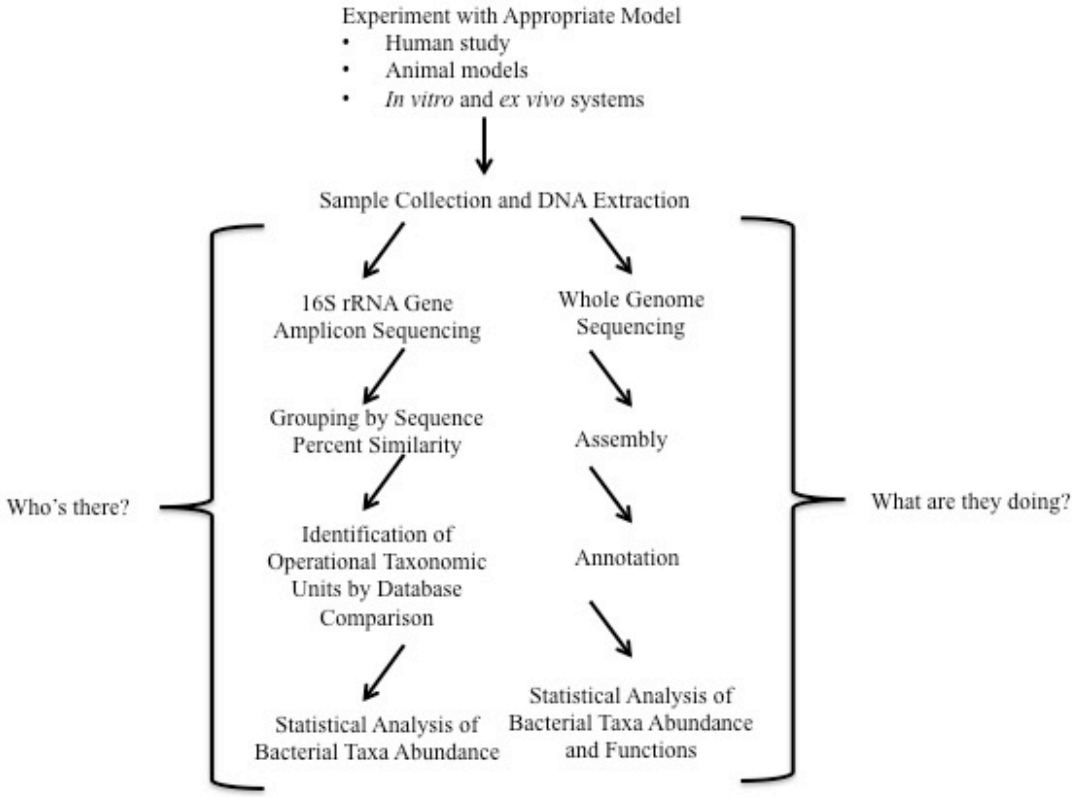

Figure 1.1: Microbiome Experimental and Sequencing Methods. Microbiome studies require selection of an appropriate model to address an experimental hypothesis, followed by sequencing methods to elucidate bacterial community composition (Who's there?) and/or function (What are they doing?).

multiple bacteria from short reads. Figure 1.1 provides an overview of experimental and sequencing techniques used by the HMP and other microbiome studies. Although alteration of the microbiome in disease states holds greater implications for human health

outcomes, the HMP recognized the need to characterize the healthy microbiome and its associated functions from a wide range of individuals in order to assess the impact of disease-induced change. The project has so far generated publicly available clinical specimens, reference genomes, sequencing and annotation protocols, and other collection, extraction, and analysis methods [20]. A paper published in 2012 details the standardized protocols established for collection, sequencing, and analysis of 5,298 samples from 15 or 18 body sites in 242 healthy individuals [20]. Bacterial 16S ribosomal gene amplicon sequencing was used on all samples and a subset was sequenced using whole-genome techniques [20]. Because suitable 16S rRNA gene sequencing methods did not exist at the start of the project, the HMP was instrumental in developing a standardized way to analyze 16S rRNA gene data for future studies. Further, the project has resulted in the development of analysis tools and statistical methods specific to microbiome data, which is key to fully understanding the implications of the results.

1.1.2 Healthy Microbiome Composition and Function

Since initiation of the HMP, microbiome research has increased exponentially. Studies are encompassing a growing number of body locations, disease states, and environmental exposures as researchers recognize the far-flung impacts of these bacterial communities. Results from the HMP provide a starting point in understanding bacterial community membership and function in healthy individuals from a broad range of backgrounds. The 242 volunteers who provided samples for this study ranged in age from 18 to 40 years old, were recruited in either Texas or Missouri, met study criteria for

healthy status, and submitted either 15 (for men) or 18 (for women) samples from body locations listed in Table 1.1.

Body Site	Specific Location
Oral cavity	Attached keratinized gingiva
	Buccal mucosa
	Hard palate
	Palatine tonsils
	Saliva
	Subgingival plaque
	Supragingival plaque
	Throat
	Tongue dorsum
Nasal cavity	Anterior nares
Skin	Left & right antecubital fossa
	Left & right retroauricular crease
Gastrointestinal tract	Stool
Urogenital tract	Mid vagina
	Posterior fornix
	Vaginal introitus

Table 1.1: Body Locations of Sampling for the Healthy Human Microbiome.

The HMP's use of both whole-genome and 16S rRNA gene amplicon sequencing provides species level abundance data for all samples and bacterial functions for a subset of the samples. Among the 242 healthy volunteers in the HMP, variation among the microbiota was greater between individuals than between body locations over time [8]. Microbiota from different locations of the body in the same person were more similar than those microbiota were to another person, highlighting the effect of inter-individual variation. A myriad of reasons may explain this, from differences in diet, to genetics, to environmental exposures and ethnic background, among others. Despite person-to-person differences in bacterial composition, levels of diversity among body sites were similar. The oral and stool microbiota were observed to have high alpha (within sample) diversity while the vagina had low diversity [8]. Comparing alpha and beta diversity metrics reveal ecological diversity and similarity among community membership. For example, the

saliva microbiota displayed high alpha diversity but low beta diversity, indicating bacterial communities rich in bacterial taxa that were similar among individuals [8].

One of the goals of the HMP was to determine whether a core microbiome exists among healthy individuals [7]. Bacterial taxa that are core to a community are consistently present among two or more individuals or habitats [21–23]. In the strictest definition, the core taxa must be present among all the communities in a group [21]. However, other studies have used membership among 95% of individuals to as few as 50% [23–25]. Identification of a core microbiome is complicated by the level of taxonomic specificity as well as sub-populations of people and environments [22]. It is easier to define core taxa at the phylum level than it is at the species level, since there is wide variation in species carriage among healthy individuals [8]. Various subject metadata could be employed to identify core taxa among groups, such as season in which the sample was taken among nasal microbiota and the development of peri-implant-associated disease in the oral microbiota [24,25]. Within the HMP dataset, no taxa were present among all subjects [8]. However, taxa present among 95% of subjects, rather than all, revealed core microbiota [23]. Samples from the oral cavity contained the greatest number of taxa present among 95% of subjects, followed by nose, stool, skin, and vagina [23]. This work revealed the difficulties in defining a core microbiota at an appropriate level of taxonomic granularity; while taxa in the mouth were defined as core at the genus level, refining the taxa to the species level resulted in distinct selection of species among individuals [23]. Regardless, identification of core microbiota which occur commonly among a group may imply their importance in the function of that particular community [21].

Despite differences in taxonomic composition among individuals, the HMP found relative stability among the functions of microbial communities inhabiting specific body sites in healthy individuals [8]. Common metabolic pathways, such as the components necessary for translation, ATP synthesis, and glycolysis, were consistently represented among healthy subjects [8]. This underscores the idea that microbial community function may be more important in understanding health outcomes than composition. Although common metabolic pathways could be correlated with subject metadata through multivariate statistical analysis, this does not explore the impact of less abundant functions, such as those associated with pathogenesis [8].

The HMP was an enormous undertaking and provided many tools previously lacking in the microbiome field, including standardized protocols, new bioinformatics methods, and an understanding of the healthy microbiome in various body sites. However, much work remains to be done to fully understand the breadth and depth of the healthy human microbiome. Some major limitations of the HMP data set are its restriction to healthy adults in the United States as well as a lack of detailed metadata from these subjects, such as diet, comorbidities, environmental and drug exposures, and human genome data. Recent studies have found changes in both microbiome composition and function that are age and population dependent, indicating that the HMP data set does not adequately cover the range of the healthy microbiome [26,27]. Diet has a clear and measurable impact on both the composition and function of the gut microbiome, making it an important factor to consider when characterizing healthy microbial communities [28]. Similarly, cigarette smoke, environmental chemicals, and xenobiotics have all been shown to alter microbiota composition and, in some cases, function [29–32]. Finally, host

genetics can shape the microbiome to influence microbial growth and metabolism and host phenotype [33,34]. For example, several taxa were identified as heritable among monozygotic twin pairs, with Christensenellaceae being the most heritable and demonstrating the ability to reduce obesity when introduced to germ-free mice in the company of obesity-associated taxa [33]. Further, specific microRNA produced by human intestinal epithelial cells have been shown to regulate bacterial gene transcripts and modulate bacteria growth [34]. Clearly, there is a complex interaction among the microbiota, host, and environment, which we have yet to completely understand. Though the HMP's work was pioneering, there remains a need to expand it to take into account a more diverse range of subjects and incorporate a rich set of metadata in order to more completely define the healthy human microbiome.

1.1.3 Role of the Microbiome in Immune System Development and Education

Recent studies have revealed the key roles microbiota play both in the development and function of the host immune system, and how dysbiosis, or perturbation of the communities, contributes to disease [16,35–38]. Millions of years of co-evolution has resulted in finely balanced cross-talk that allows the host to shape microbial communities and these communities, in turn, to influence host immune responses [9]. Colonization of mucosal and other surfaces by bacteria occurs immediately following birth, and new work indicates that exposure may even begin *in utero* [14,39–41]. Though the impact of *in utero* bacterial exposure on later health outcomes remains unknown, this demonstrates early interaction among the microbiome and the immune system during critical developmental periods. Disruption of these interactions, through diet, antibiotic

and drug treatments, and environmental exposures, may predispose infants to inflammatory, allergic, and other diseases later in life [14,42]. Germ-free (GF) mice, which are born in a sterile environment and deprived of microbial interactions, display stunted immune maturation and impaired responses to pathogenic bacteria, demonstrating the importance of commensal microbes in development of a functional immune system [39]. Large-scale studies with human children show reduced risk of allergies if they grow up in a farm environment, suggesting that early exposure in life to diverse microbes aids in educating the immune system to respond appropriately to foreign antigens [43]. This is incorporated in the hygiene hypothesis, which encompasses the idea that modern-day sterility reduces early-life exposure to microbes to a degree that results in poor development of immunity and increases the risk of allergic disease [44]. Inflammation in the absence of microbiota, which is ameliorated by introduction of innocuous bacterial products, supports this hypothesis from a microbiome perspective [42]. This work shows that early life microbial exposures are important to appropriate immune development and function later in life.

In an appropriately educated immune environment, microbiota play key roles in regulation of inflammation and prevention of infection. In the intestine, a loss in microbial diversity or perturbation of the community predisposes the gut to infection, highlighting the importance of commensal species in limiting pathogen growth and stimulating immune response [39]. Treatment of mice with antibiotics impairs adequate innate and adaptive immune responses to viral infection, reducing clearance of the virus [45]. In the healthy microbiome, segmented filamentous bacteria (SFB) drive differentiation of $CD4^+$ Th17 differentiation and protect against infection by *Citrobacter*

rodentium [39]. In the nasal microbiome, the commensal bacteria *Lactobacillus murinus* is less effective in stimulating $CD4^+$ T cells than the pathogen *Streptococcus pyogenes* and does not induce disease [46]. This demonstrates the immune system's ability to differentiate between commensal and pathogenic organisms and elicit appropriate responses. The immune system can also differentiate between the commensal and pathogenic potential of the same organism depending on its location in the body. If staphylococcal species within the skin microbiome are present below the dermis, an inflammatory response is invoked [10]. If they are instead on the epidermal surface, no inflammation is produced due to inhibition of Toll-like receptor 3 (TLR3) signaling of keratinocytes by staphylococcal lipoteichoic acid [10]. These studies imply that interactions between microbiota and the immune system are finely tuned to maintain homeostasis and their disruption can lead to inflammation and infection. Ongoing work continues to explore these interactions and how they might be manipulated to prevent and/or treat disease.

1.1.4 The Diseased and Injured Microbiome

Though the healthy microbiome and its interactions with the immune system remain to be entirely elucidated, study of the diseased and injured microbiome has revealed previously unknown roles for specific taxa that may hold keys for therapeutic manipulation of bacterial communities. A large body of work exists exploring changes in the gut microbiome under conditions such as irritable bowel disease (IBD), colon cancer, diabetes, allergy, and obesity. In all of these diseases, shifts in microbial community composition have been observed concomitantly with a loss of overall bacterial diversity.

At the coarsest level, more than 90% of phylotypes (sequences with >97% similarity; also known as operational taxonomic units or OTUs) in the healthy gut belong to one of two phyla; either the Bacteroidetes or the Firmicutes [47]. The phyla Proteobacteria, Actinobacteria, Fusobacteria, and Verrucomicrobia have also been consistently detected in the gut at lower abundance [23,47]. In the presence of IBD, the abundance of Actinobacteria and Proteobacteria increase, Bacteroidetes decrease, and these changes are accompanied by a loss of bacterial diversity [47]. In colorectal cancer, Fusobacteria and Proteobacteria increase while overall diversity decreases [48]. Patients with type two diabetes display a loss of bacteria in the Firmicutes phylum with an increase in Betaproteobacteria [49]. Fewer studies on the role of the gut microbiome in allergies exist, but a study of infants showed increases in anaerobes and lactobacilli and decreases in bifidobacteria and enterobacteria [42]. Finally, obesity is known to induce shifts in the gut microbiome, with specific increases in Actinobacteria and Bacteroidetes accompanied by an overall loss of diversity as compared to lean individuals [50]. Though these disparate diseases show alteration in different bacterial phyla, they all display a loss in diversity among the gut microbial communities. This has been a consistent finding among other gut microbiome studies as well, leading to the general acceptance of the idea that a loss of microbial diversity is associated with poor health outcomes. Supplementation with additional bacteria, such as in probiotics or fecal microbiome transplants, is a possible therapeutic intervention to ameliorate the effects of these gut diseases.

The association between loss of diversity and poor health outcomes holds true for other body locations, such as the oral and nasal microbiome. In the oral microbiome,

outgrowth of a single, low-abundance bacterium, *Poryphyromonas gingivalis*, is known to induce periodontitis [17]. Periodontitis is a polymicrobial disease characterized by inflammation and bone loss [17]. *P. gingivalis* was shown to induce periodontitis through the complement pathway only in the presence of other commensal microbiota, despite it being a low-abundance organism [17]. Outgrowth of *P. gingivalis* and development of periodontitis was accompanied by a loss of oral microbiome diversity [17]. In a similar manner, *Corynebacterium tuberculostearicum* mediated severity of chronic rhinosinusitis in the nasal microbiome [16]. Here, *C. tuberculostearicum* was enriched in the presence of decreased nasal microbiome diversity but could be inhibited by *Lactobacillus sakeii* [16]. These studies not only confirm the association between loss of diversity and poor outcomes, but also demonstrate the ability of single, low-abundance species to mediate disease. Interventions aimed at controlling the growth of these specific bacteria, possibly through other microbiota known to inhibit their growth, could prove effective in managing and preventing disease.

The vaginal microbiome is one of the few known body locations in which low diversity is associated with better health outcomes. In healthy, non-pregnant women, the vaginal microbiome has low alpha diversity and is dominated by one or two of the following *Lactobacillus* species; *L. crispatus, L. gasseri, L. iners,* or *L. jensenii* [41]. Its composition fluctuates with changes in age, hormones, infection, and sexual behavior [41]. During bacterial vaginosis (BV), vaginal microbiome diversity increases, which indicates poor health outcomes for this body site [51]. In contrast, the vaginal microbiome during pregnancy is more stable but less diverse [15,41]. Preterm birth and the development of chorioamnionitis are associated with changes in both the vaginal and

placental microbiome. In preterm births, the vaginal microbiome displays a decrease in the abundance of lactobacilli with a dominance of the *Prevotella* and *Peptoniphilus* genera [52]. Healthy term births are associated with increased *Enterobacter* and *L. crispatus* in the placenta, but preterm births with severe chorioamnionitis display enrichment of *Ureaplasma parvum, Fusobacterium nucleatum*, and *Streptococcus agalactiae* in the placenta [40]. Though the role of these bacteria in BV and preterm birth is unclear, intervention strategies which supplement the vagina with *Lactobacillus* depending on vaginal pH level have shown decreases in preterm births [41]. Larger clinical trials are necessary to confirm the positive impact of *Lactobacillus* supplementation in pregnant women.

Injury to the microbiome, such as by surgery, smoke inhalation, or toxic exposures, can also induce dysbiosis. Victims of burn injury often experience disruption of the intestinal epithelium, which increases the risk of bacterial translocation and sepsis [53]. The gut microbiota in these patients shows increased dysbiosis accompanied by specific enrichment of aerobic Gram-negative bacteria [53]. However, alterations in the gut microbiome after injury do not always lead to infection and disease. In germ-free mice lacking a healthy microbiome, cellular regeneration after colonic injury is slowed, indicating that microbiota play important roles in epithelial repair [36]. They may also protect the cells; cytoprotective genes are upregulated by microbiota through detection by pattern recognition receptors (PRR) and subsequent generation of reactive oxygen species (ROS) through the Nrf2/ARE pathway [36]. Environmental chemicals have been shown to alter microbiota composition and function as well. Exposure to chemicals in personal care products can alter microbial community composition, even at low doses [32]. Rats

were exposed to low and high doses of diethyl phthalate, methylparaben, triclosan, or a mix of these three from birth through adulthood [32]. Significant increases in Bacteroidetes and decreases in Firmicutes were seen at the adolescent stage in these rats, but these changes disappeared by adulthood [32]. Arsenic, a common drinking water contaminant, also shifts microbial composition and metabolic phenotype [54]. Arsenic demonstrated time and dose-dependent changes in Bacteroidetes and Firmicutes along with microbial and host nitrogen metabolism, possibly inducing conditions favorable to infection and disease [54]. Cigarette smoke, which contains known cytotoxic compounds such as acrolein and polycyclic aromatic hydrocarbons (PAH), alters the nasal and oral microbiome [55–59]. Comparison of the upper and lower respiratory microbiota in smokers and non-smokers revealed overall enrichment of certain taxa in the lung but the effect of smoking was only significant in the oral microbiome [58]. Another study comparing the effect of smoking on the nasal and oral microbiota revealed increased microbial diversity in smokers and clustering first by body site and then by smoking status, indicating the greater effect of location over smoke exposure [59]. These studies demonstrate that injury, whether physical or chemical, can induce changes in the microbiota in a variety of body locations, but their immediate and long-term effects on health are unclear and require additional investigation.

1.2 Review of Strategies for Investigating the Microbiome

The HMP established a precedent for future microbiome studies. Development of standardized protocols provided guidelines for other researchers undertaking human microbiome studies, enabling comparison among work from various groups. Mouse and

in vitro cellular models have been developed, allowing for mechanistic investigation into observations from human studies, as well as various bioinformatics pipelines and statistical methods specific to the unique requirements of microbiome data. Microbiome studies require the unification of biological experiments with computational data analysis, encouraging the formation of multi-disciplinary groups to carry out the most effective research. Effective communication between experts in these fields is crucial to optimization of research strategies.

1.2.1 Experimental Models

The first step in any microbiome project is selection of an appropriate model system in which to carry out the study. For research in the human microbiome and its role in health and disease, human subjects research is the ideal model. However, certain manipulations are impossible to perform in people and must instead be carried out in animal or cell culture models. Animal models, such as mice, rats, and ferrets, recapitulate the whole organism, allowing for interactions among microbiota and cell subtypes. However, animals are genetically and metabolically different from humans, making them unsuitable for certain types of studies. Cell culture models allow study of the interaction of cells and bacteria in isolation, revealing detailed mechanisms that may shed light on therapeutic targets. However, this isolation means crucial cellular interactions may be missed and promising therapeutic targets *in vitro* may not work at all *in vivo*. A combination of these model systems is necessary for a complete understanding of the microbiome, its interactions with the host, and how it may be manipulated to improve health outcomes.

1.2.1.1 Human Studies

Of all the model systems, human subjects research may be the most critical to accurately understanding microbe-host interactions and their roles in health and disease. Study of microbiota in humans is the most clinically relevant system but may be limited by the ability to recruit appropriate subjects, compliance with National Institutes of Health (NIH) guidelines, and lack of fine experimental control [60]. Protections for human subjects in biomedical research in the United States are heavily based upon the 1979 Belmont Report and were subsequently expanded in the Protection of Human Subjects Law and the Common Rule [61]. The Belmont Report is composed of three principles: (1) Respect for persons, (2) Beneficence, and (3) Justice. Institutional Review Boards (IRB) were created to uphold these principles in resect to research with human studies. An IRB is composed of both scientists and members of the community whose job is to review research protocols and ensure they comply with human research protection laws [61]. Therefore, it is crucial to design human microbiome studies with these principles in mind. In compliance with the principle of beneficence, a study must do no harm to the subject and provide a benefit either to the individuals involved in the study or to society as a whole [61]. Within respect for persons, researchers must provide informed consent documents to study volunteers which explain, in clear terms, what they are agreeing to do and why [61]. Finally, to comply with the principle of justice, researchers must not discriminate in selection of study participants and must make results of the study publicly available [61]. Compliance with these rules makes certain types of studies impossible in human volunteers, necessitating the use of other model systems.

1.2.1.2 *In Vivo* Animal Models

Animal models allow study of manipulations to both the host and the microbiome that may not be ethically possible in human studies. Animals can be treated with pathogens and toxic compounds that are known to do harm to people, and they can be euthanized and dissected to study the impact of such compounds on specific body locations [62]. Although such methods provide crucial knowledge that ultimately benefits human health, their use in people would clearly be unethical and violate the principle of beneficence laid out by the Belmont Report. Therefore, animal models provide a critical link between observations in human studies and specific manipulations that aid in understanding the mechanism behind them.

Of existing animal models, mice are used most frequently due to ease of working with and maintaining them, genetic tools available, and their genetic similarities to humans [60]. Mice share ninety-nine percent of their genes with humans, and various knock-out, knock-in, and transgenic models have been developed to study the function of these genes and their impact on human biology. Environmental conditions in mouse models, such as diet, exercise, wake/sleep cycles, and stress are easy to alter and assess the resulting impact on microbiota and health outcomes. Mice have played key roles in the study of transgenerational inheritance, which is the idea that epigenetic alteration to previous generations, through environmental exposures and/or disease, can be passed on to the offspring [63]. Though not widely explored in the context of the microbiome, recent mouse studies demonstrate the role epigenetics may play in appropriate colonization of the microbiome after birth as well as the influence these microbes have on

host epigenetic mechanisms [64,65]. Mice provide a convenient model system in which to explore the interaction of host genetics, epigenetics, and the microbiome.

The most widely used mouse model in microbiome studies is the germ-free mouse (GF). These mice are born in sterile isolators that keep them free from colonization with any microorganism; bacteria, virus, or fungi [66]. These animals can then be colonized with individual bacteria, selected communities, or even donor communities from humans in order to link them to specific functions [66]. Once colonized, these animals are known as gnotobiotic mice [60,66]. The utility of this model is exemplified in a recent study on the composition and function of the gut microbiota in twins [67]. Here, fecal microbiota from one monozygotic twin pair and three dizygotic twin pairs, all of which were discordant for obesity, were transplanted into GF mice. The mice were fed a low-fat, high-plant polysaccharide diet and assessed for adiposity and metabolic changes. Because mice are coprophagic, cohousing one mouse with the lean microbiota and another with the obese microbiota allowed study of transmission of the microbiota between them. The model revealed increases in adiposity in mice colonized with the obese microbiota, along with decreased metabolism of short-chain fatty acids (SCFA), and increased metabolism of branched-chain amino acids. Cohousing prevented adiposity in obese mice and shifted their microbiota to a lean-like state, which included increased abundance of Bacteroidales [67]. Although compositional and functional differences can be measured from human microbiota, the use of GF mice in this study provided evidence that the bacteria themselves, rather than diet, contributed to increased body mass and adiposity. Further, the ability to cohouse these mice demonstrated transferability of the phenotype and identified bacterial taxa that may be important in

preventing obesity. This level of fine experimental control is not possible in human studies and provides direct evidence for functional roles of the microbiota in health and disease.

Though GF mice play crucial roles in elucidation of microbial functions, they are not without their limitations. In comparison to conventionalized (CONV) mice, GF animals demonstrate stunted immune responses, altered gene expression profiles in epithelial intestinal cells, and reduced renewal of epithelial cells after injury [39,66]. Several recent studies imply that microbial colonization after birth is critical to development of appropriate immune responses as well as brain development [14,68–70]. In the prefrontal cortex of GF mice, genes involved in myelination and myelin plasticity were upregulated and axons showed hypermyelination, suggesting that microbiota are necessary for appropriate regulation of this process [68]. Though the GF model allows elucidation of functional roles of microbiota, particulary during development, it may not be appropriate for studies investigating differences in microbial composition at the adult stage. Since these animals have not had exposure to microbes from birth, their responses to various microbiota may be skewed and not accurately represent differences induced by microbial communities. For example, though the previously mentioned twin study used GF mice to identify differences in function and composition between lean and obese microbiota, the response may be entirely different if CONV mice had been used instead. Despite these issues, the GF mouse remains an important model for establishing the importance of specific bacteria and communities of bacteria on host development.

Due to limitations of the GF model as well as limitations in access to it, depletion of microbiota in mice has also been done through antibiotic treatment. This allows the

mouse to be exposed to microbiota during development, eliminating the deficiencies seen in the GF model [71]. For studies examining the impact of microbiota on disease later in life, this may be a more appropriate model. Chronic treatment of mice from weaning onward with antibiotics (ABX) resulted in altered gut microbiota, reduced anxiety, cognitive deficits, and significantly reduced expression of neuromodulators in the adult brain [69]. The similarity of these results to those in GF mice implies that ABX treatment may be a valid alternative to GF mice. As with GF mice, microbiota can be transplanted into ABX-treated mice, but differences in phenotype may not be as strong and the donor community composition may not be maintained for as long [71,72]. Further, antibiotics may selectively inhibit specific bacterial species while allowing outgrowth of others rather than eliminating the entire community, resulting in variability among ABX-treated mice in baseline remaining microbiota before introduction of the microbiota of interest [72]. Depletion of intestinal microbiota in mice by oral gavage with ABX has been shown to significantly decrease bacterial load and result in a GF-like phenotype but a lack of 16S rRNA gene sequencing in this study does not address the issue of variability in post-treatment microbiota [73]. The solution may be post-ABX and pre-transplant sequencing of ABX-treated mice in order to statistically address the impact of the pre-existing microbiota. The transplant could also be given to GF parents and the offspring used in subsequent studies, but this introduces additional uncertainty in efficient transfer of the transplant to the offspring [71]. ABX are also known to directly impact host physiology, may contribute to an increase in transfer of ABX-resistant genes, and do not eliminate viruses or fungi, both of which play important roles in interactions with the microbiome and the host [71]. Despite these limitations, ABX-treated mice provide an

easily maintained and accessible model for investigating the role of the microbiome in various life stages.

Additional considerations in using mouse models are strain and vendor specific differences in microbiota and cohousing. C57BL/6 mice obtained from two different commercial vendors, Jackson Laboratory and Taconic Farms, were observed to have significant differences in the proportion of Th17 cells in their small intestinal lamina propria [74]. Transfer of microbiota only from Taconic mice into a GF model induced accumulation of Th17 cells, and cohousing of mice from the two vendors allowed Th17 cell accumulation in Jackson mice. The hypothesis that this was due to the presence of a specific bacterial taxa in Taconic mice was confirmed; segmented filamentous bacteria were shown to induce Th17 cells and play significant roles in immune modulation [74]. Sequencing analysis of fecal microbiota of various strains from several vendors confirmed the major impact of vendor on microbial composition and development over 24 weeks of the mouse's life [75]. Strain-specific differences are well documented in mouse models, but vendor-specific differences at the level of the microbiota are often overlooked. These studies emphasize the importance of taking this into consideration when designing studies; results with mice from one vendor could be completely different with mice from another. The way in which mice are cohoused also has a demonstrated impact on the gut microbiome. Various studies have shown that cohousing mice results in transfer of the microbiota with healthy, advantageous species dominating the community [71]. Further, mice caged together have microbiota which are more similar in composition than mice in other cages, indicating the development of a cage microenvironment which can confound microbiome effects [76]. Introduction of an initial

common microbiome by gavaging mice altered microbial community composition but did not eliminate the impact of the cage microenvironment. Meaningful microbiome studies must not only choose an appropriate mouse model system, but must take into account the source of the mice as well as their living environment in assessing the impact of alterations in microbial communities.

Both GF and ABX-treated models exist in other species, and similar limitations apply. Each has specific advantages, but none are as similar genetically and physiologically as mice are to humans. Zebrafish are popular model systems due to their transparency until they reach adulthood, ease of maintenance and generation, the suite of genetic tools available, and the similarity of their gastrointestinal tract to mammals [60,66]. Rats contain many similarities to mice, including various disease-specific and genetically altered strains. The fruit fly *Drosophila* has powerful genetic tools available but lacks an adaptive immune system [60]. Various other mammals, such as pigs, dogs, and the bobtail squid have also been used to model the microbiome [60,66].

1.2.1.3 *In Vitro* and *Ex Vivo* Systems

Once observational and functional studies have been done in human and animal models, *in vitro* and *ex vivo* systems can be applied to examine the mechanism behind them. *In vitro* model systems for the gut consist of bioreactors and/or microchannels that have been designed to reproduce distinct functions of the microbiota but typically lack human cells [60,62]. *Ex vivo* systems incorporate cells taken from human donors and co-culture them with microbiota, allowing study of specific interactions between bacteria and cell subtypes [77,78]. These models provide the highest degree of control of

experimental variables, are less costly and higher throughput, and allow use of high-resolution molecular analyses [66].

In vitro gut bioreactors consist of either simple short-term incubations and single stage reactors or multi-compartmental continuous systems [62]. Short-term incubations and single reactors act as screening tools in which the ability of microbiota to metabolize or interact with various substrates can be measured. Continuous systems are made of multiple reactions and mimic the varying digestive capabilities of the human intestinal tract [62,66]. One of the earliest models consisted of three reaction vessels simulating the ascending, transverse, and distal colon at a pH of 6.0, 6.5, and 7.0, respectively [62]. More complex models, such as the Simulator of the Human Intestinal Microbial Ecosystem (SHIME), incorporate additional reaction vessels and internal components to more accurately model the length of the digestive system [62,66]. SHIME has five interconnected reaction tubes containing mixtures of luminal microbes that mimic digestion by acid and pepsin in the stomach, monosaccharide metabolism in the small intestine, and fermentation by microbes in the varying regions of the colon previously modeled [62,66]. Models such as the TIM2 build on the SHIME model by adding computer-controlled peristaltic mixing through application of pressure to the tubes, and absorption of water and microbial metabolites through use of a dialysis membrane [62,66]. These models are useful in evaluating the roles of the microbiota in digestion and metabolism of dietary components and drugs as well as studying the functional capacity of specific microbiota from healthy or diseased states. However, their major limitation is the lack of incorporation of human cells, which prevents study of host-microbe interactions.

There are a limited number of studies that have successfully co-cultured human and microbial cells [66]. Immortalized vaginal epithelial cells (VEC) grown at air-liquid interface (ALIF) form tight junctions and multilayers that resemble the stratified squamous epithelium of the vagina and respond in a similar manner to pro-inflammatory stimuli [79]. Normal commensal members of the vaginal microbiota colonized only the apical layer of the cultures and did not induce cytokine secretion but *Staphylococcus epidermidis*, a skin commensal, did. This system was used to determine the impact of vaginal microbial community composition on replication efficiency of human immunodeficiency virus-1 (HIV-1) and application of an antiretroviral medication, demonstrating its utility in reproducibly modeling the interactions of vaginal microbiota with both normal and infected host cells [77]. More complicated and recent model systems have used organoids and microfluidics [78,80]. Intestinal organoids are three-dimensional structures that differentiate into epithelial subtypes when cultured in a gel matrix [80]. Cells from the intestinal crypt, which include stem cells, are harvested from human donors for this *ex vivo* model. Gene expression quantification revealed that this model reflects *in vivo* expression levels of genes involved in the serotonin pathway. Supplementation of microbiota-derived factors induced expression of other serotonin-related genes, making it a useful tool in studying this pathway [80]. When exposed to cultured media from abundant commensal gut bacteria, organoids respond in a strain-specific manner [81]. Although organoids have not been exposed to complex communities of commensal bacteria, these studies imply that the system has the potential to elucidate specific microbe-host interactions in a physiologically relevant model. Finally, the most complicated of these models, called human-microbial crosstalk

(HuMiX) involves three co-laminar microchannels for co-culture of human intestinal epithelial cells with microbiota [78]. These three chambers each have inlet ports for inoculation of cells and perfusion media as well as outlet ports for collection of eluates for down-stream molecular analyses. They are separated by porous membranes to allow perfusion of media between the chamber layers but prevent cross-contamination, allowing long-term co-culture. The system also has integrated sensors to measure oxygen concentrations and allows for measurement of epithelial integrity by insertion of a chopstick-style electrode [78]. This complex model allows HuMiX to recapitulate *in vivo* transcriptional responses to bacteria, making it the most physiologically relevant of the *in vitro* and *ex vivo* models described here while allowing for high-resolution molecular analyses to understand host-microbe interactions. Models that integrate the multiple cell types that interact with the microbiome, allow for in-depth mechanistic studies, and simulate appropriate physiologic processes are necessary to understanding the role of microbiota in health and disease. Fine experimental control that provides reproducible data in the most physiologically relevant system will result in concrete strategies to manipulate the microbiota *in vivo* for improved health outcomes.

1.2.2 Sequencing Strategies and Technology

After extraction of DNA from microbiome samples, identification of individual community members is achieved through the use of sequencing methods. Due to rapid advancement in sequencing technology, next-generation methods are now used with more frequency than classic capillary Sanger sequencing [82]. Of existing sequencing methods, Sanger remains the gold standard due to its long read length, low error rate, and

larger insert sizes [82]. However, Sanger sequencing requires shearing of DNA into fragments that are then clonally amplified within a plasmid vector [83]. This process is laborious, time-consuming, and expensive [22,82,83]. Further, library preparation may be biased due to toxicity of gene content to the vector expressing it [83]. Next-generation sequencing (NGS) methods produce shorter read lengths at a fraction of the cost of Sanger sequencing, implement the use of polymerase chain reaction (PCR)-based amplification methods, and take much less time to run. Comparison of NGS and Sanger sequencing has demonstrated that shorter read lengths are similar in accuracy to longer Sanger read lengths in clustering analysis based on environment the sample came from [84]. Shorter reads also give comparable results at the level of microbial community composition as Sanger reads in samples from lean and obese people [85]. These advantages have led to a shift in the metagenomics field to use of NGS over Sanger sequencing for identification of microbiota.

1.2.2.1 Next-Generation Sequencing

The term NGS refers broadly to high-throughput methods that allow sequencing of millions to billions of DNA strands in parallel [86]. These methods encompass varying combinations of amplification, detection, and sequencing chemistry methods, providing researchers with an array of platforms to choose from that best meet their projects' needs. Table 1.2 provides a summary of the platforms applied to microbiome studies specifically. Though many of these have been used in microbiome studies, the Roche 454 pyrosequencing platform dominated the field, followed by take-over of Illumina's solid-state amplification and sequencing-by-synthesis technology.

Both Roche and Illumina platforms use sequencing-by-synthesis chemistry, in which a single nucleotide is incorporated into the sequence per cycle [2,83,86]. On the Roche 454 platform, DNA fragments are attached to microscopic beads for clonal amplification by emulsion PCR and then deposited into a plate for parallel pyrosequencing [82]. For each cycle, DNA polymerase incorporates the complementary deoxynucleoside triphosphate, releasing pyrophosphate [2,82,83]. Pyrophosphate is converted to adenosine triphosphate (ATP) by the enzyme sulfurylase in the presence of adenosine 5'-phosphosulfate [2,83]. ATP acts as a substrate for the enzyme luciferase, which produces light that is detected by a charge-coupled device camera that allows conversion to the template sequence [2,82,83]. The emulsion PCR used in this method is prone to generation of artificial replicate sequences, which can be dealt with appropriately using bioinformatics tools [82]. When the polymerase encounters homopolymers, the Roche platform often has difficulty in correlating the amount of light produced to the correct number of nucleotides [82]. Illumina's technology is superior in that it does not have this homopolymer issue due to its use of uniquely fluorescently labeled deoxynucleotides (dNTP). Unlike the Roche bead system, Illumina sequencing takes place on a flow cell to which a lawn of oligonucleotides is hybridized [87]. Single-stranded sequences complementary to these oligos, known as adapters, are attached by PCR to both ends of the DNA template to allow bridge amplification on the flow cell. The opposite end bends over to bind to its complement on the flow cell surface, creating a bridge-like structure. All sequences are amplified and then one copy cleaved off to allow additional amplification [87,88]. Bridge amplification can generate 800,000 – 1 million clusters per mm^2 of the flow cell surface, allowing massively parallel sequencing

of PCR clones of the original DNA template [87]. As with Roche 454 technology, Illumina employs the sequencing-by-synthesis method, adding all of the four possible dNTPs (A, C, T, or G) fluorescently labeled with a unique color at each cycle. Only the complementary dNTP is incorporated into the growing sequence, which is imaged and the color recorded as the appropriate base. A decrease in quality may appear

Platform	**Method**	**Read Length**	**Run Time**	**% Total Error Rate**	**16S rRNA**	**Whole Genome**
Sanger	Capillary-based, fluorescent dideoxy terminator	750 - 800 base pairs	2 hours	0.001	Full length; 2-3 reads	Long reads enable database comparisons
Illumina	Fluorescent sequencing-by-synthesis	36 – 151 base pairs	4 hours – 15 days	<1	1 variable region per read	Short reads aren't limiting
Roche 454	Pyrosequencing light emission	300 – 600 base pairs	9 – 23 hours	1	3 variable regions per read	Long reads enable database comparisons
IonTorrent	Proton detection	200 base pairs	2 – 3 hours	2	4 variable regions per read	Short reads aren't limiting
PacBio	Fluorescent single-molecule sequencing	250 base pairs to 40 kilobases	1.5 hours	15	Full-length reads	Long reads assist in assembly
Oxford Nanopore	Single-molecule sequencing detected by DNA passing through pore	230 – 300 kilobase pairs	1 minute – 48 hours	~30%	90% of full-length reads	Long reads assist in assembly

Table 1.2: Comparison of Sequencing Technology used in Microbiome Studies [3,86,89–91]**.**

at the end of Illumina reads, which is due to extension of some of the sequences either falling behind or getting ahead of the others [87,92]. Illumina has become the sequencing platform of choice for microbiome studies due to its increased accuracy, lower cost, and faster run time [82].

1.2.2.2 Microbiome Sequencing Methods

The HMP used two techniques to sequence the human microbiome; whole-genome sequencing (WGS) and 16S rRNA gene amplicon sequencing [8]. Both of these methods have been extensively used in other microbiome studies, with 16S rRNA gene sequencing much more widely employed. WGS is a powerful tool that captures the entirety of the bacterial genome, allowing identification of genes and bacterial functions following genome assembly [93]. It does not suffer from PCR bias, as 16S rRNA gene amplicon sequencing does, but it is less sensitive, more expensive, and genome assembly can be difficult and computationally complex [3,91,93]. 16S rRNA gene sequencing relies on amplification of variable regions within the bacterial 16S ribosomal gene for bacterial identification. Due to the short length of the read, functional information cannot be inferred from this method. However, 16S rRNA gene sequencing has become the method of choice due to its cost-effectiveness, speed, and accuracy in identifying bacterial communities [93].

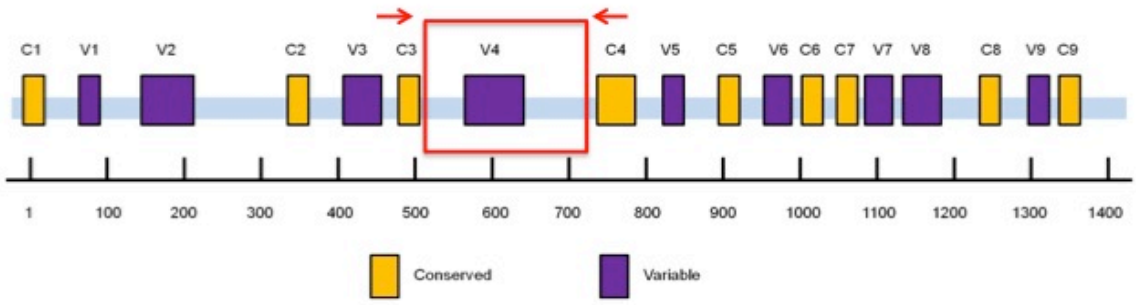

Figure 1.2: The Bacterial 16S rRNA Gene. The 16S rRNA gene contains nine conserved and nine hypervariable regions. Primers (red arrows) are designed that anneal to universally conserved regions and amplify a variable region (red box) that allows identification of the organisms present.

1.2.2.2.1 16S rRNA Gene Amplicon Sequencing

16S rRNA gene amplicon sequencing is dependent on amplification of the universally conserved bacterial 16S ribosomal gene [2]. It contains nine hypervariable regions between which nine conserved regions are interspersed, allowing design of primers that anneal to conserved regions and amplify a specific variable region for sequencing [2,3]. Figure 1.2 depicts conserved and hypervariable regions within the 16S rRNA gene. The primer set in Figure 1.2 corresponds to the highly used 515F/806R set developed by Caporaso *et. al* [94]. These primers specifically amplify the V4 hypervariable region which Caporaso *et. al.* have determined to yield optimal results based on the length of Illumina sequencing reads. However, the optimal variable region for amplifying and detecting the range of bacterial species in a microbiome sample has yet to be determined [22,91]. Despite the conserved nature of the gene, there is some variability among the conserved regions between bacteria, making certain primer sets better suited to anneal to these regions in specific populations rather than others. This results in over- or under-representation of certain taxa depending on the primer set used

[22]. For example, the previously mentioned primer 515F, as well as several others, matches greater than 95% of sequences of bacterial phyla in the gut, while 784F is biased against Verrucomicrobia, and 967F matches less than 5% of Bacteroidetes, a major constituent of the gut [22]. Primer bias clearly alters detection of bacteria within the microbiota, leading to significant differences in diversity metrics depending on region sequenced and making it impossible to compare studies using different primer sets [22,95]. Standardization of primer sets among microbiome studies has been suggested to address this problem, but ignores the issue of taxonomic under-representation by the specified primers in communities where the missing taxa could be crucial to functionality. WGS, which does not depend on amplification and therefore bypasses primer bias, is a viable alternative to amplicon sequencing for microbiome studies, but is not widely used due to cost and computational complexity.

The short read length of Illumina technology was initially a concern in the microbiome field, which had previously relied upon Roche 454 reads that were nearly twice the length [94]. Longer 454 reads allow up to three hypervariable regions to be sequenced per read, whereas Illumina's 150 base pair length allows a single region per read [3]. The shorter read length prevents high resolution of individual taxa (*i.e.* down to species level) but the greater number of reads and sequencing depth makes Illumina's technology superior in overall community resolution [95]. Caporaso *et. al.* compared microbiome amplicon sequencing using both Roche and Illumina platforms and discovered that between-sample beta diversity metrics did not significantly differ, leading them to conclude that the shorter read length would not significantly alter sequencing results [94]. Further, Illumina's greater number of reads makes it possible to compare

either thousands of samples per run or fewer samples with much higher coverage, which has not been possible with previous sequencing technology [94]. Illumina's fluorescently labeled dNTPs solve Roche's homopolymer issue, making it more accurate. The advantages of Illumina's platform have made it the leader in microbiome amplicon sequencing studies.

1.2.2.2.2 Whole Genome Sequencing

As previously mentioned, WGS avoids the primer bias experienced by 16S rRNA gene amplicon sequencing and allows investigation of both community composition and function [2,3,93]. Similar to amplicon sequencing, the first step in WGS is to extract DNA from experimental samples. However, unlike amplicon sequencing, the DNA from the entire population is fragmented to a specific size and sequenced without amplification of a hypervariable region of the 16S rRNA gene [1]. Due to its requirement for a greater amount of starting material, some studies have used whole genome amplification on low-DNA samples and found that amplification bias by this method can be minimized [2]. Samples used in WGS contain DNA from the host species as well as bacteria, viruses, fungi, and bacteriophages. Particularly in the case of whole genome amplification, the host DNA may out-number microbial DNA, which can bias sequencing due to limits in depth and coverage [91]. Though Illumina's ability to generate many thousands of reads per sequencing run provides greater coverage for complex communities it still only covers fragments of genes from each species within the community [83]. Assembly of sequenced genomic fragments is necessary to elucidating functional content of the microbes within the community as it can reveal open reading frames, operons, and

transcriptional elements with their associated promoters and binding sites [83]. In complex microbiome samples, genomes from each community member are only partially sequenced in short fragments and a lack of bacterial reference genomes creates difficulty in appropriately mapping the reads to the correct species. Despite these challenges, the HMP has successfully used WGS to demonstrate the relative consistency of microbial community functions from a range of body sites in healthy adults [8]. Although computationally complex, WGS provides an additional layer to metagenomic community sampling that aids in understanding the functional roles of different microbial communities as well as their composition.

1.2.3 Data Analysis

After sequencing, raw reads require a number of processing steps before the bacteria they represent can be identified and statistically analyzed. Various open-source tools have been developed to complete distinct steps in this process but, prior to 2009, they did not exist in a single package that could take raw reads as input and give publication-quality figures as output. To address this issue, metagenomics pipelines were developed. These pipelines integrate other independently developed tools into a streamlined platform, providing ease of use and increased reproducibility. Once WGS and amplicon reads have been identified as bacterial taxa and/or functions, statistical analysis incorporating sample metadata (such as patient physical characteristics, disease state, and diet) can be completed either within the pipelines or using other tools. Pie charts, heatmaps, and dot plots are commonly used for visualization of similarity among bacterial communities, and specific statistical methods have been developed to determine

significance of differences among bacterial abundances. Though choice of pipeline may be based on user preference, care must be taken in use of statistical methods for data analysis. An understanding of both computational and statistical methods is necessary in choosing the appropriate test to make valid biological conclusions.

1.2.3.1 Analysis Pipelines

Metagenomics pipelines were developed to streamline the analysis of both WGS and amplicon sequencing data. Very generally, a pipeline is a series of data input/output steps that is automated to run with just a few commands. Metagenomics pipelines take raw reads in FASTA and/or FASTQ file format, filter out low-quality reads, demultiplex sample barcodes, and trim off primer sequences [2]. After this pre-processing step, the samples are treated differently depending on whether they are WGS or 16S rRNA gene amplicon sequences. If they are WGS, assembly of the read fragments is required followed by comparison to a database to annotate the reads. If they are 16S rRNA gene amplicons, they are grouped into operational taxonomic units (OTU) by percent sequence similarity. This serves as a proxy of species-level taxonomy and reduces computational complexity [96]. Once grouped into OTUs, the sequences are compared to one of several 16S rRNA gene databases in order to identify the bacteria present. The number of sequences corresponding to an OTU is used to construct a raw count table, which can be normalized to provide relative abundances of bacterial taxa. This normalized abundance table is then used in downstream statistical analyses, such as calculation of species diversity and comparison of between- and within-community similarity [2,93].

The earliest and most popular of the pipelines are mothur [97], developed by Pat Schloss's group at the University of Michigan, and Quantitative Insights into Microbial Ecology (QIIME, pronounced 'chime') [98], developed by Rob Knight's group at the University of Colorado at Boulder. A less popular but more user-friendly platform was developed by Susan Huse, David Welch, and Mitch Sogin at the Marine Biological Laboratory and is known as Visualization and Analysis of Microbial Population Structures (VAMPS) [99]. Tools for analysis of microbial communities are constantly evolving and improving, and, as such, the creators of each of these pipelines are continually releasing updated versions, resulting in increasing accuracy and reproducibility of data analyzed by each. The choice of which pipeline to use for data analysis depends mainly on the user's experience in bioinformatics, as each is excellent but has its own advantages and disadvantages.

Both QIIME and mothur are implemented in the command line and require the user to be able to operate in a Linux environment as well as have some basic knowledge of the programming language each is written in. The scripts within QIIME are written in the language Python, while those within mothur are written in C++. Python is a flexible programming language that is easy to learn and use, but can be slow due to the nature of its implementation [100]. C++ is more complex than Python, making it more difficult to learn and use, but its implementation makes it much faster [97]. When analyzing large metagenomic datasets, computational power and speed are factors that must be considered when choosing an appropriate analysis pipeline.

Despite mothur's increased analysis speed, QIIME is more frequently used, likely due to the extensive tutorials, workshops, and support offered by Dr. Knight's group. At

the time of this writing, the original article describing QIIME was cited 2,146 times in Pubmed, while the article introducing mothur was cited 1,899 times. Both QIIME and mothur are free and open-source and contain scripts for parsing FASTQ files, demultiplexing barcodes, trimming sequences, denoising sequencing errors, and identifying and eliminating chimeric sequences that result from PCR errors. Mothur contains scripts to bin sequences into OTUs by percent similarity, but QIIME takes this a step further and provides three separate strategies for grouping OTUs that implement external OTU clustering tools. The first two of these strategies are *de novo* and closed-reference OTU picking. *De novo* OTU picking algorithms cluster sequences together without comparison to reference sequences. Closed-reference strategies cluster reads against reference sequences and eliminate those that do not match any of the references. Open-reference OTU picking combines these two strategies; reads are matched to reference sequences and clustered *de novo* if they do not match. QIIME's default external tool for this process is UCLUST, which is an algorithm developed by Robert C. Edgar [101]. The UCLUST algorithm improved on the speed, computational power, and quality of clustering as compared to other commonly used methods, such as CD-HIT. External OTU clustering tools are also wrapped into mothur, and include DOTUR and CD-HIT [97]. Although the OTU picking methods of QIIME and mothur are valid, differences in the algorithms used has been demonstrated to give different clustering results. He *et. al.* recently showed that many commonly used OTU clustering methods produce unstable OTUs, where membership changes based on the number of sequences clustered [96]. Varying assignment of sequences to OTUs could result in very different biological conclusions and make reproducibility impossible. Closed-reference OTU picking

produces the best OTU stability, but eliminates discovery of new species by discarding reads without a previously sequenced match. Open-reference picking, with a *de novo* clustering algorithm that is more stable, may be the best solution to this issue [96]. Understanding the nuances behind options such as OTU picking strategies is critical to getting high-quality data from metagenomics pipelines.

Beyond quality filtering and OTU picking, both mothur and QIIME contain tools for statistical data analysis and visualization. Mothur groups its scripts into OTU-based and hypothesis testing approaches, while QIIME has a variety of scripts for analyzing microbial diversity. In mothur, OTU-based approaches encompass calculation of microbial community diversity based on ecological measures. Mothur's hypothesis testing approaches include statistical analysis of distance metrics, analysis of variance, and co-occurrence. The QIIME website includes tutorials on using analysis of variance (ANOVA) to compare categories, distance metric comparison, network building, supervised learning algorithms, and microbial source tracking. QIIME also integrates software that displays the data in visually appealing formats. Both pipelines produce publication-quality figures from statistical analyses.

Though QIIME and mothur are powerful analysis pipelines, they can be daunting to begin using due to their command-line interface. VAMPS was developed to address this issue and operates on a more intuitive web-based platform [99]. Like QIIME and mothur, VAMPS incorporates externally developed tools as well as scripts written by the developers to take raw sequence reads and produce publication-quality statistical analyses and figures. It also allows use of the statistical programming language R and code from

QIIME and mothur. VAMPS is an excellent starting point for sequencing analysis for biologists with limited coding experience.

QIIME, mothur, and VAMPS were all initially developed to analyze 16S rRNA gene amplicon sequencing data, not WGS reads [97–99]. QIIME currently contains code under development to analyze WGS reads but mothur and VAMPS do not. As with amplicon sequencing, many tools exist to perform the various steps involved in WGS assembly and annotation but there are few pipelines that streamline these tools in order to take raw reads to publication-ready figures. WGS reads are generally either assembled by mapping to reference genomes or *de novo*, without a reference genome [102]. Reference-based mapping is limited by existing reference genomes while *de novo* assembly requires more computational power and memory. Once assembled, reads are binned into taxonomic groups in one of two methods. The first is based on the distribution of the k-mers, or fixed-length 'DNA words', among genomes. Different genomes have unique k-mer distributions, which allows grouping independent of a reference. In the second method reads are aligned to a reference and binned based on similarity. Finally, assembled genomes can be annotated to identify coding, noncoding, and other regulatory regions.

Analysis of WGS reads from microbiome samples is challenging due to the randomness with which the genomes are sampled [103]. WGS reads are generally short and represent a randomly selected portion of individuals in the community. These reads may or may not overlap, which is necessary to assemble the genome and identify the bacteria present. In answer to this issue, C. Titus Brown's group has developed khmer, a pipeline for analyzing WGS reads from microbial communities as well as mRNA [104].

Unfortunately, truncating sequences to create k-mers and considering them simultaneously for assembly requires a large amount of computational memory. Khmer deals with this issue by partitioning the reads into different files. The program counts k-mers and calculates their abundance and performs 'digital normalization,' in which k-mer abundance is normalized by eliminating redundant reads covering the same portion of the genome and keeping just enough to allow efficient genome assembly. Khmer does not include code to annotate assembled genomes, so users will need to seek external tools to perform this step. Though other assembly tools exist, khmer is a good choice for efficient, reproducible WGS read assembly and is relatively easy to use. Ultimately, the choice of analysis pipeline depends on the goals of the study and the researcher.

1.2.3.2 Statistical Analysis of Compositional Data

Once amplicon or WGS reads have been transformed into OTU tables, statistical analysis can be done. The complex nature of microbiome datasets makes them challenging to analyze appropriately and improved methods are constantly being developed to do it better. Various unsupervised, exploratory methods have been used, such as clustering and resampling methods, as well as univariate and non-parametric models [105]. Multivariate statistics have been developed and applied to microbiome datasets but may fail to be appropriate for the data as they tend to assume linearity when microbiome data is generally curved [105,106]. Other statistical challenges include the compositional nature of the data and its sparseness [106,107]. Due to variation and error in PCR and sequencing, it is not possible to get absolute abundances of bacterial taxa from microbiome sequencing data. However, relative abundances can be calculated in

which the percent composition of each taxa totals to 100% for each individual sample. This compositional nature means that changes in abundance of one taxa will drive changes in the others, since the data is forced to sum to a constant [107]. Rarefaction, which randomly resamples to the size of the smallest library, has been used to correct for this compositional data, but it has been argued that rarefaction is inappropriate to use for this purpose [108]. McMurdie *et. al.* demonstrated that rarefaction results in high false positive rates when identifying significant differences in species abundance and it eliminates sequences that can be appropriately clustered using other methods. The continued use and prevalence of rarefaction in the microbiome field highlights how important it is for biologists to understand the theory behind statistical and computational methods to analyze microbiome data. Inappropriate application of statistical models can lead to conclusions that do not support the underlying biology.

Development of appropriate statistical analysis methods for microbiome data is challenging. Corrections for the compositional nature of the data include log-ratio transformations which, in theory, do not alter underlying covariance or correlation among the data and allow application of traditional statistical analyses [107]. However, the sparseness of microbiome data often makes this transformation problematic, as it requires dividing by the geometric mean of the taxa. If the mean is zero, the value becomes undefined. Pseudo-counts have been used to correct for this, in which the same random, small number is added to all counts so that none are zero. This poses problems too, as division by 1 is the same as analyzing unnormalized data and the consequences of using other values is not well understood, particularly in light of the importance low-abundance taxa may play in microbial communities. Despite these issues, transformation of

compositional microbiome data enables use of traditional statistics methods to determine significant changes in microbial populations.

Microbiome statistical analysis methods draw heavily on diversity methods from the ecology field. Species diversity indices are widely used to simplify complex microbial communities by assigning values that represent overall trends in the population [109,110]. These indices have been used to compare changes in microbiota diversity according to relevant community variables, such as environment and patient disease state. They fall into one of two categories; alpha diversity, which quantifies within-sample taxa diversity, and beta diversity, which quantifies between-sample diversity [106]. Several methods exist to calculate alpha diversity, including the widely-used Shannon and Simpson indices [110]. Both of these indices combine measures of taxa richness (the number of different taxa) and abundance but do so with different underlying theoretical foundations. The Shannon index is abstract and represents uncertainty in identifying unknown taxa, while the Simpson index is more intuitive and indicates the probability of two randomly chosen taxa belonging to different species [110]. Species evenness, or the number of individuals within each taxon, can be derived from both of these indices. The Chao 1 index is used with less frequency but is a non-parametic method that can estimate OTU richness and performs well with low-abundance communities [109,111]. Several R packages can be used to calculate alpha diversity, including vegan [112] and phyloseq [113]. QIIME and mothur contain scripts for alpha diversity as well, as does the open-source software Explicet [114]. Several studies have compared the usefulness of these indices when applied to metagenomic datasets and generally agree that all three are appropriate, despite their varying foundational theories, and suggest that studies may

benefit from using and comparing all of them to determine interactions within communities [109–111].

Beta diversity measures allow comparison of similarity and dissimilarity among microbial communities. They are particularly important in identifying trends over time within large datasets [115]. Commonly used beta diversity measures include Morisita-Horn similarity and Bray-Curtis dissimilarity [114]. Both of these beta indices and the previously described alpha indices are based on normalized counts of taxa and do not take phylogenetic relationships into account. Phylogeny indicates the evolutionary history of organisms, and trees can be built in order to represent these relationships [106,116]. Fast UniFrac is a popular beta diversity method that calculates distance of relatedness of microbiota based on the branch lengths of phylogenetic trees [116]. The fact that it is a distance metric allows analysis of the resulting data with standard multivariate methods, such as clustering and principle components analysis (PCA). Fast UniFrac was developed by the Knight lab and is included in both the QIIME and mothur pipelines. Visualization of Fast UniFrac data with PCA plots allows easy identification of community similarity by clustering. Fast UniFrac has been cited in over 200 papers and has been used to compare similarity among environmental and host-associated microbial communities. Several recent papers have used it to compare bacterial communities in sludge systems [117], subtropical rainforests [118], and recurrent aphthous stomatitis, an oral mucosal disorder, in patients [119]. Each of these studies also employs a range of alpha diversity indices to compare microbiota. Beta diversity measures are useful in understanding overall trends and changes between samples in a study.

Though diversity indices are useful in assessing overall trends among microbiota, identifying differential abundance of individual taxa among groups may indicate specific bacteria that play important roles in the environment or disease state. Besides the compositional nature of microbiome data, both its sparseness and its tendency to be dominated by a few taxa make appropriately modeling this data difficult [106]. As mentioned previously, log-ratio transformations can be used in order to apply standard downstream statistical analyses. Dirichlet multinomial mixtures have been developed that take into account data sparsity as well as the presence of diverse and rare taxa [120]. Two-sample t-tests have been employed to determine differential abundance among abundant taxa and Fisher's exact test has been used for rare taxa [106]. Variations on the Wilcoxon rank-sum test have been used as well [114]. Specific tools have been developed to manage the challenges of microbiome data and identify significantly enriched taxa. Curtis Huttenhower's group at Harvard has developed a suite of analysis tools written in a combination of Python, R, and Perl that perform both compositional and statistical data analysis. These tools have been implemented in the Galaxy platform, which is a web-based environment that allows researchers without a programming background to analyze high-throughput data [121]. The Huttenhower group's programs LEfSe and MaAsLin can be used within Galaxy to determine significant enrichment of bacterial taxa based on relevant biological information [122]. LEfSe employs a combination of the Kruskal-Wallis rank sum test, the Wilcoxon rank-sum test, and linear discriminant analysis in order to rank significant enrichment of bacteria between two biological classes, such as diseased and healthy. Wu *et. al.* recently used LEfSe to detect bacteria significantly enriched among gut microbiota of normal control mice and those

exposed to lead [123]. MaAsLin takes this a step further and allows detection of enriched taxa among multiple biological classes. The previously mentioned pipelines and R packages also contain methods to detect differentially abundant taxa and visualization options to compare them. Given the issues in appropriate statistical methods to detect enriched taxa, experimental methods such as quantitative PCR (qPCR) should be used to confirm bacterial abundances.

While diversity indices provide overall trends among microbiome data and differential abundance detects changes in specific taxa, co-occurrence relationships and network analyses aim to understand how the microbes in a community interact with each other or respond to specific variables [124]. Rather than describing how and to what degree microbial communities change, network and co-occurrence analyses predict how taxa influence each other or are altered by outside variables through the use of correlation coefficients and networks. These methods are a type of dimensionality reduction, in which complex microbiome data can be mathematically condensed into a simpler version that is easier to interpret and understand. Various studies have used both Pearson and Spearman methods to calculate correlation coefficients for changes in microbial taxa and external factors, such as exercise [125] and bacterial metabolites [126]. Though the Pearson method is appropriate for parametric data and Spearman for non-parametric, neither of these methods takes the compositional nature of the data into account [127]. Sparse Correlation for Compositional Data, or SparCC, was developed to determine pairwise correlations between microbial taxa while correcting for the data's compositional nature [128]. It relies on log-ratio transformation of the data and has been shown to produce fewer false correlations than the Pearson method. Another method,

Sparse Inverse Covariance Estimation for Ecological Association Inference (SPIEC-EASI), dispenses with pairwise correlations and instead attempts to infer the entire correlation network simultaneously [129]. It does this through use of a graphical model inference framework that assumes the data is compositional and sparse. Though SPIEC-EASI is more reproducible than SparCC, results from the methods are not directly comparable due to the different ways in which they calculate microbial correlations. Each of these tools is useful in determining microbe-microbe and microbe-external variable correlations that could indicate their potential in predicting interactions or outcomes.

Statistical analysis of microbiome data is complicated and requires knowledge of both the biology behind the study as well as the mathematics driving the models and algorithms employed. Appropriate models are still under active study, making it crucial that biologists collaborate with statisticians and bioinformaticians in order to choose the best model for the data to generate valid biological conclusions.

CHAPTER 2: BURN AND INHALATION INJURY AND ITS RELATION TO THE AIRWAY MICROBIOME

2.1 The Airway Microbiome

Development of standardized methods and generation of data from healthy subjects by the HMP provided an invaluable resource to other microbiome researchers. Metagenomic profiling of various body sites from healthy individuals gave a baseline for healthy microbial community structure, allowing comparison to other disease and exposure states [20]. However, the HMP's sole sampling site for the airways was the nasal cavity [20]. The mucosal surfaces of the nasal passages are known colonization sites for commensal bacteria but little work had been done on their roles in health prior to the HMP [130–132]. The HMP revealed microbial composition and functional dynamics in the nose, but lacked sampling further down the airways. Traditionally, the respiratory tract below the larynx was considered sterile due to inability to culture organisms here [35,133]. Culture-independent sequencing techniques have challenged this belief and suggest that there is a diverse but low-abundance population of bacteria present in the healthy lungs [35,133–135]. Investigation into their roles in health and disease is ongoing, and several theories exist as to how a healthy population of bacteria is maintained in the lungs. Like the gut, the lower airways are lined by mucosal surfaces and contain ciliated cells which beat to move mucus through the airway lumen [136]. This mucociliary escalator, in combination with airway innate immune responses, is

thought to keep the lungs relatively free of foreign particles, including bacteria [136,137]. However, the airways are constantly exposed to inhaled air, which contains suspended microbes as well as dust particles to which the organisms can adhere [138]. A study in Japan found concentrations of bacteria in dusty air to be as high as 1.6×10^7 cells m^{-3}, which was two orders of magnitude higher than non-dusty air [138]. Subclinical microaspiration is also a significant source of microbial immigration to the airways [139]. This implies that bacteria are constantly entering the airways, and, though they may be quickly cleared, they are interacting with airway cells. An important question is how this interaction takes place and its impact on airway physiology and host health outcomes. A recently published theory suggests that airway homeostasis is maintained in the presence of bacteria through a balance of microbial immigration and elimination mirroring the equilibrium model of island biogeography [139]. This model is taken from ecology and posits that the diversity of species on an island is dependent on the balance of immigration and extinction [140]. Since an island is a closed system and contains limited resources, the immigration rate will fall as the number of species increases and the extinction rate will simultaneously increase [140]. If the lungs are regarded as an island, this model explains the consistently diverse yet low-abundance communities claimed to be found in healthy people at homeostasis [139]. Though not investigated by the HMP, recent work has explored the possibility of healthy airway microbiota as well as their roles in disease.

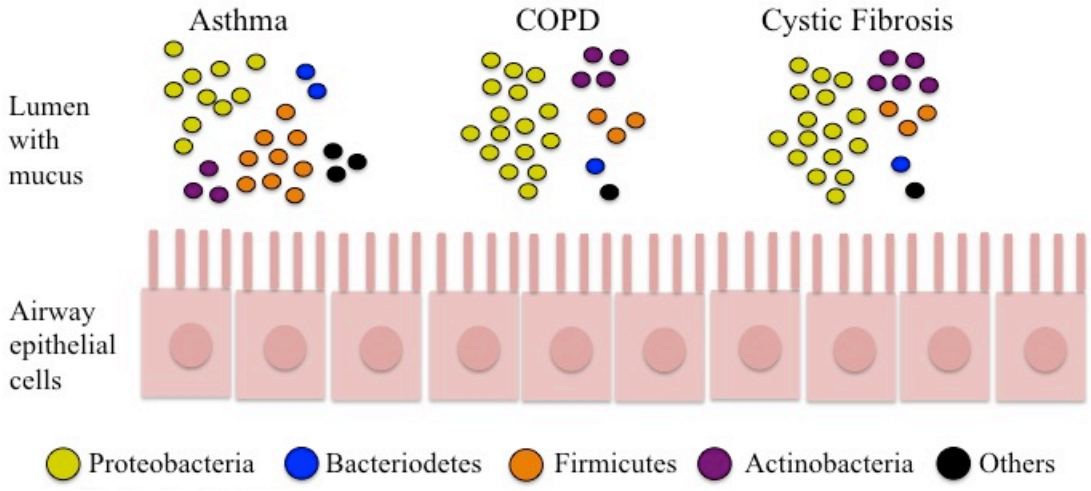

Figure 2.1: Dysbiosis in the Diseased Airway Microbiome. Outgrowth of specific phyla accompanied by an overall loss of microbiota diversity is observed in asthma, COPD, and cystic fibrosis. Adapted from Marsland *et al*, 2013.

Homeostasis within the airways can be disrupted by both injury and development of disease, either of which may lead to conditions favorable for bacterial colonization [139]. Under conditions of homeostasis, the immigration/extinction balance of the microbiota is maintained through scarcity of nutrients, mucociliary clearance, and innate immune defenses [133,137,139]. Disease or injury may alter conditions in the airway to favor bacterial growth through an influx of nutrients with concurrent inhibition of immune defenses, leading to unchecked bacterial growth [139,141]. An increased burden of bacteria within the airways induces inflammation and results in a positive feedback loop, furthering airway injury and bacterial growth, which could lead to dysbiosis and infection [139,141]. In agreement with this theory, bacteria are found at much higher abundance in airways diseased and damaged by conditions such as asthma, bronchiectasis, chronic obstructive pulmonary disease (COPD), cystic fibrosis (CF), idiopathic pulmonary fibrosis (IPF), and smoking [58,59,135,142–150]. Figure 2.1 illustrates changes at the phylum level in asthma, COPD, and CF [35]. Bronchial

brushings from patients with asthma show an increase in bacterial diversity and the specific families Comomonadaceae, Sphinomonadaceae, and Oxalobacteraceae that are correlated with severity of bronchial hyperresponsiveness [145]. Transcriptomic analysis of host and microbiota function in children with asthma demonstrate metabolic differences and association of microbial adhesion factors and increases in Proteobacteria with the cytokine IL1A, suggesting that the microbiome modulates host inflammation and immune response in this airway disease [142]. The lung microbiota of smokers is often studied in conjunction with COPD since smoking increases the risk of this disease. During COPD exacerbations, diversity in the airway microbiome is increased, patient samples are more similar by inhaled corticosteroid and bronchodilator treatment, and several oral taxa are enriched [143]. A separate COPD microbiome study, which also examined location-specific communities in the lungs, found lower diversity in two out of four patients with COPD and extensive overlap in bacterial taxa among healthy patients, healthy smokers, and those with COPD [144]. This overlap lead the authors to propose a core pulmonary microbiota consisting of *Pseudomonas, Streptococcus, Prevotella, Fusobacteria, Haemophilus, Veillonella,* and *Porphyomonas* genera [144]. A study comparing the upper and lower respiratory tract in non-smokers and smokers found enrichment of *Enterobacter, Haemophilus, Methylobacter* and *Ralstonin* in the lower airways, which were also present in the upper airways [58]. Only *Tropheryma* was unique to the lower airways. Further, this study found no difference in lower airway microbiota among smokers but did find changes in the upper airways [58]. In IPF, disease progression has been specifically associated with increases in *Streptococcus* and *Staphylococcus* genera [146]. Patients with cystic fibrosis are prone to polymicrobial

bacterial infections, which tend to be dominated by a specific genus but remain relatively stable during exacerbations [148]. Finally, injury to the airways by intubation and mechanical ventilation demonstrates decreasing microbial diversity with increasing time spent on the ventilator [135]. Though these diseases are known to predispose patients to airway bacterial colonization, these studies show their influence on airway microbial communities, which provides insight into bacterial interactions that may influence therapeutic treatment strategies.

Study of the airway microbiome is relatively recent, and the studies above demonstrate some of the challenges in accurately sampling and sequencing these low-abundance communities. Whereas loss of diversity in the gut is clearly associated with increased severity of disease, its role in the airway is not as straightforward. In these studies, airway diversity was seen both to increase with disease, opposite to the trend in the gut, and to decrease with disease. This lack of consistency may reflect true biology, but it is more likely that differences in sampling, extraction, and sequencing methods are contributing to these disparate results.

In the gut, where bacterial abundance is high, these technical differences do not contribute as drastically to the results as they do in the airways. Table 2.1 compares sampling, extraction, and sequencing methods for the airway studies. Though there are similarities among the individual steps, the differences make each study's entire protocol unique. Generally, airway samples were taken by bronchial washings or brushings, with just two using sputum and swabs from endotracheal tubes. During bronchoscopy, a sterile tube is inserted either through the nasal passages or the oral cavity, passed through the trachea and larynx, and wedged into the lung region of interest.

Study [Reference #]	Airway Sampling Method	DNA Extraction Method	qPCR to Confirm Bacterial DNA?	Sequencing Method + Variable Region
Kelly, 2016 [135]	Frozen swabbing and suctioning of endotracheal tubes	Bead beating + Mo Bio PowerSoil Kit	No	V1 – V2 Diseased: Illumina MiSeq Healthy: Roche 454 GS-FLX
Carmody, 2015 [148]	Frozen sputum	Sputolysin + bead beating + MagNA Pure kit	No	V3 – V5 Roche 454 with HMP protocols
Morris, 2013 [58]	Bronchial + oral washings	Standard at each center	No	V1 – V3 + V3 – V5 Roche 454 FLX Titanium
Han, 2014 [146]	Frozen bronchial washing pellet + some explanted lung tissue	Bead beating for tissue only + Qiagen DNeasy Blood & Tissue Kit	No	V3 – V5 Roche 454 GS Junior
Huang, 2011 [145]	Bronchial brushes	Bead beating + Qiagen AllPrep Kit	Yes	Microarray + 16S rRNA gene clone libraries of hybridized DNA ABI PRISM 3730 capillary sequencing
Pragman, 2012 [143]	Frozen and fresh bronchial washings	Bead beating + RNase treatment + own protocol	No	V3 Multiple displacement amplification Roche 454 FLX
Erb-Downward, 2011 [144]	Bronchial washing pellet + whole lung tissue	Bacterial lysis buffer + bead beating + Proteinase K treatment + MagNA Pure Kit	Yes	V1 – V3 Roche 454 FLX Titanium

Table 2.1: Sampling, Extraction, and Sequencing Methods Among Lung Microbiome Studies.

From here, a bronchoalveolar lavage (BAL) is performed by flushing sterile saline through the bronchoscope into the right or left lobe and then removing it by suction. A brushing can be done by instead inserting a small, sterile, wire-bristled brush into the same location and physically scraping the airway surface to collect epithelial cells. There are advantages and disadvantages to both methods. With a bronchial washing or BAL, a larger area of the lungs can be sampled since the fluid can travel throughout the tissue. With brushings, only a small area of the tissue is sampled, but the physical scraping removes more mucus and some tissue beneath it, allowing detection of bacteria that may be within the mucus layer or adhered to the epithelium below. Bronchial brushes are usually done with a protected brush and thus are considered less prone to contamination since the brush is only brought into the distal airway once in the region of interest. With either method, insertion through the nasal or oral cavity invites contamination from these densely population areas, creating a need for simultaneous oral and/or nasal washes in order to detect non-lung organisms. Intubation can significantly reduce the risk of contamination but has more potential side effects. Of the studies listed in Table 2.1, only Morris *et al*. sequenced both oral washings and BAL and compared them in order to detect bacteria unique to the lower respiratory tract [58]. Among the 64 study subjects, they found enrichment of several bacterial taxa in the lungs that were also present in the mouth, and only one taxa uniquely associated with the lungs [58]. Based on application of the neutral model [151], they hypothesize that these enriched and uniquely represented species may be part of a healthy lung microbiome. However, this study did not entirely eliminate the possibility that these species are contaminants from the upper airways pushed down by bronchoscopy. The subjects used antiseptic mouthwash prior to

bronchoscopy in an attempt to control for this, but re-sampling of the oral microbial population to identify bacteria present after the mouthwash was not done. None of the other studies in Table 2.1 indicated any attempt to control for upper airway contamination during bronchoscopy, making it possible that the so-called lower airway microbiome actually reflects that of the upper airways.

After collection, samples are either immediately extracted for sequencing or frozen for later extraction. The impact of extracting fresh or frozen samples on microbial diversity has not been examined extensively, but a single study found significant differences in two bacterial genera among gut samples that were extracted either fresh or frozen [152]. The authors conclude that freezing preserves the integrity of microbial diversity, but this does not take into account the impact that specific species and even strains can have on host physiology. This is particularly important in lower airway samples, which have much fewer bacteria than the gut, enhancing the impact of individual, low-abundance bacteria. The use of frozen samples in three of the studies in Table 2.1 may not reflect true biology among the microbiota of their subject populations as accurately as fresh samples.

After sample collection but prior to DNA extraction, it is wise to employ a method to measure the ratio of live to dead cells. Though dead cells may elicit responses from both other bacteria and the host immune system, live cells may be more important in community dynamics and host-microbiome interactions. Regardless, differentiation between these states may reveal the roles of specific bacteria within the airway ecosystem. Unfortunately, this is rarely done in microbiome studies and can often be difficult to do. None of the studies in Table 2.1 assessed viability of bacterial cells in their

samples. Metagenomic sequencing methods are incapable of detecting, in single samples, the viability of the organisms present [153]. All bacterial DNA is detected and identified, regardless of whether it came from a live or dead organism. The biochemical processes of living and actively reproducing organisms will have vastly different effects on host physiology than dead organisms, whose metabolic activities have stopped. Differentiation among live and dead bacteria in the studies in Table 2.1 may have led to different conclusions.

Following sample collection, an appropriate DNA extraction method is necessary to release nucleic acids from all bacteria present. Bead-beating in combination with phenol:chloroform:isoamyl alcohol has been commonly used [34,154–156] along with extraction kits developed specifically for either microbiome studies or purification of DNA from bacteria [16,32,59,73,135,157,158]. Efficient extraction of the diverse bacteria present in a community while preserving nucleic acid integrity for sequencing can be challenging, particularly for low-abundance samples. Bead beating is commonly used as a physical lysis method followed by chemical or enzymatic lysis and DNA purification methods. The HMP employed MoBio Laboratories' PowerSoil DNA Isolation Kit, in which samples are added to tubes pre-filled with 0.7mm garnet beads [159]. A solution containing sodium dodecyl sulfate (SDS) is added to chemically lyse cells before the samples are homogenized in a 10 minute vortexing step. This step can alternatively be done in 30 seconds using a bench-top homogenizer. Because it is designed for soil samples, the PowerSoil kit contains several steps designed to remove non-DNA and inorganic substances such as humic acid. DNA purification is done on a silica spin column, to which DNA binds after addition of a high-salt solution. Other DNA

kits, such as those from QIAGEN and MP Biomedicals, follow this general protocol. Several studies have tested the effectiveness of these varying protocols, including the necessity of bead beating. Two studies compared the effectiveness of physical and enzymatic lysis methods of bacterial DNA in soils through denaturing gradient gel electrophoresis and assessed purity by absorbance ratios [160,161]. de Lipthay *et. al.* compared bead beating, sonication, and grinding-freeze-thawing methods and found that bead beating gave consistently diverse and higher molecular weight DNA than the other methods [161]. Further, the results were more reproducible. Yeates *et. al.* compared bead beating with sonication and enzymatic lysis and also found that bead beating gave the highest bacterial diversity as assessed by number of bands on the gel [160]. de Boer *et. al.* recognized the need for optimization of bead beating protocols, as too vigorous methods result in DNA shearing [162]. Here, silica beads of 0.1mm in diameter resulted in increased detection of Gram positive species without compromising detection of the Gram negative. Though these methods demonstrate the ability of bead beating to extract sufficient quantity and quality of DNA for sequencing, they do not indicate whether the retrieved DNA accurately reflects the members of the bacterial community. More recent studies have compared variation in bead beating extraction methods as well as commercial kits in 16S rRNA gene amplicon sequencing results. Yuan *et. al.* compared six different extraction methods and evaluated resulting DNA yield, shearing, representation of microbial diversity, and reproducibility [163]. A wide range of DNA extraction methods are used in microbiome studies and the airway studies in Table 2.1 are no exception. Yuan *et. al.* recognized this and set out to compare the more common variations in DNA extraction strategies. While most studies evaluate the effectiveness of

the method by DNA yield and quality, Yuan *et. al.* created specific communities with known quantities of bacteria so they could calculate expected abundance after 16S rRNA gene amplicon sequencing to determine the ability of each method to accurately reflect the diversity of the community. This is an important measure since the goal of microbiome studies is to draw conclusions about biology based on the composition of microbial communities. Confirming the work of Yeates *et. al.* and de Boer *et. al.*, bead beating was found to yield significantly higher quantity and quality of DNA [163]. However, higher yield of DNA did not correlate with better representation of the original microbial community. Instead, bead beating in combination with enzymatic digestion by a mixture of mutanolsyin, lysozyme, and lysostaphin gave the best representation of the microbial community regardless of quantity, implying that a higher yield of DNA does not indicate a superior extraction method. Each of these enzymes cleaves the bacterial cell wall through different mechanisms. Lysozyme cleaves the glycosidic bond between N-acetyl-glucosamine and N-acetyl-muramic acid while mutanolysin does it here when peptidoglycan is O-acetylated. Lysostaphin cleaves the pentaglycine cross-link in the peptidoglycan cell wall of staphylococci. Differences in the peptidoglycan layer of both Gram positive and Gram negative bacteria make them variably sensitive to each of these enzymes so using a mixture of each lyses a more diverse range of bacteria. For low-density airway samples, an effective, reproducible method is critical to accurate representation of the original community. Evaluation of extraction methods on upper airway human samples [164] and BAL from lower airway pediatric samples [165] demonstrated more significant variation in microbial community composition by extraction method than by technical replicate. Further, low-abundance airway samples are

more prone to influence by contaminating microbial sequences known to be present in kit reagents, making reagent controls crucial for these samples. Of the studies in Table 2.1, only Morris *et. al.* sequenced a reagent control in order to detect reagent contaminants [58]. However, reagent contamination may not contribute significant bias to the results, even in low-abundance samples [166]. Each study employed bead beating for DNA extraction but no two studies used the same method. Although comparison of samples extracted by the same method may give useful results, comparison between the studies in Table 2.1 is not valid due to the demonstrated variability among extraction methods.

Bacterial load in microbiome samples must be quantified independently of sequencing, as sequencing can only quantify relative abundance of bacteria. It is also useful to confirm the presence of bacterial DNA prior to amplification and sequencing, as both PCR and sequencing errors in low-abundance samples may give false positive results if there is little to no DNA present. Inclusion of negative and reagent controls in sequencing runs can help detect this error. Methods such as qPCR, which is another relative quantification method, give a better estimate of the initial load of bacteria present in a sample. Though useful, this step is not necessary, which is reflected in only two studies in Table 2.1 employing it. Nevertheless, quantification by qPCR must be performed cautiously, as underlying differences among bacteria can lead to incorrect quantification. Most universal quantification methods, similar to amplicon sequencing, rely on the bacterial 16S rRNA gene. Primers are designed which anneal to conserved regions within this gene in all bacteria present, allowing, in theory, amplification and quantification of all bacterial DNA in the sample. In reality, each primer set contains bias that amplifies certain bacteria preferentially over others [167]. The most commonly used

universal primer set was designed by Nadkarni *et. al.* [168] and has good coverage of most bacterial taxa but does not amplify those in the phyla Spirochetes or Chlamydiae well. Though its common use may indicate that those studies employing this primer set may be compared, there are other complications that prevent this. The bacterial 16S rRNA gene is known to vary in copy number among bacterial species, which means that uncorrected quantification data will over-represent species with more copies and under-represent those with fewer [169]. Methods to correct for this have been developed for sequencing data [170,171] but doing so for qPCR data is more difficult since the community composition is unknown. Sequencing data could be used to elucidate bacterial community structure and qPCR data could be corrected retrospectively, but this could introduce more bias due to primer bias, PCR and sequencing error, and variation in OTU generation and identification. Use of other genes for universal quantification has been explored, such as the single copy rpoB gene, which provides better species-level resolution over the 16S rRNA gene [172,173]. Despite these advantages, the 16S rRNA gene remains dominant in the field for amplicon sequencing, likely due to the high number of resources already available for its use. However, the rpoB gene may hold promise for increased accuracy in quantification of bacterial load.

Though use of the 16S rRNA gene is ubiquitous in microbiome studies, the best variable region for amplicon sequencing is still up for debate [22]. This is clearly demonstrated in Table 2.1, in which studies use multiple regions in the V1-V5 range. This further makes direct comparison between the studies impossible, since the primer sets that target these varying regions will be biased against different bacterial taxa. Use of a standard set of primers targeting the same region would eliminate this issue, but does

not account for variability in detection of specific bacteria. The use of an appropriate variable region that detects all bacteria present in a community and reflects their true abundance is an on-going challenge in the microbiome field that is actively being investigated.

Though all the studies in Table 2.1 identify bacterial communities in the airways in various disease states, the heterogeneity of methods used in detecting these communities make them impossible to compare directly, and may imply that none accurately reflect the true community composition. Continued standardization and optimization of amplicon sequencing methods will improve these issues in the future, but make current metagenomics studies questionable, particularly for low-abundance airway communities. Despite these challenges, comparison of community changes in disease states can give valuable insight into host-microbe and microbe-microbe interactions that, if validated experimentally, may generate novel research questions and lead to new therapeutic targets. Metagenomic studies in combination with experimental validation provide a powerful set of tools to explore the roles of airway microbial communities in disease and health that may lead to improvements in treatment, prevention, and detection strategies.

2.2 Burn and Inhalation Injury

Research in the airway microbiome field has focused on changes in microbiota during and after disease as well as determination of existence of a lower airway microbiota at homeostasis. Few studies have examined changes in airway microbiota after injury and none have examined the microbiota following burn and inhalation injury.

Treatment of burn injury has improved markedly over the past 35 years, resulting in a 78% decrease in mortality since the 1970s [174]. This is illustrated in the mortality rate associated with total body surface area (TBSA) burn, where 30% TBSA in the 1970s resulted in a 50% mortality rate while 80% TBSA results in the same mortality rate today [174,175]. This is likely due to improvement in initial shock resuscitation, airway management, nutrition, wound care, and infection control [174]. Burn injury is defined by the type of injury, burn depth, TBSA, and injury severity [176]. Type of injury includes thermal, electrical, chemical, and radiation. In between the years 2005 and 2014, 42.6% of burn injuries reported to the American Burn Association were caused by fire or flame, while scalding accounted for 34.0%, making thermal injury the most common type [175]. Burn depth is measured by the layer of skin the injury penetrates through and ranges from first to fourth degrees [176,177]. First and second degree burns are generally superficial and heal with minor intervention, while third and fourth require more care. Increasing TBSA is associated with increased mortality rates. Though burns with %TBSA of less than ten compromised over 75% of burns within the years 2005 – 2014, patients with 70% to 80% TBSA had a 55% mortality rate while those with greater than 90% TBSA had an 85% mortality rate [175]. Burn severity encompasses several metrics and is used to triage patients into minor, moderate, and major categories. TBSA, burn depth, patient age, location, injury type, and pre-existing conditions all factor into severity. Burns categorized as severe with greater than 20% TBSA are associated with systemic changes similar to those seen in trauma and surgical patients [177]. Patients experience decreases in metabolic function and cardiac output within the first 48 hours of burn injury which gradually increase to become hypermetabolic within five days of injury [177]. Though

patients experience a systemic inflammatory response, global immune function is compromised, increasing susceptibility to bacterial, viral, and fungal infections. The top three complications burn patients experienced between 2005 and 2014 include pneumonia, cellulitis, and urinary tract infections, with respiratory failure coming in at a close fourth place [175]. Unfortunately, infection can be difficult to diagnose in this patient population, as the systemic changes described above make traditional infection indicators, such as elevated temperature and white blood cell count, unreliable [174,178]. The presence of inhalation injury (II) complicates matters by increasing mortality by up to 25%, while pneumonia alone increases it by 40%, and both together by 60% [179,180]. Prevention of both complications is challenging due to ineffectiveness of prophylactic antibiotic treatment and a lack of standardized procedures for diagnosing, scoring, and treating II [178,179]. Diagnosis of II is highly subjective, as it requires visual examination of damage to the airways through a bronchoscope accompanied by a history of burn injury in an enclosed space [181,182]. After smoke inhalation, soot and other particles and gases come into contact with airway epithelial cells, inducing damage and causing them to slough off [183]. Damage signals initiate a cascade of inflammatory responses that result in airway edema, bronchoconstriction, and poor gas exchange [184,185]. Combined with the immune suppression and hypermetabolic state induced by severe cutaneous burn, impairment of airway defenses by II increases susceptibility to bacterial infection and subsequent pneumonia [183]. In order to improve survival rates of patients with concurrent burn and II, improved methods for detection and identification of infecting bacteria is critical. Application of NGS methods, particularly 16S rRNA gene amplicon sequencing, to burn patient airway samples may reveal changes in airway

communities after injury that can be therapeutically targeted to prevent infection and pneumonia.

2.3 Toxic Effects of Smoke Exposure

Smoke is generated by the process of incomplete combustion, in which a fuel source burns in the presence of an insufficient oxygen supply and heat [184]. Thermal decomposition and vaporization produce a complex, heterogeneous mixture of gases and particles that can be inhaled in large concentrations in an enclosed space, such as a burning building [186]. The toxic components of smoke can be categorized as asphyxiants, respiratory irritants, or systemic toxins [184,186]. Carbon monoxide and hydrogen cyanide are asphyxiant gases common to all fires that are causes of early smoke inhalation-associated morbidity [181]. Both gases interfere with the body's ability to utilize oxygen. The iron-containing protein haemoglobin, to which oxygen binds within red blood cells for transport through the body, has 250 times the affinity for carbon monoxide than oxygen [187]. Binding of carbon monoxide displaces oxygen, reducing the oxygen-carrying capacity of the blood, and it interferes with cellular respiration by inhibiting binding of oxygen to cytochrome oxidase. Hydrogen cyanide's toxic effects are due to its ability to inhibit electron transport and cellular respiration by binding to trivalent iron in the mitochondrial a3 complex. Due to their systemic effects, both of these gases can also be classified as systemic toxins. Heavy metals inhaled from combustion of various materials are also classified as systemic toxins. However, it is the components classified as respiratory irritants that are most likely to induce II. Respiratory irritants include gases that may be inhaled alone or adhered to the surface of carbon-

containing particles, such as ammonia, acrolein, and formaldehyde. Heat injury may seem an obvious culprit in induction of II, but the efficiency of heat dissipation of the upper airways prevent heat injury from traveling much further down than past the vocal cords [188]. Irritant gases, conversely, can travel as far as the alveoli, even when adhered to particles. The presence of specific irritant gases depends on the materials burned. Acrolein, which can bind to particles, is produced by combustion of materials containing cellulose, such as wood and paper products, as well as acrylics such as wall coverings and textiles [189]. Acrolein is known to be present in cigarette smoke as well and has been shown have a pro-inflammatory effect on primary nasal epithelial cells [55]. Hydrogen chloride may be more toxic to the airways when bound to particles than in its gaseous form and is produced in large amounts during combustion of polyvinyl, which is present in floor and furniture coverings as well as wire and pipe coatings. Phosgene, also produced by burning of polyvinyl, is a strong irritant that injures the small airways and alveoli. Aldehydes, free radicals, ammonia, and polycyclic aromatic hydrocarbons, among others, all of which may contribute to II, are also produced in varying amounts during a house fire [184,189].

Regardless of the composition of the inhaled smoke, II results from the interaction of the gases and particles with the airway mucosa and lung parenchyma, inducing an inflammatory cascade that results in airway injury and pulmonary edema [187]. Specifically, damage to the airways by irritant gases causes production of inflammatory mediators such as interleukin-1 [185]. These inflammatory mediators induce the complement cascade, draw neutrophils and macrophages to the airways, and activate fibrinogen. Activation of the complement cascade may lead to cellular dysfunction, while

inducible nitric oxide synthase by macrophages and neutrophils contributes to pulmonary edema and airway casts. The coagulation cascade, which activates fibrinogen, also results in formation of airway casts. Each of these responses decreases the ability of the airways to oxygenate effectively and inhibits airway defenses and immune responses, increasing susceptibility to infection. Treatment involves mechanical ventilation to assist oxygenation and supportive care, including washing the airways with saline to remove soot and toxic particles.

2.4 Immune Response and Infection Risk

As mentioned above, severe burn injury with greater than 20% TBSA results in traumatic injury with systemic effects, including hypermetabolic changes and a massive inflammatory response [177]. Recognition of this inflammatory response, particularly in the septic patient, and the need to identify it early led to defining the systemic inflammatory response syndrome (SIRS) in 1991 [190]. SIRS is characterized by a massive pro-inflammatory response after traumatic injury that can inhibit patient recovery if not appropriately addressed. The compensatory anti-inflammatory response syndrome (CARS) is thought to directly follow SIRS, and consists of inhibition of immune responses in order to restore homeostasis [191]. Recent studies suggest that SIRS and CARS may occur simultaneously rather than successively [192]. Blunt trauma or burn injury was found to alter 80% of the leukocyte transcriptome, including activation of inflammatory mediators and genes involved in pattern recognition and antimicrobial functions, and suppression of antigen presentation and T and NK cell function [192]. These genome-wide changes were induced regardless of type of injury, suggesting a

common immune response to traumatic injury. Further, induction of both pro- and anti-inflammatory responses simultaneously implies that SIRS and CARS can occur together, rather than one after the other.

Dysregulation of pro- and anti-inflammatory responses has also been demonstrated in II. Several studies have found blunted immune responses in the airways in burn patients with II early after injury and that the magnitude of the response is associated with the degree of injury [193,194]. Further, these studies have associated excessive amounts of the cytokine interleukin-1 receptor antagonist (IL-1Ra) specifically with immune dysfunction in II, allowing it to serve as either a biomarker of injury or a therapeutic target. Leukocytes from patients who do not survive burn and II do not produce as many immune mediators as those from patients who survive, while macrophages, though increased in number in the airways after II, also display decreased function [193,195]. The increased number of macrophages follows increases in the inflammatory cytokines tumor necrosis factor alpha (TNF-α) and interleukin-8 (IL-8), both of which induce tissue damage in the lung parenchyma. Neutrophils are also drawn to the airways in larger numbers following II and release oxygen radicals, inflammatory cytokines, and proteases that can also damage the lung parenchyma. This damage leads to increased pulmonary vascular permeability, resulting in edema that, in addition to inflammation, alters ventilation and perfusion and leads to acute lung injury (ALI) and acute respiratory distress syndrome (ARDS) [195,196]. The tissue damage and inflammation associated with ALI and ARDS increase susceptibility of the airways to infection, particularly pneumonia, which was present twice as frequently in patients with II than those without [180,195]. Bacterial pneumonia in patients with both burn and II is

associated with mortality rates as high as 68%, emphasizing the importance of early and appropriate antibiotic treatment in this patient population [183,195]. This is complicated by challenges in identifying infections early as well as the organism responsible for infection. Routine cultures of bronchial washings as well as endotracheal tubes have been used to identify infecting organisms but use of specific culture media identifies only those organisms capable of growth in those conditions rather than all organisms present. Use of NGS methods could overcome this bias, as it allows identification of a broader range of organisms and is not dependent on the organism's growth conditions. Although these methods are not yet fast enough to be clinically beneficial, they could identify the communities of organisms that play roles in the development of ALI and ARDS following burn and II in the airways, providing therapeutic targets that may improve patient outcomes.

CHAPTER 3: OPTIMIZATION OF DNA EXTRACTION AND SEQUENCING METHODS

3.1 Patient Population

The goal of this project was to use NGS methods to identify bacterial DNA present in the airways of patients with burn and inhalation injury. Through collaboration with the North Carolina Jaycee Burn Center at the University of North Carolina Hospital, a repository was created in order to store bronchial washings from burn patients. Within 24 hours of hospitalization, patients with suspected inhalation injury (II) underwent therapeutic bronchoscopy in order to flush soot and other debris from the airways. After clinical use of the sample, what remained was usually discarded. Creation of the repository allowed frozen storage of these airway washings for future studies instead. The repository was approved by the UNC Institutional Review Board (IRB) under study #10-0959. Consent for retaining samples was obtained from the patient or their legally authorized representative. After bronchoscopy, samples were placed on ice and processed within 72 hours. The washing was spun down and the pelleted cell fraction was stored separately from the supernatant at -80°C. Special permission from the IRB was obtained to use the pelleted portion of the sample to extract bacterial DNA (IRB #12-2475).

Samples were collected over a three-year period and de-identified before storage in the repository. Patient clinical and demographic data was also collected and stored in the electronic Red Cap database. Information such as patient gender, race, comorbidities,

clinical bacterial cultures, and measurement of inflammatory cytokines was collected and stored. Table 3.1 lists the information available within the database. For this study, DNA was extracted from 277 samples from a total of 102 patients.

Demographics	**Injury Information**	**Bronchoscopy Clinical Data**	**Bronchoscopy Laboratory Data**	**Blood Data**
Study ID	Date Injured	Date of 1st Bronchoscopy	Date	Blood Obtained?
Date enrolled	Item First Ignited	Day Post Injury	Volume	Tube Types
Sex	Flame Spread	Ventilator Mode	HSP-70	GSTM1 Genotype
Race	Fire Location	FiO_2	Hyaluronic Acid	Serum Urea
Age at Injury	COHb	Mean Airway Pressure	HMGB-1	
Height	Tracheostomy	PaO_2	Total Cells per Milliliter	
Weight	Days Trached	FiO_2/PaO_2	Differential Cell Count	
BMI	Hospital Days	Oxygenation Index	Inflammatory Cytokines	
Comorbidities	Ventilator Days	Secretions		
Smoking Status	Discharged on Ventilator?	Soot		
	Cause of Death	Mucosa Condition		
		X-Ray Results		
		Bacterial, Viral, and Fungal Culture Results		

Table 3.1: Patient Demographics and Clinical Data Collected. BMI = Body mass index, COHb = carboxyhemoglobin levels, FiO_2 = fraction of inspired oxygen, PaO_2 = partial pressure of arterial oxygen, HSP-70 = Heat shock protein-70, HMGB-1 = high mobility group box 1 protein

Besides patient demographics, information recorded in the Red Cap database included general injury information as well as biomarkers and clinical results specific to airway injury. General information on how the injury occurred, as listed under 'Injury Information' in Table 3.1, included the type of injury (flame, scald, electrical, etc.), the

material first ignited, the location of the fire (house, office building, trailer, work site, etc.), and how the fire spread. Specific clinical parameters included whether a tracheostomy was performed for mechanical ventilation, the number of days the patient spent on a ventilator, the number of days spent in the hospital, as well as levels of carboxyhemoglobin (COHb), which indicate carbon monoxide poisoning, and cause of death, if the patient died. Clinical information specific to the bronchoscopy included how long after the injury it was performed, the ability of the airways to oxygenate and ventilate, the physical appearance of the airways during bronchoscopy (sloughing of epithelial cells, secretions, presence of soot, etc.), x-ray results indicating edema and/or pneumonia in the lungs, and the results of clinical cultures. After clinical use of the bronchoscopy samples, specific biomarkers were assayed for research purposes, including HSP-70, HMGB-1, hyaluronic acid, and pro- and anti-inflammatory cytokines. The samples were also analyzed for the number and type of cells present. HSP-70, hyaluronic acid, and HMGB-1 all play roles as damage-associated molecular patterns (DAMPs) which can activate innate and adaptive immune responses in the airways [197]. HSP-70 is an important heat shock protein released in the cytoplasm following thermal or chemical injury in order to protect protein integrity and prevent cytotoxicity [198]. HMGB-1 plays important roles in assembly of nucleoprotein complexes in the nucleus but induces pro-inflammatory responses when in extracellular space [197]. Hyaluronic acid is derived from damage to the extracellular matrix and induces inflammatory responses. All three of these molecules activate TLR2 and TLR4. A panel of cytokines was selected for measurement to represent inflammatory and immune responses as well as tissue damage and repair [199]. Blood samples were taken in order to determine the

patient's glutathione S-transferase Mu 1 (GSTM1) genotype and serum levels of urea. GSTM1 is part of the glutathione S-transferase family and protects the lungs from oxidative damage [200]. It has a null polymorphism that is present in a median 50% of the population and is associated with increased risk of inflammatory lung disease. Serum urea has been associated with mortality and is an indicator of kidney function [201,202]. Together, this data enables a range of observational studies in relation to the airway microbiota identified within these samples.

Recent airway microbiome studies indicate the possibility of a low biomass yet diverse microbiota in the lower airways of healthy individuals [203–205]. We collected bronchial washings from healthy volunteers in order to determine if we could replicate these results. Collection of these samples was previously approved by the UNC IRB [206] and their use in this study was approved in IRB #12-2475.

3.2 Challenges in Extraction of Bacterial DNA from Bronchial Washings of Burn Victims

Methods for extraction of DNA from the bronchial washing samples in the repository were developed. Prior to this study, no specific method existed for extraction of bacterial DNA from burn patient bronchial washings samples. Increased interest in gut and soil microbiota over airway microbiota has led to development of extraction methods specific to these environments. Airway samples from the nasopharyngeal region contain approximately five to seven orders of magnitude fewer bacteria than do gut samples [164]. This low abundance of bacteria makes use of efficient extraction methods and reagent controls crucial to ensuring accurate representation of the original airway

community. Bacterial biomass within the respiratory tract is highest in the nasopharyngeal region and decreases from the upper respiratory tract down through the lower respiratory tract (from trachea to bronchi to alveoli) [134]. In bronchial washings from burn victims, the presence of soot, sloughed epithelial cells, blood, and mucous posed additional challenges in ensuring efficient extraction of bacterial DNA. Due to the possibility of bacteria adhering to any of these contaminants, separation by centrifugation was not considered. Use of dithiothreitol (DTT), which reduces disulfide bonds and maintains thiol groups, was considered for reducing mucous viscosity for increased extraction efficiency of bacteria adhered to it. DTT at a concentration of 0.1% has been shown to increase reproducibility of cell counts in sputum samples from patients with CF [207], and it has been used for detection by culture of anaerobic bacteria in CF patient sputum [208] as well as extraction of DNA for 16S rRNA gene amplicon sequencing of CF patients [148]. To compensate for overabundance of human DNA as compared to bacterial DNA, methods for enrichment of bacterial DNA were considered. Specifically, the New England BioLab's NEBNext kit was tested with burn patient samples. This kit depends on differential methylation of human and bacterial DNA, and employs a methyl-CpG binding domain and magnetic beads in order to separate human DNA from bacterial DNA, which has a higher percentage of methylation at adenine nucleotides [209]. Mechanical, chemical, and enzymatic lysis methods were tested in order to determine which combination of techniques resulted in the highest extraction efficiency. The most appropriate method was determined based on the quality and quantity of DNA extracted. Appropriateness of these methods for this sample type could be further confirmed

through 16S rRNA gene amplicon sequencing in order to determinate how well they reflect the structure of the original bacterial community.

3.3 DNA Extraction Methods

Commercial bacterial DNA extraction kits from MP Biomedicals, Mo Bio, and Qiagen were compared with extraction using phenol:chloroform:isoamyl alcohol (PCI). Each method incorporates some form of physical lysis (through homogenization or vortexing) as well as subsequent chemical and enzymatic lysis steps.

Dr. Scott Plevy's group at the University of North Carolina at Chapel Hill designed the PCI protocol used here. Prior to use of this protocol, the Qiagen DNeasy kit was used according to the manufacturer's instructions but gave poor quality and quantity DNA from nasal lavage samples. This method was not specifically designed for microbiome research but Dr. Plevy's PCI method was. The PCI method employs enzymatic lysis with lysozyme and physical lysis by homogenization with 0.1mm silica beads followed by additional chemical lysis with SDS and isolation of DNA using PCI. The Qiagen DNeasy kit is then used to further purify the precipitated DNA. Homogenization with bead beating and through a syringe was tested with this method. Bead beating consistently gave four to ten-fold higher DNA quantity with comparable quality. Digestion using lysozyme and lysis with SDS was tested using the Gram positive organism *Bacillus subtilis.* Gram positive organisms contain a thick cell wall composed of peptidoglycan, while Gram negative organisms contain a much thinner layer [210]. This thick cell wall makes lysis of Gram positive organisms difficult and necessitates lysis methods in addition to detergent-based techniques. Lysozyme is an antimicrobial

factor that is abundant in the airways and cleaves the glycosidic linkage between N-acetylglucosamine and N-acetyl muramic acid within peptidoglycan, effectively inducing cell lysis in Gram positive bacteria [211,212]. SDS is a negatively charged amphipathic detergent composed of exposed hydrophilic heads and hydrophobic tails tucked within [213]. This bipolar nature allows it to disrupt the phospholipid bilayer of cell membranes, resulting in lipid-detergent micelles in solution with hydrophilic heads pointing out and hydrophobic tails pointing in. The necessity of both SDS and lysozyme was tested with *B. subtilis* samples. $1.4x10^9$ *B. subtilis* cells per mL were placed into six tubes containing silica beads. These were treated either with or without lysozyme and SDS and incubated at 37°C for 5, 15, and 30 minutes. Prior to DNA extraction, treated bacteria were placed on slides and Gram stains performed to visually confirm cell wall digestion. Gram positive *B. subtilis* was expected to appear purple if the cell wall was intact and pink if not. Figure 3.1 is a representative picture of *B. subtilis* treated with 1% SDS and lysozyme incubated for 30 minutes and Figure 3.2 is of *B. subtilis* without SDS or

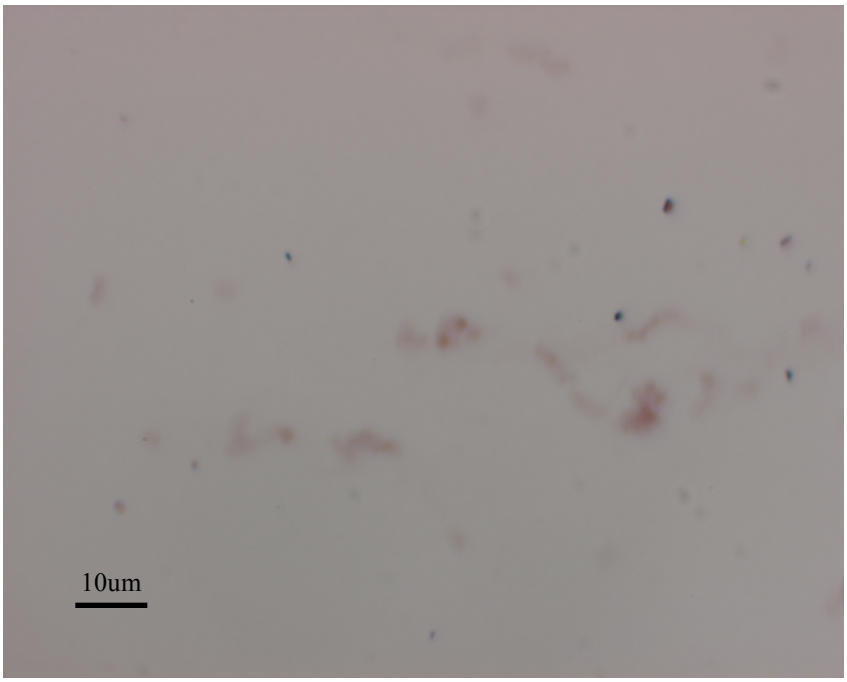

Figure 3.1: *B. subtilis* Gram Stain after SDS and Lysozyme Treatment. *B. subtilis* was treated with 1% SDS and lysozyme and incubated at 37°C for 30 minutes. Though some purple can be seen, the majority of the cells appear pink, indicating efficient lysis of the Gram positive cell wall.

lysozyme treatment. The majority of *B. subtilis* cells treated with 1% SDS and incubated at 37°C for 30 minutes after addition of lysozyme are pink, indicating efficient cell wall lysis. In Figure 3.2, lack of SDS and lysozyme treatment does not result in cell lysis, even after bead beating, as evidenced by the presence of mostly purple cells.

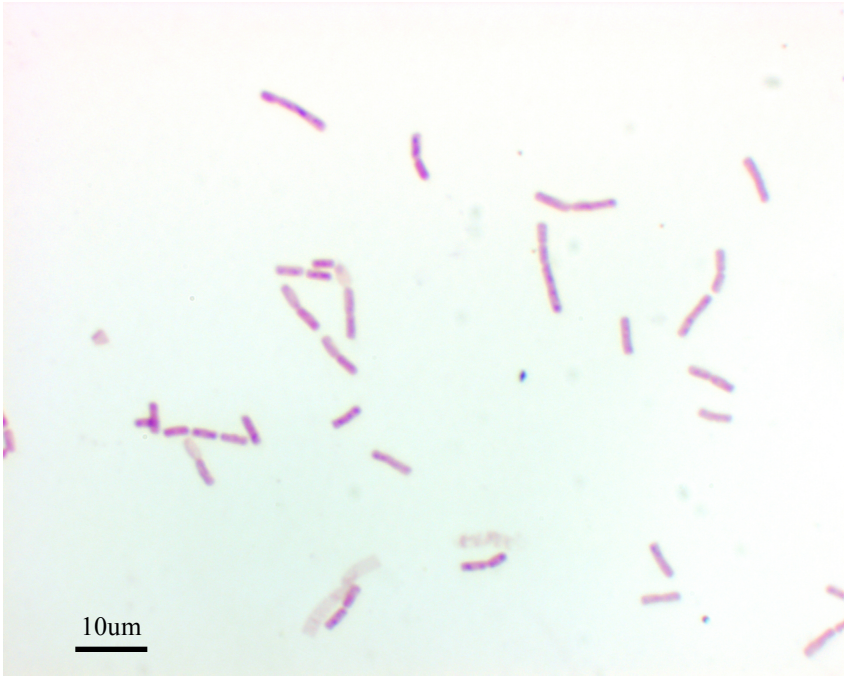

Figure 3.2: *B. subtilis* Gram Stain without SDS or Lysozyme Treatment. *B. subtilis* was not treated with SDS or lysozyme prior to physical bead beating. The resulting Gram stain shows mostly purple cells, indicating an intact cell wall.

Based on these results, physical bead beating in combination with lysozyme enzymatic digestion and SDS chemical lysis was included in the method prior to DNA extraction.

To address increased mucous viscosity in the burn patient samples and possible inaccessibility of bacteria within it, DTT was tested in conjuncture with the PCI extraction method. The PCI method was used to extract bacterial DNA from nasal lavage samples and the Sputolusin method (EMD Millipore) was used to treat the samples with DTT before extraction. Two samples from the same individual were taken for a total of four individual subjects and eight samples. One sample per subject was treated with DTT and the other sample was not. For two of the subjects, DTT treatment resulted in

increased DNA quantity but for the other two it did not seem to make a difference (Table 3.2). Overall, DNA extraction quantity was variable, leading to questions about the effectiveness of the extraction method.

Sample	Treatment	DNA (ng/ul)	260/280 Ratio
1a	Sputolysin	242.3	1.86
1b	No sputolysin	252	1.82
2a	Sputolysin	12.4	1.56
2b	No sputolysin	2.6	0.82
3a	Sputolysin	30.7	1.59
3b	No sputolysin	33.7	1.64
4a	Sputolysin	7.5	1.9
4b	No sputolysin	1.1	0.46
Control-a	Sputolysin	3.2	0.87
Control-b	No sputolysin	8.6	1.39

Table 3.2: DNA Quantity and Quality after Phenol:Chloroform:Isoamyl Alcohol Extraction Prior to DTT Treatment. Sputolysin is the commercial name for DTT.

Despite optimization of cell lysis methods, the phenol:chloroform:isoamyl alcohol procedure was inconsistent in quality and quantity of DNA extracted and was prone to contamination with bacterial DNA from the environment and the low amounts present in reagents. To overcome these issues, standardized commercial kits designed for microbiome samples were tested. Three kits were compared, including MP Biomedicals' FastPrep DNA kit, Mo Bio's PowerFecal kit, and Qiagen's UCP Mini Pathogen kit. For the FastPrep kit, lysing matrix A, which included garnet and zirconium beads, was used for physical lysis. For the PowerFecal kit, 0.7mm garnet beads were used, and the Qiagen kit uses a proprietary mix of beads that appear similar to silica beads. Initial tests of the PowerFecal kit gave very low DNA quantities as compared to the phenol:chloroform:isoamyl alcohol method, so this kit was not tested further. The Gram positive bacteria *Staphylococcus aureus* was used to compare extraction efficiency of the Qiagen and MP Bio kits. Samples were pre-treated with lysozyme, mechanically lysed

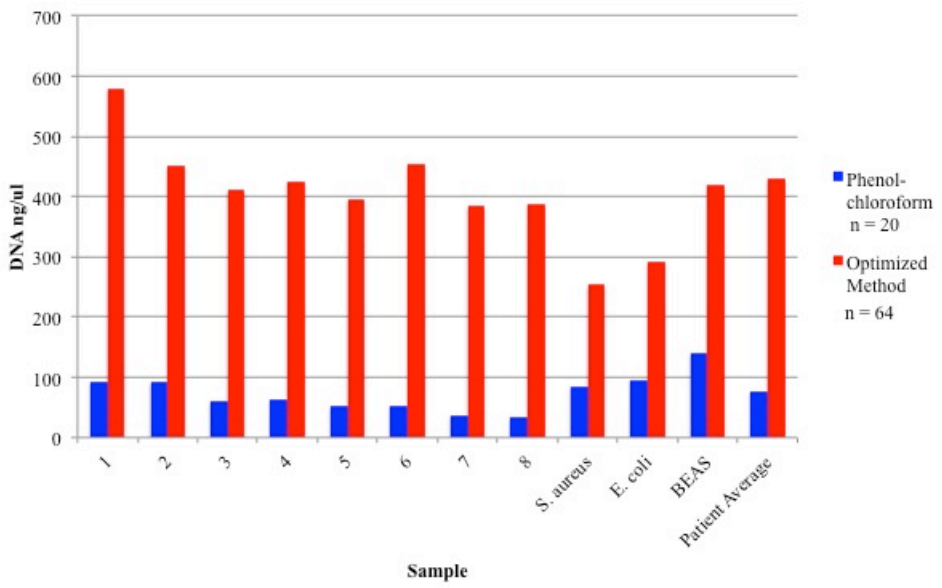

Figure 3.3: Quantity of Extracted DNA. Burn patient samples and bacterial controls were extracted using the phenol:chloroform:isoamyl alcohol method and the optimized protocol.

using the beads that came with each kit, and treated with 1% SDS. The same initial quantity of *S. aureus* resulted in six times more DNA when extracted with the Qiagen kit. This DNA was also of better quality. Addition of RNase A resulted in better quality DNA and additional tests demonstrated consistent quantity and quality with this kit. Figure 3.3 compares DNA quantity using this optimized extraction protocol to that obtained using the phenol:chloroform:isoamyl alcohol method. Figure 3.4 compares the

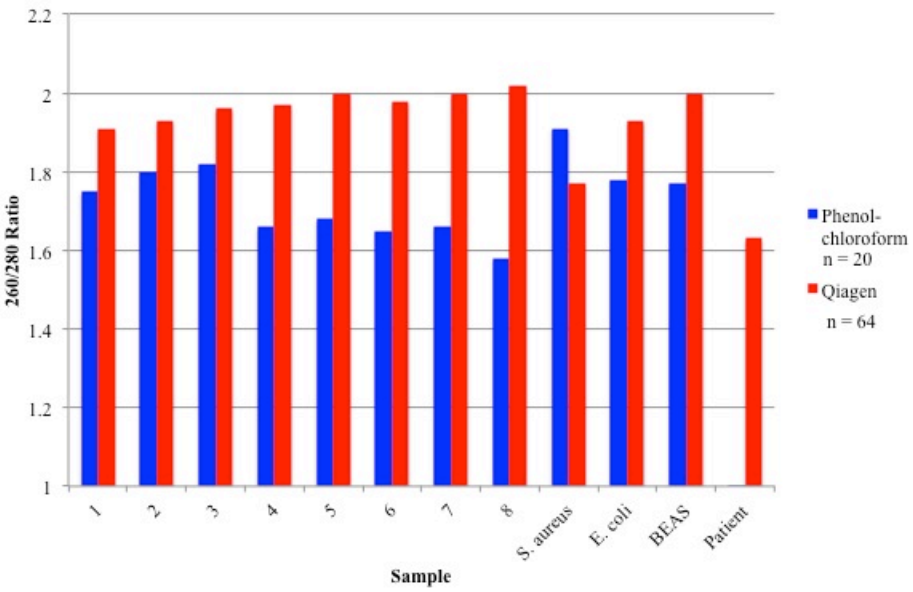

Figure 3.4: Quality of Extracted DNA. Burn patient samples and bacterial controls were extracted using the phenol:chloroform:isoamyl alcohol method and the optimized protocol.

quality of DNA obtained from both of these methods. The final, optimized method used to extract bacterial DNA from burn patient bronchial washings consisted of enzymatic lysis by lysozyme, physical lysis through bead beating with a vortex, chemical lysis by SDS, and DNA purification using the spin-column based Qiagen UCP Mini Pathogen kit. The protocol can be found in Appendix 1.

Overabundance of human DNA was a concern in the burn patient samples. Although the sequencing primers target the bacterial 16S rRNA gene specifically, excess human DNA may interfere with efficient amplification and sequencing. To enrich the bacterial DNA present in the samples and remove as much human DNA as possible, the New England BioLabs NEBNext kit was tested. This kit makes use of differences in

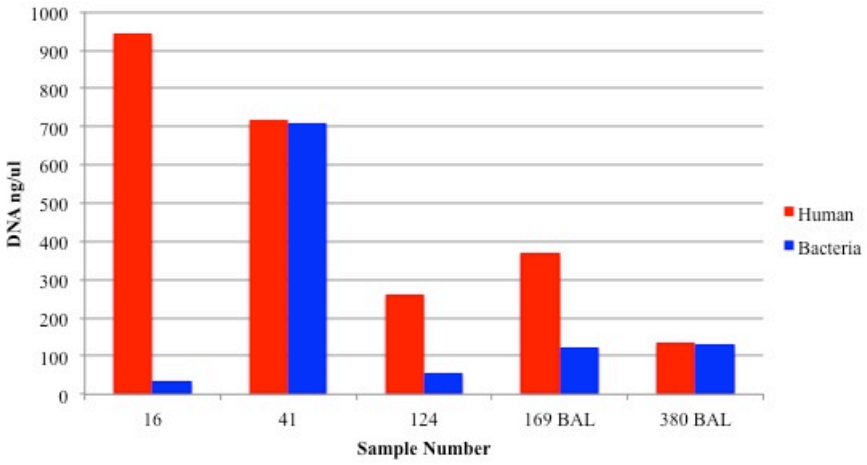

Figure 3.5: Quantity of Human and Bacterial DNA Before Enrichment. Bacterial and human DNA was quantified prior to enrichment to determine loss of DNA due to the method.

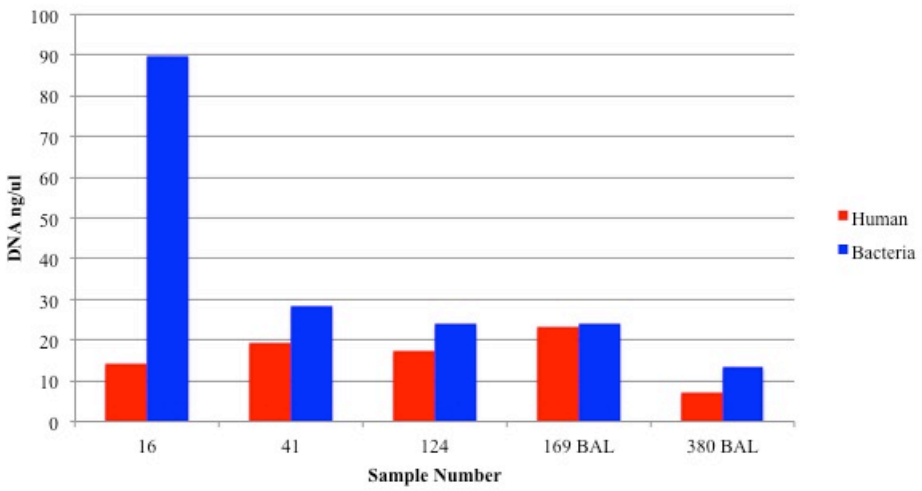

Figure 3.6: Quantity of Human and Bacterial DNA After Enrichment. Human and bacterial DNA present in patient samples was quantified after enrichment.

methylation location between human and bacterial DNA [209]. A CpG-binding domain fused to the constant region of immunoglobulin G (IgG) is used to specifically bind human DNA. This protein is bound to magnetic beads prior to introduction of DNA, allowing removal of CpG-methylated DNA and leaving adenine-methylated bacterial DNA behind. Figures 3.5 and 3.6 compare human and bacterial DNA quantity before (Figure 3.5) and after (Figure 3.6) enrichment using the NEBNext kit. Although the kit successfully removes human DNA, leaving behind a majority of

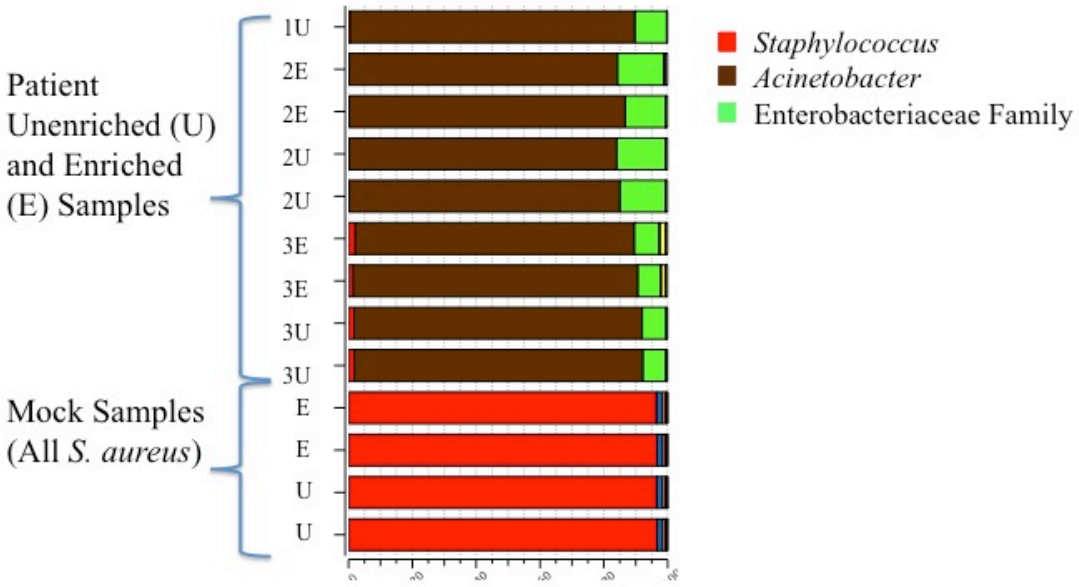

Figure 3.7: Enrichment Does Not Alter Bacterial Community Composition After 16S rRNA Gene Amplicon Sequencing. Although enrichment results in a 10-fold loss of bacterial DNA within the patient samples, it does not significantly alter community composition. Due to low abundance of bacterial DNA in some patient samples, enrichment was not used.

bacterial DNA, it depletes bacterial DNA quantity by more than one full order of magnitude. In order to determine whether this loss of DNA quantity impacted the bacterial composition, both unenriched and enriched samples were sequenced using 16S rRNA amplicon sequencing. Mock samples containing only *S. aureus* were included as controls. Figure 3.7 shows the sequencing results as the bacterial community composition of each sample normalized to 100%. Samples from the same patient did not display significant differences in community composition based on enrichment of samples. Since unenriched samples contained more bacterial DNA, which will increase sequencing accuracy, and they are not significantly different from enriched samples, the enrichment method was not used in the final extraction method prior to sequencing.

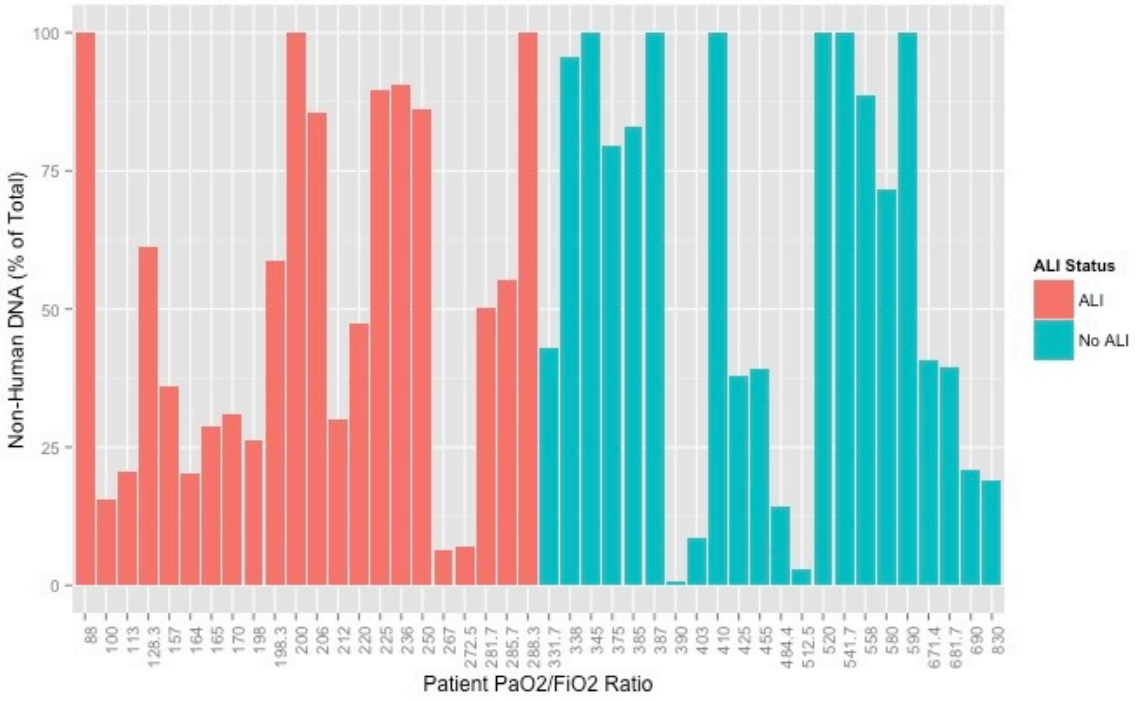

Figure 3.8: Non-Human DNA in Burn Patient Bronchial Washings. Non-human DNA was indirectly quantified by subtracting total DNA from human DNA. This method quantifies bacterial as well as viral and fungal DNA but avoids bias associated with direct bacterial DNA quantification.

3.4 Quantification Methods

Quantification of bacterial DNA after extraction and prior to sequencing can measure total bacterial load as well as specific bacterial species of interest. 16S rRNA gene amplicon sequencing can only quantify relative abundance of bacteria, making additional methods necessary for more accurate measures. Universal primers targeting the bacterial 16S rRNA gene are widely used with qPCR to quantify bacterial load. Nadkarni *et. al.* has developed a set that is commonly used and successfully detects a majority of bacteria present [168]. However, at the phylum level, this primer set does not detect any Chlamydiae and misses the majority of Spirochetes. A set designed by Maeda

et. al. successfully detects 44% of Spirochetes but also misses all Chalmydiae [214]. Measuring primer coverage at the phylum level does not take into account additional variation at lower taxonomic levels, such as genus and species. It is likely that coverage is even worse at these levels, which means these primers do not accurately measure total bacterial load. Further, it is well known that copy numbers of bacterial 16S rRNA genes vary between a single to as many as fifteen copies among different species [169]. If this variation is not corrected for it results in overrepresentation of species with higher copy numbers and underrepresentation of those with lower. Although algorithms have been developed to correct for this, they all depend on knowledge of the bacterial community composition and are dependent on sequencing results. A non-biased method, independent of sequencing, is needed to accurately quantify total bacterial load.

To address this challenge, we indirectly quantified total bacterial load by qPCR quantification of human DNA and total sample DNA using the double-stranded DNA dye PicoGreen. Subtraction of the quantity of human DNA from total DNA in the sample gives the quantity of non-human DNA and is not biased by primers, bacterial 16S rRNA gene copy number, or sequencing results. Clearly, this method is limited in that it will quantify viral and fungal DNA in addition to bacterial and it depends on two distinct methods of quantification. However, we find increased consistency in quantification with this method as compared to quantification of the 16S rRNA gene. Figure 3.8 shows the percent of non-human DNA present in bronchial washing samples from burn patients taken within 72 hours of burn and inhalation injury. These samples are divided based on whether the patient had hypoxia as indicated by their PaO_2/FiO_2 ratio.

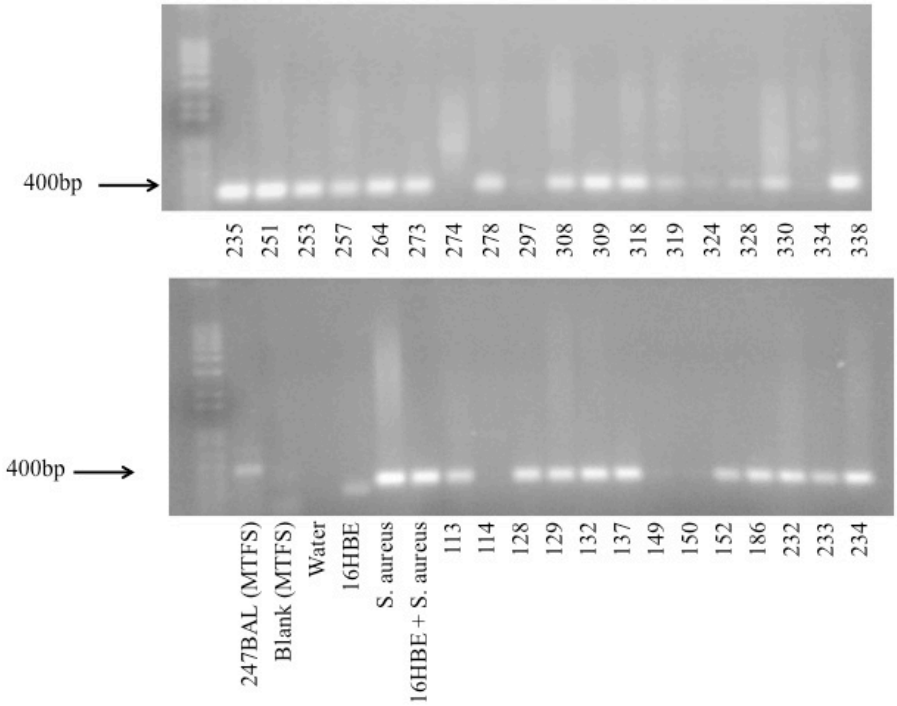

Figure 3.9: Detection of Bacterial DNA with Universal Primers. DNA from patients was PCR amplified using universal primers designed by Maeda *et. al.* Expected product size was slightly less than 400bp. Lanes are labeled by patient ID numbers.

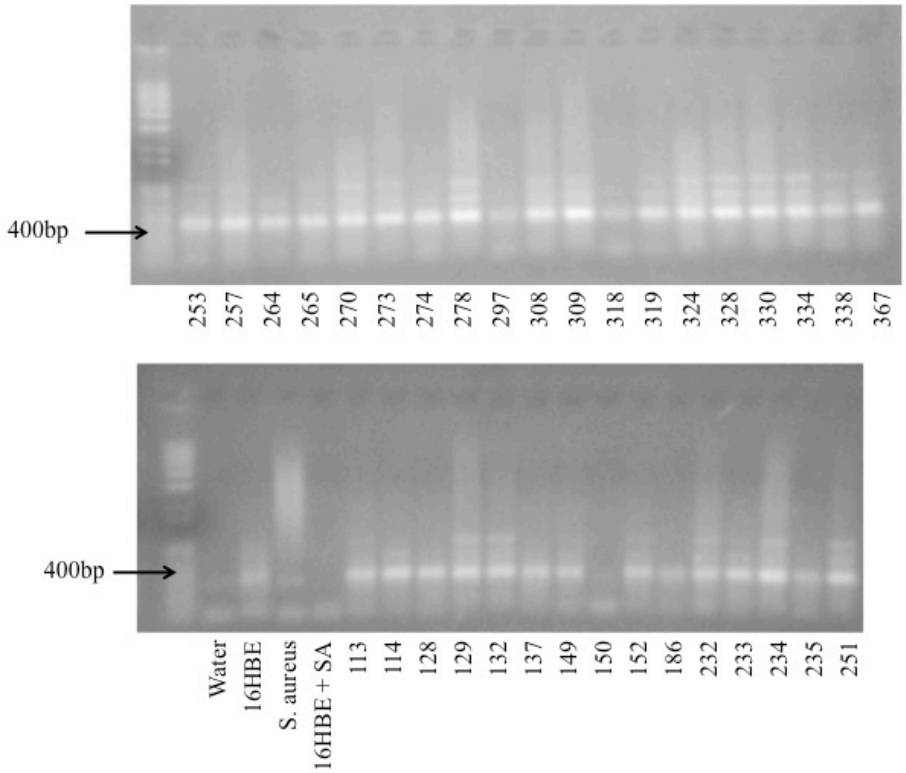

Figure 3.10: Detection of Bacterial DNA with MTFS Sequencing Primers. DNA from patients was PCR amplified using the MTFS sequencing primers. Expected bacterial product size was less than 400 bp. Lanes are labeled by patient ID numbers.

Although the indirect quantification method removes bias associated with direct quantification of bacterial DNA, direct detection of it was necessary to ensure its presence before sequencing. This was done through PCR amplification using a set of universal primers and the primers to be used for sequencing followed by gel electrophoresis. Figure 3.9 shows detection of bacterial DNA using the universal primer set designed by Maeda *et. al.* and Figure 3.10 shows detection using the sequencing primer set (Molecule tagging frame shifting; MTFS). Both PCR products were slightly less than 400 base pair long, as indicated in both figures. Negative and positive controls were included, as well as a mock sample containing both human (16HBE) and bacterial

(SA or *S. aureus*) DNA. The MTFS sequencing primers display brighter bands in some patient samples that appear lighter in the Maeda set (Figure 3.10), such as for patients 297, 324, and 328, possibly indicating higher sensitivity of the MTFS set. The Maeda set does not detect any DNA in the negative water control while the MTFS set shows a faint but small band. This may indicate contamination of the water with very low-abundance bacterial DNA that the Maeda set is not sensitive to. The Maeda set shows a brighter band for both the positive *S. aureus* control and the mock sample with both human and *S. aureus* DNA, which may indicate that it detects staphylococci better than the MTFS set. Both detect a product in the human DNA negative control, but the MTFS set detects a product similar in size to that in the water control for the mock sample. This could indicate that human DNA present in the samples will not interfere with amplification of bacterial DNA by the MTFS primer set. Low-level contamination is common in microbiome samples, especially since reagents may contain small amounts of bacterial DNA. To account for this, control samples that include reagents and water, human DNA, bacterial DNA, and human and bacterial DNA together will be sequenced with the patient samples. Detection of bacterial DNA by the MTFS sequencing primer set ensures the presence of template for 16S rRNA gene sequencing.

Bacterial DNA in control samples from healthy volunteers was quantified using the Maeda *et. al.* primer set (Figure 3.11). The positive bacterial controls show that the primers were able to quantify bacterial DNA. However, the healthy samples did not have significantly more DNA than the human-only (16HBE) control. Due to this low quantity of DNA, these samples were not submitted for sequencing.

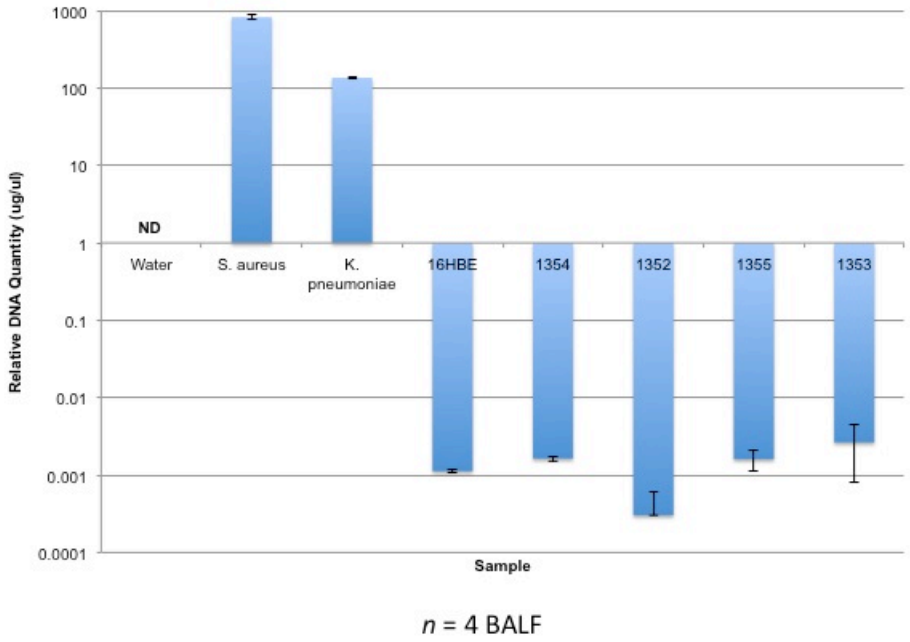

Figure 3.11: Quantification of Bacterial DNA in Healthy Lower Airways. Bronchoalveolar lavage fluid (BALF) from healthy individuals was extracted using the optimized method described above and quantified using primers designed by Maeda *et. al.*

3.5 Molecule Tagging Method

The use of barcodes for DNA template from the same sample has become common among 16S rRNA gene sequencing projects [22]. This allows pooling of multiple samples into a single lane, decreasing sequencing time. An addition to this, called molecule tagging (MT), was developed by Lundberg *et. al.*, in which the MTFS primer set previously mentioned allows labeling of individual DNA molecules with unique tags [215]. The MTs are added prior to PCR amplification and sequencing, and can be used to group resulting sequences with identical tags and generate a consensus sequence. This minimizes amplification and sequencing errors and has been shown to decrease the number of operational taxonomic units (OTUs) generated as compared to methods that only barcode the samples. For the MT method, the samples are barcoded as well so that samples may be pooled before sequencing on the Illumina MiSeq platform. We chose to use the MT method due to its increased accuracy in sequencing. The low

abundance of bacterial DNA present in the burn patient samples makes sequencing accuracy critical to accurate representation of the original bacterial community. Therefore, although it is more expensive than traditional barcoding, we chose to implement the MT method. Figure 3.12 shows how sequencing libraries are created with this method. In a short PCR step, MTs and barcodes are added to each original DNA molecule per sample. In a subsequent round of full PCR, primers and the required Illumina adapters are added. These libraries are checked to ensure the expected PCR product is present and then they are loaded onto the MiSeq for sequencing.

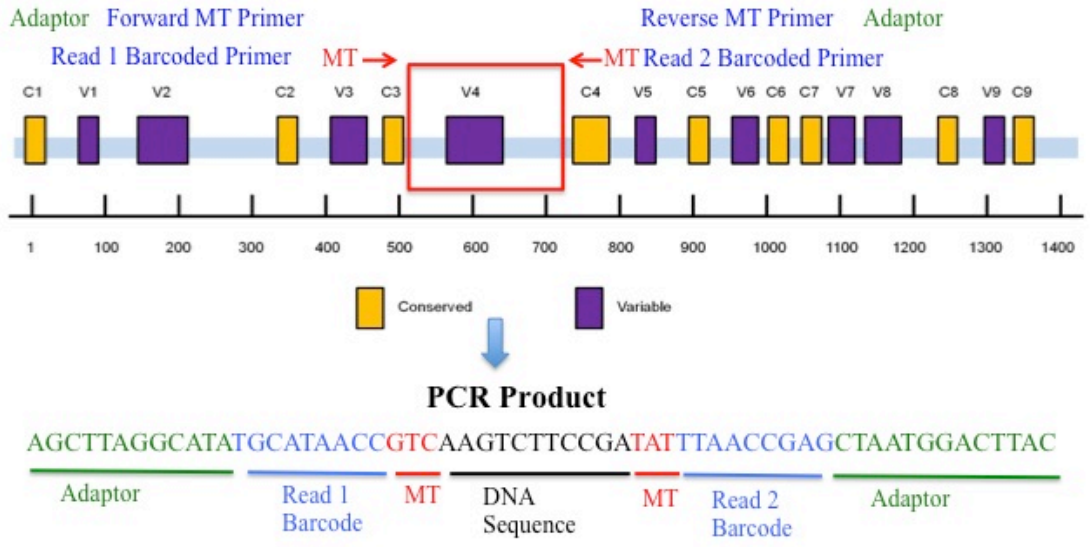

Figure 3.12: Sequencing Library Creation with the Molecule Tagging Method. MTs are attached to each DNA molecule in a brief round of PCR. A full round of PCR follows, which attaches the primers and adaptors. The resulting library is loaded onto the Illumina flow cell for sequencing on the MiSeq instrument.

After sequencing, raw reads are quality filtered using Illumina's CASAVA software [216]. The sequences can then be analyzed using MT-Toolbox, a pipeline created specifically for the MT method [217]. MT-Toolbox joins paired-end reads and generates consensus sequences from molecule-tagged DNA. OTUs are generated using

UPARSE at 97% sequence similarity [218] and 16S rRNA copy number is corrected for. OTUs are compared to the GreenGenes 16S rRNA database in order to identify them to the lowest possible taxonomic level [219]. The final output consists of sequence quality information (such as the percent of reads that merged successfully and MTs per sample) and an OTU table with patient samples in columns and bacterial taxa in rows. The number of sequences identified per OTU is represented as a count per patient. OTU counts are typically normalized to 100% since they are relative counts of the bacteria present. Further statistical analysis can be performed on the OTU table, as outlined in Figure 3.13, in order to elucidate biological roles of the microbiota. After analysis, control samples contained very low percentages of total sequences and MTs (Table 3.3). Normalized bacterial community composition is shown in Figure 3.13. The *S. aureus* (SAUR) control contains only sequences identified as *S. aureus*.

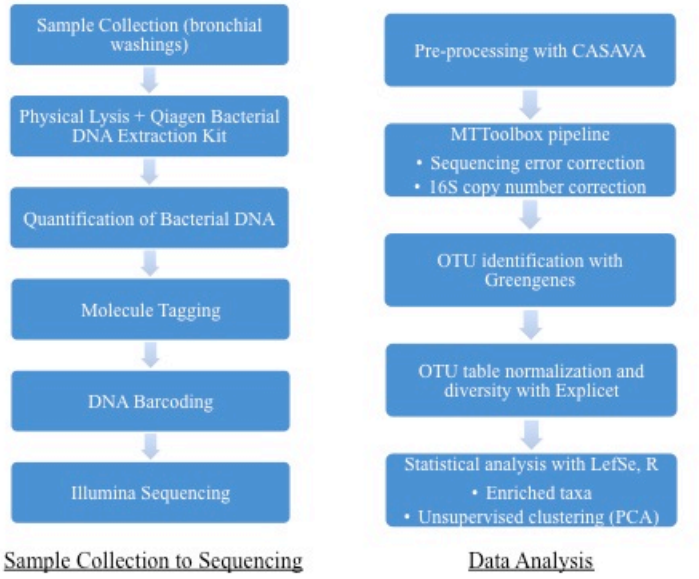

Figure 3.13: From Sample Collection to Data Analysis. Methods used for sample collection, DNA extraction, sequencing, and statistical analysis are outlined.

Control	Percent of Total Sequences	Percent of Total Molecule Tags
Human (16HBE)	1.1	0.002
Staphylococcus aureus (SAUR)	1.2	2.3
Reagent (CNTRL)	0.44	0.002

Table 3.3: Percent of Total sequences and Molecule Tags for Human (16HBE), *Staphylococcus aureus* (SAUR) and Reagent (CNTRL) Controls.

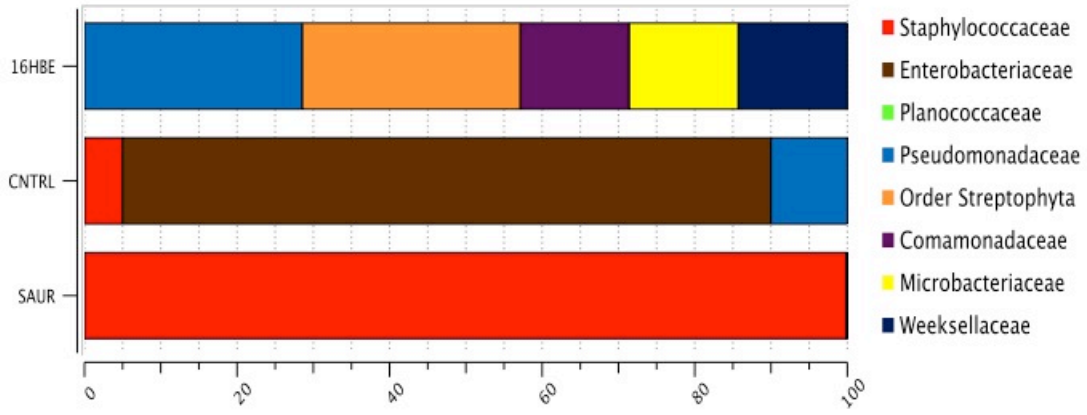

Figure 3.14: Family Level OTUs Detected Among Human (16HBE), *Staphylococcus aureus* (SAUR), and Reagent (CNTRL) DNA Controls.

The reagent control (CNTRL) contains a majority of sequences identified as Enterobacteriaceae while the human DNA control (16HBE) is split between Streptophyta, Pseudomonadaceae, Comamonadaceae, Microbacteriaceae, and Weeksellaceae. However, these two control samples only contained 0.002% total MTs, while the *S. aureus* control contained 2.3%. Overall, each had very low percentages, but those that were expected to contain no bacterial DNA had 1,150 times fewer total MTs. Such a low amount of contamination is unlikely to significantly alter sequencing results.

CHAPTER 4: ALTERATIONS IN AIRWAY MICROBIOTA IN PATIENTS WITH LOW P/F RATIOS AFTER BURN AND INHALATION INJURY

4.1 Introduction

Smoke-induced inhalation injury occurs in up to 43% of burn victims, increasing death rates by up to 20% as compared to patients with burn injury alone [220]. Inhalation injury predisposes these patients to respiratory failure, acute respiratory distress syndrome (ARDS), and pneumonia. Pneumonia, in combination with burn and inhalation injury, further increases patient mortality to 60% and is a contributing risk factor to development of ARDS [181,196]. ARDS is a life-threatening condition resulting from either direct or indirect injury to the lung, and is diagnosed clinically by the presence of bilateral opacities on chest imaging and airway hypoxia [196,221]. Hypoxia is determined by the ratio of the partial pressure of arterial oxygen (PaO_2) to the fraction of inspired oxygen (FiO_2). To meet the Berlin definition of ARDS, this ratio must be less than or equal to 300 mm Hg, with a minimum positive end expiratory pressure (PEEP) of 5 cm H_2O [221]. Although bacterial infection is frequently the first step towards pneumonia and sepsis, and can induce direct injury to the lung and contribute to the pathogenesis of ARDS, its relationship with the disease is complex and not well understood [222].

Early antibiotic therapy is critical to improved patient outcomes once infection and pneumonia occur, but identification of the organisms can be challenging [183]. Current methodologies rely on culture or polymerase chain reaction (PCR) techniques to identify the causative agent [223]; however, these methods require specific knowledge of the organism's growth and metabolic requirements and a period of 1 – 2 days for identification and susceptibility testing, which are prone to false positive results [223]. These limitations often result in broad-spectrum antibiotic treatment that may have little to no impact on the target organism, promote the development of antibiotic resistance, and ultimately increase mortality [183,223].

To address these limitations, we utilized next-generation sequencing to characterize the bacterial communities (collectively known as microbiota) in the airways of burn patients following smoke inhalation with or without a PaO_2/FiO_2 (P/F) ratio $\leq$ 300, regardless of the presence of ARDS. Study of the microbiota has revealed the key roles they play in the development and function of the host immune system, and how dysbiosis, or perturbation of the communities, contributes to disease [35,37,16]. Although host-microbiota interactions are complex and poorly understood, recent studies underscore the importance of low-abundance species in dysbiosis and disease progression, particularly in the airways [16,17]. We hypothesized that inhalation injury and a low P/F ratio ($\leq$ 300) would create conditions within the airways that favor distinct communities of bacteria. We show that facultative anaerobic taxa are enriched among all burn patients, and that specific, low-abundance bacterial taxa are associated with low P/F ratios within the first 24 to 72 hours after injury.

4.2 Methods

4.2.1 Patients and Sample Collection

Therapeutic bronchial washings from patients hospitalized for burn and inhalation injury at the North Carolina Jaycee Burn Center were collected as previously described [199]. Briefly, patients with suspected inhalation injury underwent clinically indicated bronchoscopy within 24 hours of admission. All patients were intubated, bronchial washes performed, and inhalation injury severity scored on the basis of examination. Clinical cultures were grown to detect bacteria within these bronchoscopy samples. Organisms detected per patient and antibiotic treatment are listed in Table S2 in the additional data According to the Berlin definition of ARDS, hypoxia was defined as the ratio of the partial pressure of arterial oxygen (PaO_2) to the fraction of inspired oxygen (FiO_2) ≤300 [199]. Ratios >300 were defined as normal oxygenation levels [196]. Other clinical information, such as patient demographics and total body surface area burned, were collected upon admission. The study protocol was approved by the Institutional Review Board at the University of North Carolina School of Medicine in Chapel Hill (IRB# 10-0959 and #12-2475). All patients or their legally authorized representative gave informed consent for collection of their bronchial washings for inclusion in a repository as previously described [199]. Analysis of the microbiota in bronchial washings was not an original part of the study and was added after completion of sample collection (IRB #12-2475).

4.2.2 DNA Extraction and Sequencing

Bronchial washes were transported on ice and processed within 24 - 48 hours. DNA was extracted from the cellular portion of the wash. Positive *Staphylococcus aureus* and negative reagent and human DNA controls were extracted simultaneously and prepared in parallel with the patient samples for sequencing. Samples were centrifuged to separate the supernatant from the cellular fraction and these were stored separately at -80°C. The cellular fraction was used to extract bacterial DNA, and these were thawed briefly in a water bath at 35°C prior to extraction. Both enzymatic and physical methods were used to lyse the bacterial cell walls; the samples were resuspended in a lysis buffer including lysozyme and were placed in a Vortex mixer for 10 minutes. Samples were treated with RNase A to degrade contaminating RNA and DNA was extracted using the Qiagen QIAmp UCP Pathogen Mini Kit according to the manufacturer's protocol. The recommended lysis step was skipped in favor of the method described above. Quantitative real-time polymerase chain reaction (qPCR) was used to quantify human DNA [224] in the samples and total DNA was quantified using the Applied Biosystems PicoGreen double-stranded DNA dye. Bacterial DNA was quantified indirectly by subtracting the quantity of human DNA from total DNA. Strains of *S. aureus* and *K. pneumoniae* used as standard curve DNA were received from Carolina Biologicals (Burlington, NC). The 16HBE14o- cell line was a gift from D.C. Gruenert [225]. DNA extraction for all samples and standards was done as described above. Bacterial DNA was quantified using primers designed by Maeda *et. al.*[214].

Sequencing of bacterial DNA was performed in duplicate by a molecule tagging method recently described by Lundberg et al., [215]. This approach allows us to

confidently identify operational taxonomic units (OTU) that diverge at the 3% threshold. Briefly, a short round of polymerase chain reaction (PCR) was performed to attach molecule tags to each DNA molecule, followed by a round of full PCR to label each individual sample with a barcode and attach the adapters necessary for sequencing. The primers targeted the V4 region of the bacterial 16S rRNA gene with forward sequence GTGCCAGCMGCCGCGGTAA (515F) and reverse sequence TAATCTWTGGGVHCATCAGG (806R) [215]. Sequencing was performed on the Illumina MiSeq platform at the High Throughput Sequencing Facility at the University of North Carolina at Chapel Hill. Quality trimming of the resulting sequencing reads was performed using the Illumina CASAVA software. The MT-Toolbox pipeline, developed specifically to handle sequences resulting from the molecule tagging method, was used to generate consensus sequences from the molecule tags, group them into operational taxonomic units (OTUs), and match them to the GreenGenes 16S rRNA gene database to identify the sequences to the lowest bacterial taxonomic level possible [217,219].

4.2.3 Sequencing Data and Statistical Analysis

We used the MT-Toolbox [217] pipeline to minimize sequencing errors and match reads to the GreenGenes 16S rRNA database [219]. Sequences that did not match a 16S GreenGenes sequence were removed from the OTU table and those remaining were corrected for variation among 16S rRNA operon number. Duplicate samples were averaged and count thresholds were set for the OTU tables using an R-squared correlation analysis as detailed previously [226]. The samples varied according to the date of sequencing and thresholds were set separately for each group. Tables with

appropriate thresholds were imported into the program Explicet [114] for normalization and subsequent diversity analyses and the Wilcoxon rank-sum and two-proportions tests. Rarefaction was performed on sample counts within Explicet before bootstrapping to calculate Chao1 diversity indices. The Wilcoxon test is a non-parametric, continuity-corrected test appropriate for analysis of differential OTU abundances [114]. The two-proportions test performs a continuity-adjusted chi-square test to determine differences in detection among OTUs [114]. One-way analysis of variance (ANOVA) was used to identify differences among the abundance of aerobic and anaerobic bacterial taxa present in patients with and without ALI (performed in R as anova = lm(Taxa_per_seq_count~Group, data=ALI)) [227]. The linear discriminant analysis (LDA) effect size (LEfSe) method [122] was used to determine the significance of differences in taxa abundance by biologically relevant classes, which included patient P/F ratio and antibiotic treatment. LEfSe first performs a factorial Kruskal-Wallis test to determine differential distribution of OTUs among the biological classes. If subclasses are present, a pairwise Wilcoxon test is done on those with p values greater than 0.05. OTUs with significant differences are then used to build a linear discriminant analysis model, which uses the relative differences of OTUs among classes to rank those that are most discriminative.

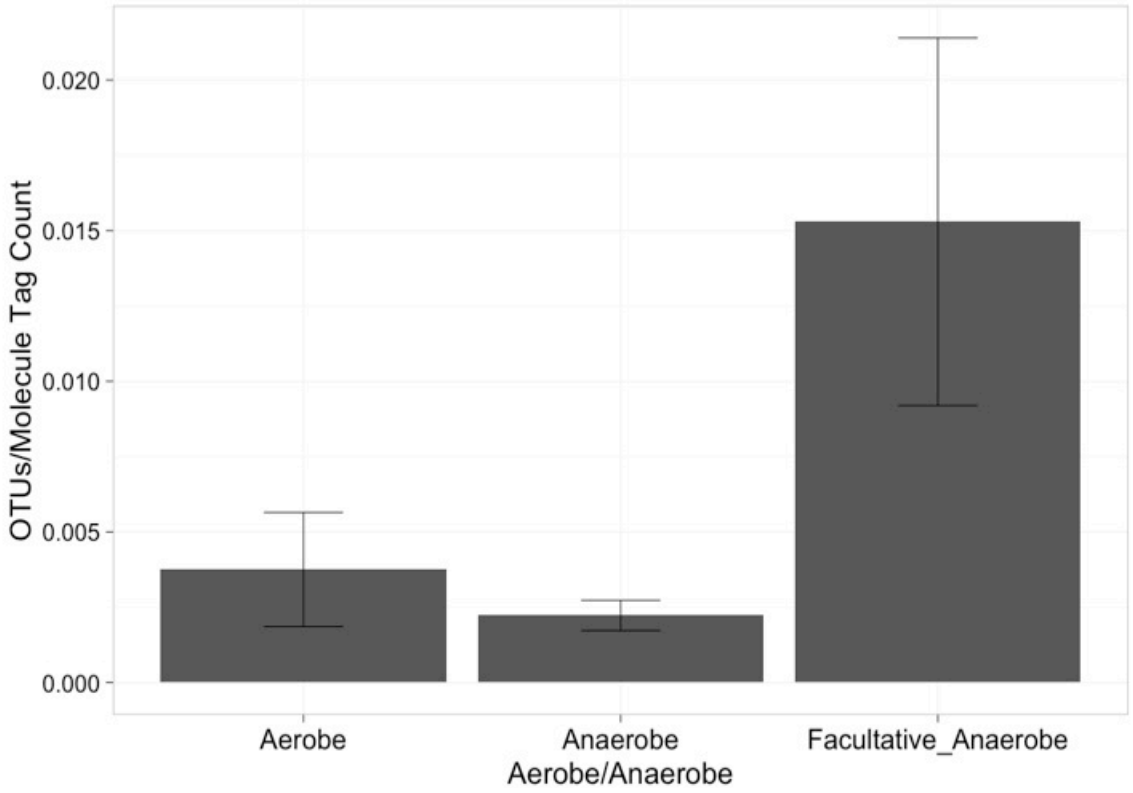

Figure 4.1: Unique Facultative Anaerobic OTUs are Significantly Enriched Among All Patients. OTUs were identified as facultative anaerobes, obligate anaerobes, or obligate aerobes among all patients. OTUs were quantified and normalized to the molecule tag count and averaged by bacterial aerobic or anaerobic capability. One-way ANOVA detected a significant difference among the mean taxa of facultative anaerobes (p = 0.029). (n = 48)

4.3 Results

4.3.1 Patients

Of the 48 patients included in this study, 50% had P/F ratios ≤ 300 (Table 4.1). The rate of positive bacterial cultures in both patients with (21%) and without (25%) a P/F ratio ≤ 300 was similar to the overall rate (23%). However, the rate of antibiotic treatment within the first 72 hours of injury in patients with a P/F ratio ≤ 300 was lower

(29%) than either patients with a higher P/F ratio (46%) or the entire group (40%). Antibiotic treatment was not associated with P/F ratio ≤ 300 (chi-square test, $p = 0.4$).

Clinical Variable	**Total**	**$PaO_2/FiO_2 \leq 300$**	**$PaO_2/FiO_2 > 300$**
Patients	48	24	24
Males	36 (75%)	18 (75%)	18 (75%)
Females	12 (25%)	6 (25%)	6 (25%)
BMI	27 (14 – 51)	30 (17 – 51)	25 (14 – 42)
Age	41 (1 – 75)	42 (8 – 76)	41 (1 – 75)
%TBSA	19 (0 – 85)	27 (0 – 85)	10 (0 – 40)
Antibiotic Treated	19 (40%)	7 (29%)	11 (46%)
Baux Score	60 (1 – 115)	71 (31 – 115)	50 (1 – 96)
Endotracheal Tube	29 (60%)	17 (71%)	12 (50%)
Days on Ventilator	35 (0 – 105)	45 (0 – 105)	25 (0 – 79)
Positive Cultures	11 (23%)	5 (21%)	6 (25%)
Survived	41 (87%)	17 (71%)*	24 (100%)

Table 4.1: Clinical Variables. Patient clinical characteristics were grouped by total population and subdivided by P/F ratio. The data are represented as mean (range) or number (percent). $PaO_2/FiO_2 > 300$ and $PaO_2/FiO_2 \leq 300$ group percentages are calculated per group total. *Cause of death was either or both cardiac and pulmonary failure.

4.3.2 The Airway Microbiota Among All Patients

Among all patient samples, OTUs identified as facultative anaerobic taxa were detected at a significantly higher rate than OTUs identified as either aerobic or obligate aerobic taxa (Figure 4.1; ANOVA p=0.029). The most abundant OTUs among all patient samples at the family level were *Streptococcaceae* and *Enterobacteriaceae*, which

accounted for 26% and 18% of total family-level OTUs, respectively. The remaining 56% of OTUs consisted of 45 additional families, each present at 7% of the total family-level OTUs or less.

4.3.3 Enrichment of Low-Abundance OTUs Among Patients with $PaO_2/FiO_2 \leq 300$

The *Streptococcaceae* and *Enterobacteriaceae* family-level OTU abundances were not significantly different between patients with and without $PaO_2/FiO_2 \leq 300$ (Wilcoxon test, $p > 0.05$). At the lowest level of OTU identification, *Enterobacteriaceae* family-level OTUs, *Streptococcus* genus-level OTUs, and *Staphylococcus* genus-level OTUs were detected in 80% of patients both with and without $PaO_2/FiO_2 \leq 300$ (Tables 4.2 and 4.3). However, when compared to patients with $PaO_2/FiO_2 > 300$, patients with the lower ratio had a 27% increase in OTUs identified as *Streptococcus spp.,* a 32% increase in Enterobacteriaceae, and an 83% increase in *Staphylococcus spp*. An additional six OTUs were detected in 80% of patients with $PaO_2/FiO_2 \leq 300$ at 3.1% or less of the total OTUs in this group (Table 4.3). All OTUs detected in 80% of patients were either facultative or obligate anaerobes. Figures 4.2 and 4.3 display OTU abundances at the family level that account for greater than 1% of the total OTUs among individual patients without and with $PaO_2/FiO_2 \leq 300$, respectively.

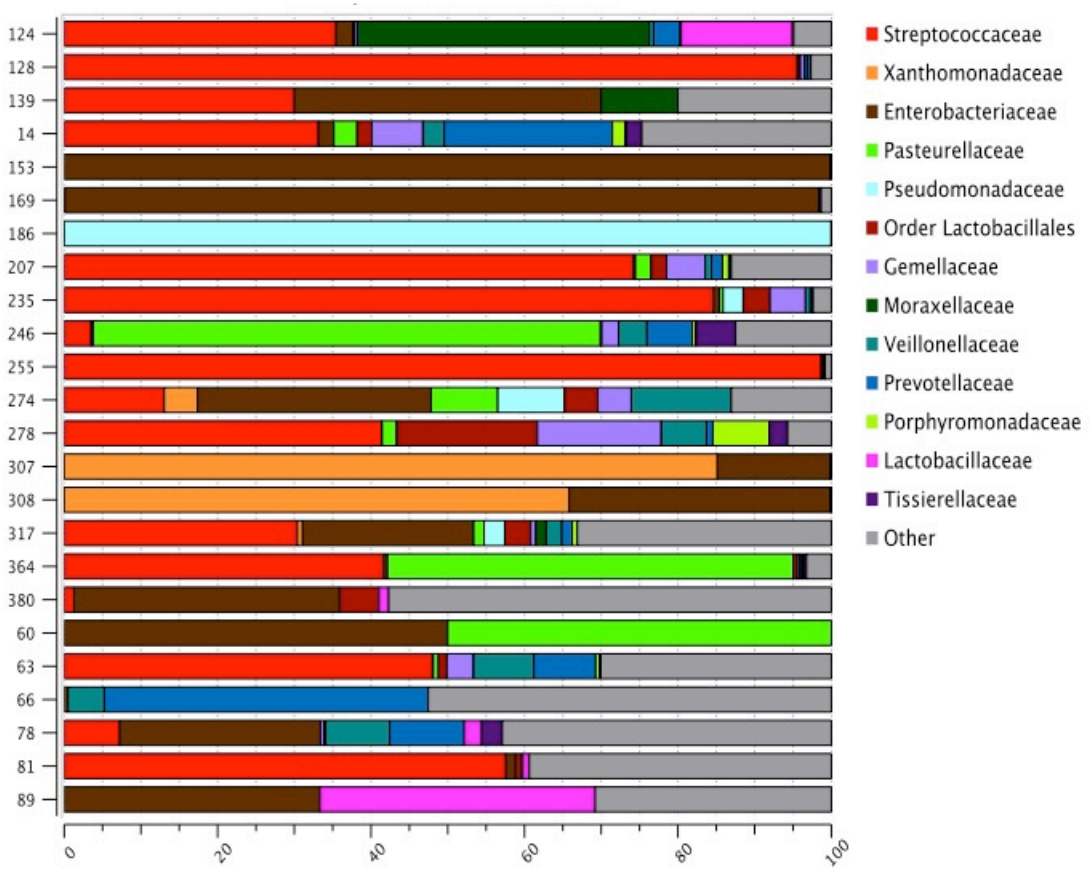

Figure 4.2: The Airway Microbiota Among Patients with PaO_2/FiO_2 > 300. OTUs identified as the families Streptococcaceae and Enterobacteriaceae dominate the airway microbiota within 72 hours following burn and inhalation injury. The category 'Other' includes bacterial taxa present at less than 1% of the total community. (n = 24)

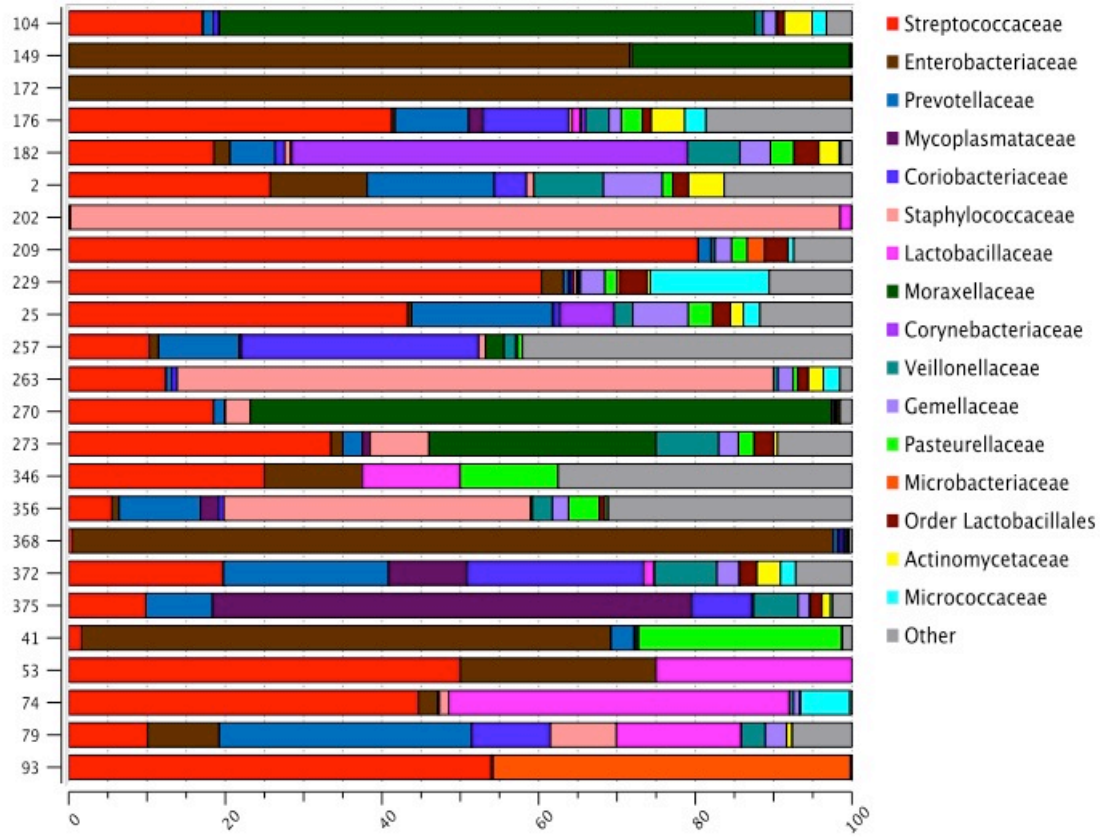

Figure 4.3: The Airway Microbiota Among Patients with $PaO_2/FiO_2 \leq 300$. OTUs identified as the families Streptococcaceae and Enterobacteriaceae dominate the airway microbiota within 72 hours following burn and inhalation injury. The category 'Other' includes bacterial taxa present at less than 1% of the total community. (n = 24)

Bacteria	Aerobe/Anaerobe	Average Abundance*
Enterobacteriaceae	Facultative anaerobe	20.0
Streptococcus spp.	Facultative anaerobe	27.8
Staphylococcus spp.	Facultative anaerobe	2.7

Table 4.2: Taxa Detected in 80% of Patients with $PaO_2/FiO_2 > 300$. Taxa names represent the lowest level of identification of the corresponding OTU. *Percent of total OTUs among 24 patients with $PaO_2/FiO_2 > 300$.

Bacteria	Aerobe/Anaerobe	Average Abundance*
Enterobacteriaceae	Facultative anaerobe	17.0
Streptococcus spp.	Facultative anaerobe	22.1
Staphylococcus spp	Facultative anaerobe	9.2
Atopobium spp.	Facultative anaerobe	3.1
Gemellaceae	Facultative anaerobe	1.8
Veillonella dispar	Obligate anaerobe	1.2
Lactobacillales	Facultative anaerobe	0.7
Prevotella spp.	Obligate anaerobe	2.4
Prevotella melaninogenica	Obligate anaerobe	2.5

Table 4.3: Taxa Detected in 80% of Patients with $PaO_2/FiO_2 \leq 300$. Taxa names represent the lowest level of identification of the corresponding OTU. *Percent of total OTUs among 24 patients with $PaO_2/FiO_2 \leq 300$.

4.3.4 Alpha Diversity Among Patients with and without $PaO_2/FiO_2 \leq 300$

The Chao1 diversity index, which is a non-parametric species richness estimator [228], did not show significant differences in number of different OTUs between patients with and without $PaO_2/FiO_2 \leq 300$ (Figure 4.4). Though the median Chao1 index in

patients with $PaO_2/FiO_2 > 300$ is less than that of patients with the lower ratio, it shows a much broader range in patients with $PaO_2/FiO_2 > 300$.

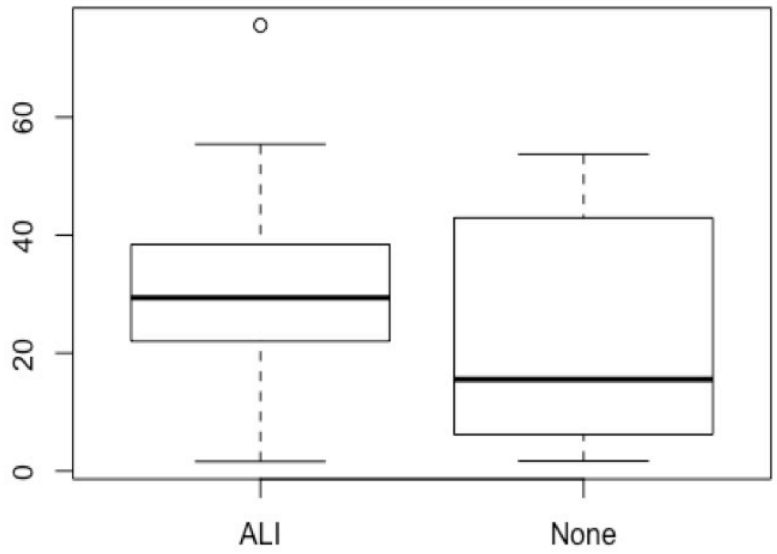

Figure 4.4: Average Chao1 Diversity Index of Patients with and without $PaO_2/FiO_2 \leq 300$. ALI = $PaO_2/FiO_2 \leq 300$, None = $PaO_2/FiO_2 > 300$ (n = 48)

4.3.5 Significant Enrichment of Specific OTUs Among Patients with $PaO_2/FiO_2 \leq 300$

Four OTUs were identified as significantly different in abundance and detection between patients with and without $PaO_2/FiO_2 \leq 300$. OTUs identified as *Prevotella melaninogenica, Mogibacterium spp.,* and *Corynebacterium spp.* were significantly increased in abundance among patients with $PaO_2/FiO_2 \leq 300$ (Table 4.4). Patients with $PaO_2/FiO_2 \leq 300$ had 72% more of the OTU represented by *Prevotella melaninogenica* than patients with $PaO_2/FiO_2 > 300$, 79% more *Corynebacterium* genus-level OTU, and 86% more of the *Mogibacterium* genus-level OTU. *Prevotella melaninogenica* OTUs were also detected significantly more frequently among patients with $PaO_2/FiO_2 \leq 300$, while *Corynebacterium* OTUs were significantly more frequent in patients with the higher ratio (Table 4.5). OTUs identified as *Fusobacterium spp.* were also detected

significantly more frequently among patients with $PaO_2/FiO_2 \leq 300$. LEfSe [122] was used to confirm these results. LEfSe identified the *Prevotella melaninogenica* OTU as most discriminative among patients with $PaO_2/FiO_2 \leq 300$ as compared to those with a higher ratio, followed by *Staphylococcus* genus-level and then *Bifidobacteriales* order-level OTUs (Figure 4.5). *Staphylococcus* OTUs were 83% more abundant in patients with $PaO_2/FiO_2 \leq 300$ as compared to those with $PaO_2/FiO_2 > 300$, while *Prevotella melaninogenica* OTUs were 72% more abundant and *Bifidobacteriales* OTUs were 50% more abundant (Figure 4.6). Additional analysis with LEfSe indicated significant enrichment of *Staphylococcus spp.* OTUs in the presence of antibiotics, while enrichment of *Prevotella melaninogenica* OTUs was not affected (Figures 4.7 and 4.8).

Taxa	**Percent Abundance Among Patients with $PaO_2/FiO_2 \leq 300$**	**Percent Abundance Among Patients with $PaO_2/FiO_2 > 300$**	**P-Value**
Prevotella melaninogenica	1.56	0.44	0.042
Corynebacterium spp.	1.53	0.32	0.037
Mogibacterium spp.	0.07	0.01	0.048

Table 4.4: OTU Level Significant Differences in Abundance as Determined by Wilcoxon Rank-Sum Test. Taxa names represent the lowest level of identification of the corresponding OTU.

Taxa	Detection Rate Among Patients with $PaO_2/FiO_2 \leq 300$ (# of patients)	Detection Rate Among Patients with $PaO_2/FiO_2 > 300$ (# of patients)	P-Value
Prevotella melaninogenica	19	11	0.037
Corynebacterium spp.	6	14	0.040
Fusobacterium spp.	17	9	0.043

Table 4.5: OTU Level Significant Differences in Detection as Determined by the Two-Proportions Test. Taxa names represent the lowest level of identification of the corresponding OTU.

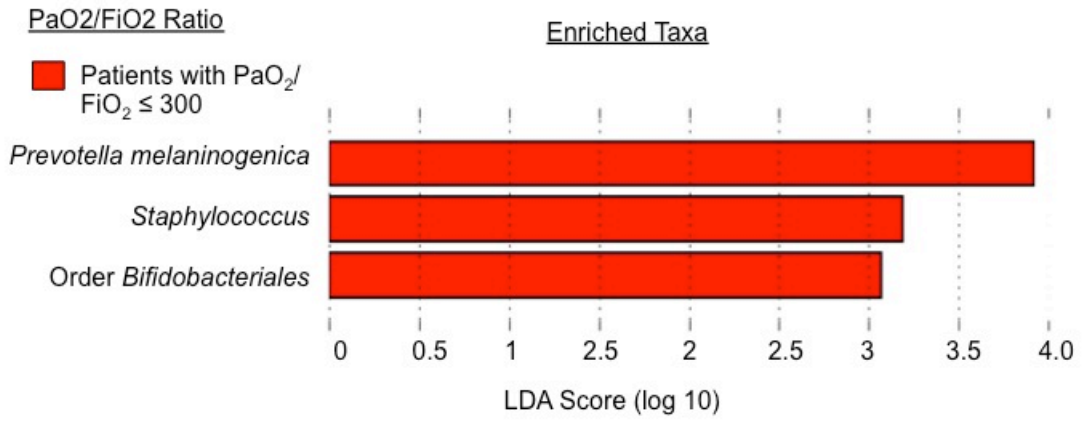

Figure 4.5: Specific Bacterial Taxa are Enriched Among Patients with $PaO_2/FiO_2 \leq 300$. LEfSe analysis detected significant enrichment of OTUs identified as *Prevotella melaninogenica*, *Staphylococcus spp.*, and the order Bifidobacteriales among patients with $PaO_2/FiO_2 \leq 300$. This tool uses a Kruskal-Wallis rank-sum test, Wilcoxon rank-sum test, and linear discriminant analysis to determine the biological relevance of significant enrichment of taxa and ranks them by effect size. LDA score indicates the magnitude of the effect size.

4.4 Discussion

Our work details differences in the airway microbiota in patients with PaO_2/FiO_2 ratios ≤ 300 and > 300 following burn and inhalation injury. A cut-off of 300 was chosen based on the Berlin definition of airway hypoxia in ARDS [221]. We identify several low-abundance OTUs with significant enrichment in patients with $PaO_2/FiO_2 \leq 300$, of which the OTU identified as *Prevotella melaninogenica* was the most significant. In addition, we show that while antibiotic treatment alters the airway microbiota, it does not explain the enrichment of a specific OTU among patients with $PaO_2/FiO_2 \leq 300$.

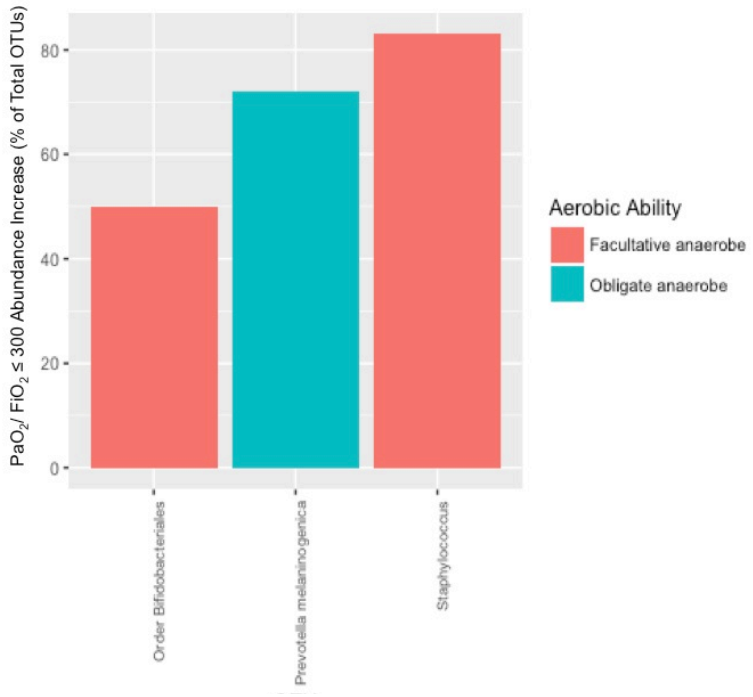

Figure 4.6: Percent Abundance Increase in OTUs with Significant Differences Detected by LEfSe. Bacterial abundances are displayed as the percent increase in patients with $PaO_2/FiO_2 \leq 300$ as compared to patients with $PaO_2/FiO_2 >$ 300. (n = 48)

Patients with a PaO_2/FiO_2 ratio that was less than or equal to 300 within 72 hours of burn and inhalation injury had consistently worse indicators of poor prognosis. Table 1 shows the average values for patients with and without $PaO_2/FiO_2 \leq 300$ for several clinical variables that are predictive of injury severity. In patients with inhalation injury, several studies have demonstrated that age, percent TBSA and PaO_2/FiO_2 ratio predict

mortality [181]. The PaO_2/FiO_2 ratio itself has been shown to be more predictive of patient outcomes on the day after patients meet the Berlin definition of ARDS rather than the day of [229]. In our study, patients with $PaO_2/FiO_2 \leq 300$ had, on average, a higher Baux score (age + %TBSA), spent longer on the ventilator, were intubated more frequently, and had lower survival rates. Only percent TBSA and the PaO_2/FiO_2 ratio were significantly different among the patient groups (Student's t test, $p = 0.002$ and 5.293e-11, respectively). While fewer patients with $PaO_2/FiO_2 \leq 300$ received antibiotic treatment than those with ratios > 300, rates of positive clinical bacterial cultures were similar between the two groups. This discrepancy may be partly due to the challenges in predicting bacterial infection and development of pneumonia in this patient population. Pneumonia is the primary complication of inhalation injury [180] and early, adequate antibiotic treatment has been shown to improve outcomes in these patients [183]. Criteria to predict pneumonia early after injury have been developed and include age > 60 years, TBSA > 20%, and initial PaO_2/FiO_2 ratio of ≤ 300 [230]. The patients with $PaO_2/FiO_2 \leq 300$ in our study meet the TBSA and initial PaO_2/FiO_2 ratio criteria, but not the age criteria, which may explain why they did not receive as many antibiotics. A major limitation of this scoring system is its failure to take into account bacteria within the airways, emphasizing the need for one that does, perhaps through a combination of clinical cultures and next-generation sequencing.

Though we have focused on the PaO_2/FiO_2 ratio in alterations of the airway microbiota, TBSA may also contribute to the differences we detected. Increasing TBSA is a known predictor of patient mortality [180], which is compounded in the presence of inhalation injury. Burns greater than 20% TBSA induce systemic changes similar to those

seen in trauma and surgical patients [177]. The injury induces a systemic inflammatory response, but compromises global immune function, increasing susceptibility to bacterial, viral, and fungal infections. Patients with $PaO_2/FiO_2 \leq 300$ in our study had, on average, 27% TBSA, indicating immune dysfunction that could predispose them to airway bacterial colonization and infection. Though we cannot determine whether the burn injury itself induces $PaO_2/FiO_2 \leq 300$ through systemic changes or if this is a direct result of inhalation injury, it is clear that TBSA may be contributing indirectly to alterations in the airway microbiota in our patient population. A mouse model of burn and inhalation injury is necessary to determine the extent to which TBSA influences changes in the airway microbiota.

Among all patients in the study, there were significantly more unique OTUs identified as facultative anaerobes than either strict anaerobes or aerobes (Figure 4.1). Anaerobic taxa are normally associated with mucosal surfaces, but may lead to infection following disruption by trauma and surgery [231]. All patients within this study, regardless of PaO_2/FiO_2 ratio, presumably experienced disruption of their mucosa through the double trauma of burn and inhalation injury. Recent work has demonstrated that the mouth serves as the primary source community for the airway microbiota [139]. Inhalation injury may have increased microbial immigration through disruption of the mouth and upper airways' mucosal surface, dislodging facultative anaerobic taxa that subsequently traveled down the airways to the bronchi. Alteration of airway conditions by inhalation injury may have selected for enrichment of facultative anaerobic taxa among all patients, which was significantly different from strict aerobic and anaerobic taxa (ANOVA, $p = 0.029$). When we subdivided the data by PaO_2/FiO_2 ratio, we did not

see significant differences in strict aerobes, anaerobes or facultative anaerobes between the two patient groups (Figure 4.9, $p > 0.05$). These results suggest that $PaO_2/FiO_2 \leq 300$ early after burn and inhalation injury does not select for overall taxa in the airways based on their aerobic or anaerobic capabilities, but that burn and inhalation injury do. Development of $PaO_2/FiO_2 \leq 300$ within 72 hours of burn and inhalation injury may not be enough time to observe significant change in the abundances of overall taxa between the two groups.

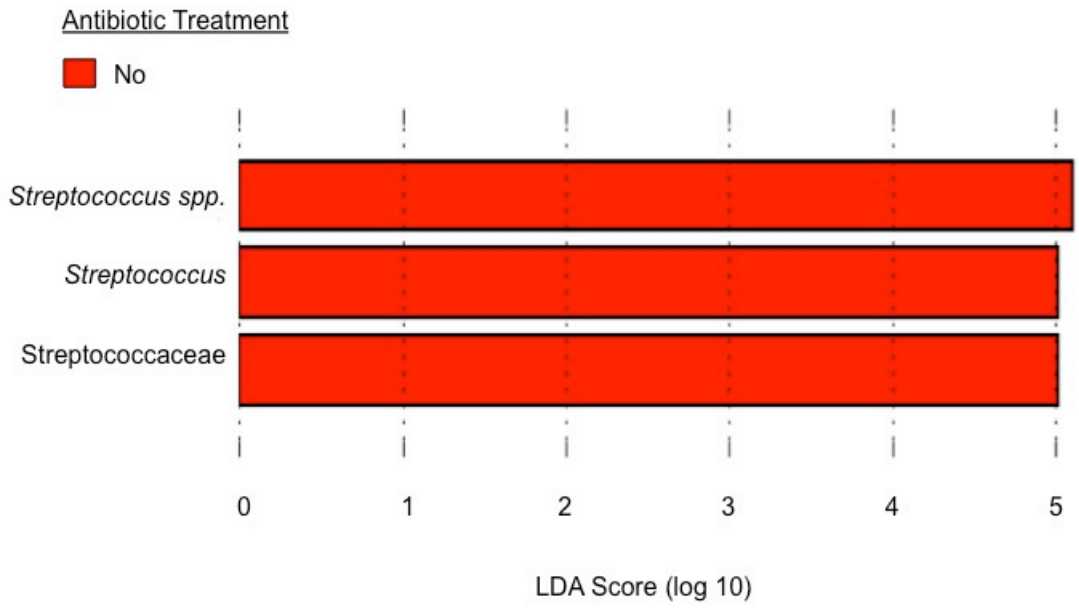

Figure 4.7: **Streptococcaceae Family Members are Enriched in Patients with $PaO_2/FiO_2 \leq 300$ without Antibiotic Treatment.** Among patients with $PaO_2/FiO_2 \leq 300$, only those not treated with antibiotics contained significantly enriched taxa, all of which were in the Streptococcaceae family.

We detected OTUs identified as *Enterobacteriaceae*, *Streptococcus spp.*, and *Staphylococcus spp.* in 80% of patients with and without $PaO_2/FiO_2 \leq 300$ (Tables 4.2 and 4.3). All three of these OTUs are facultative anaerobes and their dominance across patients implies similarity in the mechanism of injury to the airways selecting for these

taxa and their related functions. Inhalation injury may induce fluctuations in oxygen availability in the airways, perhaps creating both aerobic and anaerobic microenvironments that favor taxa that can withstand these changes. Our finding of overall significant enrichment of facultative anaerobic taxa supports this idea. Patients with $PaO_2/FiO_2 \leq 300$ demonstrated a 32%, 27%, and 83% increase in *Enterobacteriaceae*, *Streptococcus spp.*, and *Staphylococcus spp.* OTUs, respectively, when compared to those with $PaO_2/FiO_2 > 300$ (Tables 4.2 and 4.3). Additionally, 80% of patients with $PaO_2/FiO_2 \leq 300$ contained six more OTUs that represented 3.1% and less of the total community among these patients (Table 4.3). This suggests that, although facultative anaerobes are enriched over strict anaerobes and aerobes among all patients, there are differences in enrichment of specific, low-abundance OTUs depending on PaO_2/FiO_2 ratio.

Enterobacteriaceae, Streptococcus spp., and *Staphylococcus spp.* have all been consistently detected in previous airway microbiome studies in both healthy and diseased airways [133]. Members of the *Enterobaceriaceae* family have been implicated in inflammation-driven colorectal cancer in the gut microbiome [48,232], are enriched in patients with COPD and asthma, but can also be detected in healthy airways [58,145]. Similarly, *Streptococcus* is consistently found in healthy airways but is enriched in COPD [143], idiopathic pulmonary fibrosis (IPF) [146], and pneumonia [233]. *Staphylococcus*, while a normal commensal in the nasal microbiome [234,235], is largely associated with disease in the lung, such as IPF [146], and cystic fibrosis, in which it is correlated with increased inflammation [147,236]. Given the inconsistency with which these three taxa are associated with health or disease, it is difficult to interpret the

importance of their detection across patients with and without $PaO_2/FiO_2 \leq 300$. They may indicate an underlying core airway microbiota among all burn and inhalation injury patients but it is not clear whether their presence is beneficial or detrimental. A longitudinal study of patients with burn and inhalation injury could clarify the role of these taxa.

Due to its association with health outcomes, overall diversity has long been a focus in many microbiome studies; however, we observed no difference in diversity between patients with $PaO_2/FiO_2 \leq 300$ and those with $PaO_2/FiO_2 > 300$ (Figure 4.4). This agrees with recent studies demonstrating that diversity (especially in the airways) is a complex, multifactorial trait that is unlikely to simply indicate positive or negative outcomes [145,236]. Many of these studies have emphasized the critical roles of specific taxa during disease and their interactions with other taxa [16,17]. They suggest that rare and less abundant taxa, which are overlooked by traditional culture methods, may play significant roles in the development of disease. Dysbiosis of the microbiota is followed by enrichment of a specific bacterial taxa that is either rarely found or present at very low abundance [16,233]. Changes in the balance of bacterial taxa alters how the microbes interact with each other along with their associated functions, allowing species that may have been suppressed by the presence of other bacteria to overgrow [16]. What was considered a harmless commensal in a healthy individual may become a harmful pathogen under dysbiosis-inducing conditions [39]. Accordingly, in our study, we observed significant differences not in the species dominating the overall community, but in less abundant taxa. While these taxa do not differ in microbial diversity, they may differ by functional diversity, which ultimately plays a greater role in patient outcomes

[133]. Most significantly, we identified enrichment of the OTU identified as *Prevotella melaninogenica* among patients with $PaO_2/FiO_2 \leq 300$ within 72 hours of burn and inhalation injury.

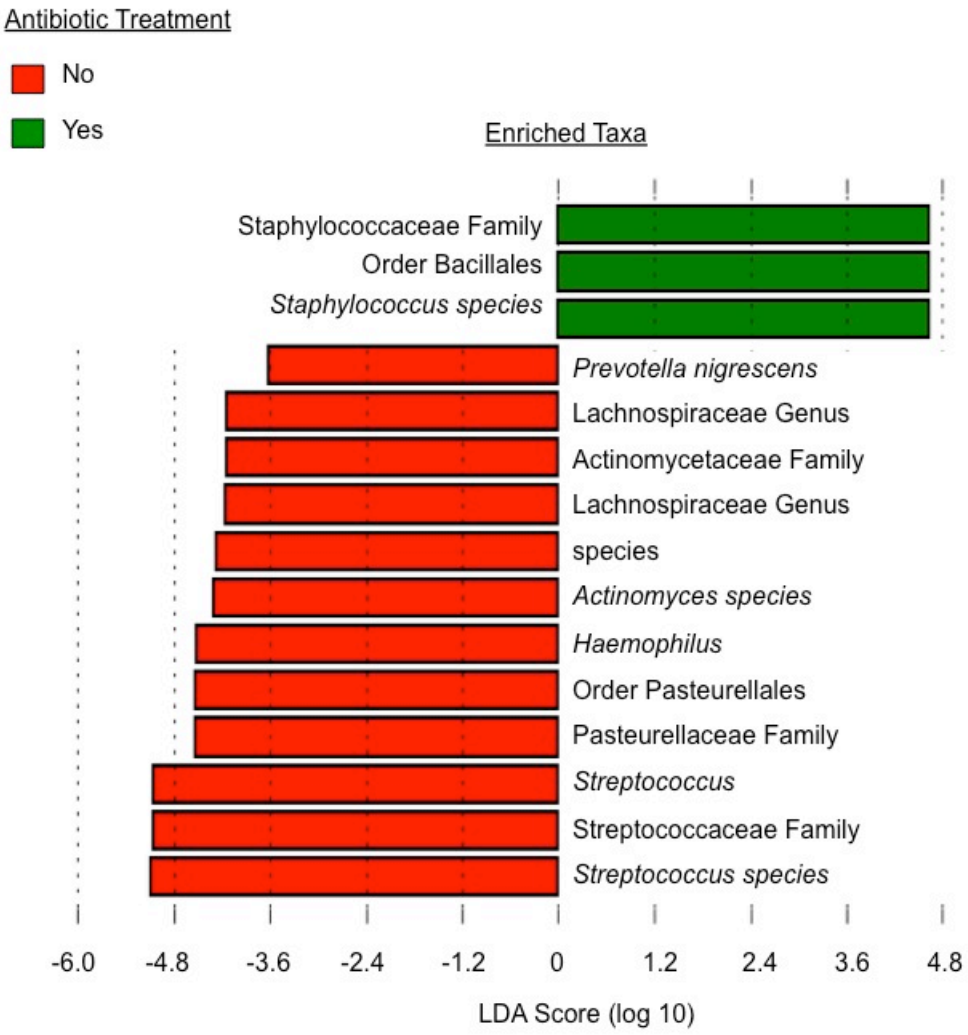

Figure 4.8: Antibiotic Treatment Alters the Microbiome but Does Not Impact Association of the *Prevotella melaninogenica* OTU with $PaO_2/FiO_2 \leq 300$. LEfSe analysis identifies a significant increase in *Staphylococcus* among all patients treated with antibiotics.

Prevotella melaninogenica is a gram-negative obligate anaerobe that is part of the normal flora but is also a significant source of infection [237]. The specifics of *Prevotella melaninogenica*'s function in the microbiome remain unclear. In the gut, it has been identified as a normal commensal family, but within dental plaque it is a potential pathogen [124]. In the upper airways, the presence of *Prevotella melaninogenica* is associated with health while lactobacilli, *Rothia spp.*, and *Streptococcus pneumoniae* dominate bacterial profiles in patients with pneumonia [233]. *Prevotella melaninogenica*'s positive role in the airways is supported by its ability to decrease production of T cell-activating IL-12p70 by dendritic cells exposed to *Haemophilus influenzae* [238]. This highlights the ability of bacteria within microbial communities to regulate each other's functions as well as that of the host immune system. Several studies indicate that *Prevotella melaninogenica* could also play a non-beneficial or harmful role in the airways under certain conditions. *Prevotella melaninogenica* was a dominant bacterial species isolated from the airways of intubated patients [239] as well as cystic fibrosis patients, where characterized species varied phenotypically over time [240]. Though present at low abundance, we identified a large, consistent, and significant enrichment of the *Prevotella melaninogenica* OTU among patients with $PaO_2/FiO_2 \leq 300$ within 72 hours of burn and inhalation injury. While facultative anaerobic taxa were enriched among all patients in the study, *Prevotella melaninogenica* was enriched specifically in patients whose airways have the lowest PaO_2/FiO_2 ratio, which may select for growth of this obligate anaerobe. Without pre-injury samples from the patients, it is not possible to determine whether enrichment of *Prevotella melaninogenica* precedes $PaO_2/FiO_2 \leq 300$ or if a low PaO_2/FiO_2 ratio precedes this enrichment. If confirmed in a

longitudinal study, the consistent presence of this OTU throughout the hospital stay of patients with $PaO_2/FiO_2 \leq 300$ would suggest that it is in some way altering the airway environment to favor *Prevotella melaninogenica*. This could be achieved through elimination of other OTUs it interacts with that cannot thrive in hypoxic conditions

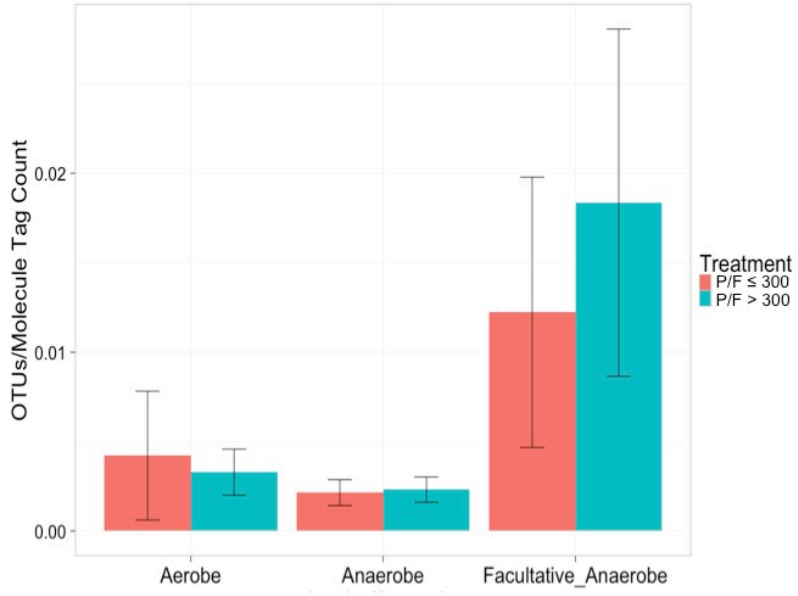

Figure 4.9: Average Unique OTUs Identified as Facultative or Strict Anaerobes and Aerobes Among Patients with and without $PaO_2/FiO_2 \leq 300$. (n = 48)

or outgrowth of those that can. Early changes in both oxygen availability and other OTUs may impact *Prevotella melaninogenica*'s ability to act as a pathogen depending on whether species it interacts with are increased or eliminated or airway conditions alter its growth and pathogenicity. Given that *Prevotella melaninogenica* is an obligate anaerobe, hypoxic conditions may favor its growth, but it is impossible to predict its pathogenicity without further study. While determining a causal link between $PaO_2/FiO_2 \leq 300$ and *Prevotella melaninogenica* is beyond the scope of the current study, future studies will examine its pathogenicity from patients with and without $PaO_2/FiO_2 \leq 300$ as well as its role in either preceding or following hypoxia.

Infection is a serious concern in these immunocompromised patients, for whom mortality rates increase to 20% with inhalation injury alone and triple to 60% when present with pneumonia [241]. Prophylactic antibiotic treatment is a common strategy to prevent infection, but results in as many as 25% of patients without infections receiving antibiotics [242], which may alter the microbiota in deleterious ways and encourage

Patient Number	Bacteria Cultured	Percent of Individual Community*	Antibiotic Treatment
14	ORSA	0	None
79	Unknown	NA	Vancomycin/iso-dex & Piperacillin/tazobactam/isoOsmoticPMB
93	OSSA	0.1	None
	S. pneumoniae	54	
104	*S. pneumoniae*	17.1	Tigecycline
124	*S. pneumoniae*	35.3	Tigecycline & Piperacillin
128	*Acinetobacter*	0.02	None
	H. influenzae	0	
169	*S. pneumoniae*	0.1	Piperacillin/tazobactam/ isoOsmoticPMB
202	*Enterobacter*	0.1	Tigecycline & Piperacillin
	H. influenzae	0	
	S. pneumoniae	0.1	
	OSSA	91.2	
308	Unknown	NA	Vancomycin
346	OSSA	0	None
	S. pneumoniae	25	
380	Unknown	NA	Tigecycline & piperacillin/tazobactam/isoOsmoticPMB

Table 4.6: Clinical Cultures. Bacteria detected by clinical culture per patient, their corresponding abundance as detected by NGS, and patient antibiotic treatment.

outgrowth of resistant bacteria [243]. Antibiotic treatment has been shown to perturb the gut microbiome and immune cell response by eliminating commensal species and allowing drug-resistant bacteria to take over [244,245]. In the airways, antibiotic treatment in asthma shows a similar response, in which elimination of certain species provides a niche for establishment of other infectious species [150]. Three months of varying types of antibiotic treatment in patients with COPD did not reduce overall bacterial load and increased antibiotic resistance across all groups [246]. While a powerful tool for controlling bacterial growth, antibiotic treatment is a double-edged sword that can create communities of bacteria resistant to treatment. Our poor understanding of bacterial interactions within the microbiota and their roles in patient outcomes combined with antibiotics' lack of specificity results in overkilling of beneficial organisms that could aid in improving patient outcomes. In our study, 18 total patients were treated with antibiotics; nine of these had negative culture results and for two, cultures were not done (Table 4.6). If negative culture results indicate absence of infection in these patients, antibiotic treatment is unnecessarily altering the airway microbiota, possibly contributing to development of resistance and poor outcomes. Among all patients treated with antibiotics, analysis with LEfSe indicated significant enrichment of bacteria in the *Staphylococcaceae* family and the order *Bacillales (*Figure 4.8). These bacteria may be resistant to the drugs or not targeted by them, leading to overgrowth of these particular species. Methicillin-resistant *Staphylococcus aureus* is a known problematic infection in hospitals, including the Jaycee Burn Center, but its role within burn patient microbiota is unknown and requires further study. Despite alteration of other taxa by antibiotic treatment, enrichment with the *Prevotella melaninogenica*

OTU among patients with $PaO_2/FiO_2 \leq 300$ was not affected, implying that its association with hypoxia is independent of antibiotic treatment, at least within 72 hours of injury. Further study is necessary to determine the role of this OTU in early development of hypoxia and whether targeted antibiotic treatment may be beneficial.

There are several limitations to the study. Although unique in its examination of a heterogeneous group of burn patients, our work is also limited by this variability. The number of patients studied is comparable to or larger than previous studies of microbiota in airway disease [16,143]. The small number of patients treated with antibiotics and cross-sectional nature of the study makes investigation of the effect of different antibiotics on the airway microbiota impractical.

In conclusion, we have demonstrated differences in the airway microbiota of patients with and without $PaO_2/FiO_2 \leq 300$ within 72 hours of burn and inhalation injury. We detected overall enrichment of facultative anaerobes among all patients with differences in specific OTUs among patients with and without $PaO_2/FiO_2 \leq 300$. Significant differences between these patients reside among the less abundant OTUs, specifically the *Prevotella melaninogenica* OTU, an obligate anaerobe whose role in the microbiome is unclear. Hypoxic conditions indicative of ARDS development may favor *Prevotella melaninogenica* enrichment and alter its pathogenicity. Alternatively, hypoxia may develop due to increased abundance of this OTU following inhalation injury. A mouse model of inhalation injury is needed to determine whether development of hypoxia drives enrichment of *Prevotella melaninogenica* or enrichment of this OTU induces hypoxia. Given the cross-sectional nature of this study, more work is necessary to determine the long-term impact of *Prevotella melaninogenica* and its role in the airway

microbiome of burn patients with inhalation injury who develop $PaO_2/FiO_2 \leq 300$ within 72 hours of injury. Importantly, antibiotic treatment did not alter this association, supporting a link between this OTU and $PaO_2/FiO_2 \leq 300$ early after burn and inhalation injury.

CHAPTER 5: PREDICTION OF BACTERIAL TAXA ASSOCIATED WITH $PAO_2/FIO_2 \leq 300$ AFTER BURN AND INHALATION INJURY

5.1 Introduction

An exciting potential of microbiome research is to predict outcomes based on community structure. The field of machine learning is increasingly being utilized to predict outcomes from medical, experimental, and environmental data sets, including those from microbiome studies [247–250]. Machine learning algorithms have been successfully applied to patient medical records to predict colorectal cancer [251], to somatic mutations, copy number alterations, DNA methylation, and gene expression in tumors to predict drug response [252], and to patients undergoing repair of an abdominal aortic aneurysm to predict mortality [253]. Machine learning evolved from the field of artificial intelligence and allows a computer program to 'learn' without being explicitly programmed [254]. Data is given as input to an algorithm, from which it learns to predict outcomes. The more data it has as input, the better the algorithm will be at predicting output. Metagenomic data sets are large and high dimensional in nature, which makes them well suited to machine learning methods.

Metagenomic data sets are organized in a typical N x p matrix, in which N is the number of samples and p is the number of features, or variables, in the data [255]. The features p are also referred to as the dimensions of the data and in this way the matrix becomes a collection of N points in a p-dimensional space. It is common in metagenomic

studies for p to be much larger than N, as the number of bacterial species or operational taxonomic units (OTUs) typically outweigh the number of subjects in the study. Classical statistical methods operate on the assumption that p is less than N and N increases towards infinity [255]. These methods fail when p is larger than N, which tends to be the case in most metagenomic studies unless they have very large sample sizes. The growing pervasiveness of this problem in modern data sets has resulted in the development of methods to deal with high dimensional data, usually through some form of dimensionality reduction prior to interpretation and analysis [256]. Both unsupervised and supervised methods exist for this purpose. Unsupervised methods, which examine relationships between the data points to reveal underlying structure independent of data response variables, can be used as a preprocessing reduction step before use of supervised methods [256,257]. Clustering and principle components analysis (PCA) are commonly used unsupervised methods. Supervised methods categorize data based on a response variable, such as patient outcome. The reduced data set can be applied to these methods in order to make predictions about the data. Regression is a well-known supervised method and machine learning algorithms such as random forests (RF), neural networks, and support vector machines (SVM) fall into this category as well. Appropriate use of unsupervised and supervised methods in the context of microbiome data can lead to novel inference about the data and accurate prediction of outcomes.

5.1.1 Unsupervised Clustering Methods Reveal Data Structure

Unsupervised clustering methods are a form of exploratory data analysis that allow discovery of relationships among data points independent of associated variables

[257]. They provide an unbiased way of determining which features are most important in the data set so that the N x p matrix can be reduced to $N > p$ in order to allow analysis with classical statistical methods. Several popular methods that are extensively used with microbiome data include PCA, hierarchical clustering, and *K*-means clustering. PCA is a form of dimensionality reduction that can be easily visualized, while hierarchical and *K*-means clustering group the data based on similarity. While each method is useful, understanding the theory behind them is necessary for choosing the most appropriate one for data analysis.

PCA summarizes a high dimensional data set with representative variables that explain the most variability within the entire data set. This is done through generation of principle components (PC) by linear combination of the features that maximize variance. The linear combination of features is similar to a weighted average that takes into account each feature and its variation [258]. Each feature is normalized with a value referred to as a loading. The total loadings for each feature comprise a loading vector, which is used to generate the scores of each PC. The loadings give a direction for the data in feature space, along which the scores for each PC can be projected. Thus, PCA allows for selection of features based on variability as well as visualization, providing easy identification of similar data points in a lower dimension space.

Clustering methods differ from PCA in that they seek to find similar subgroups among the dataset, rather than dimension reduction based on the best explanation of variance [257]. Both hierarchical and *K*-means clustering methods cluster the data based on similarity. *K*-means requires specification of the desired number of clusters while hierarchical does not. Hierarchical clustering results in an easily interpreted dendrogram

that is usually built in an agglomerative, bottom-up way. All observations begin as their own cluster at the bottom of the tree. Those that are most similar are fused, and the algorithm moves up to the next leaf. Here, the previously formed clusters are compared to the remainder of the data, and those that are most similar are grouped again. This process continues until all samples are grouped in a single cluster or tree. In *K*-means clustering, observations are grouped into the specified number of clusters iteratively until the smallest within-cluster variation is found. Due to the requirement of cluster size specification, *K*-means is used less frequently with microbiome data than hierarchical clustering is. Like PCA, both provide a way to group the data based on similarity in order to find trends within it.

Discriminant analysis of principle components (DAPC) improves on the previously described unsupervised methods by combining PCA and discriminant analysis (DA) [259]. PCA incorporates both between-group and within-group variation, which does not allow assessment of the relationship between clusters. DA maximizes the impact of between-group variation and minimizes that of within-group variation in determining discriminative variables within the data. Unlike PCA, it has no method for reducing dimensionality of the data, and it cannot handle data with p larger than N. It also cannot compensate for the compositional nature of the data, which will result in false correlations. DAPC uses PCA to transform the data as a prior step to DA. This allows both data transformation and feature selection, which addresses the issues associated with DA analysis of high dimensional data sets. DAPC optimizes the variance between groups while minimizing that within groups, allowing identification of discriminant features [260]. After application of PCA to reduce dimensionality, DAPC uses *K*-means

clustering with an increasing number of clusters to select the number that best represents the data. This method was developed for high dimensional genetic data but also works well for clustering metagenomic data.

5.1.2 Machine Learning Algorithms Predict Outcomes

Several machine learning algorithms exist that have been applied to microbiome data. Regression methods are commonly used to assess correlations between variables and can be useful in selecting subsets of predictors related to the response, shrinking coefficients to reduce variance, and reducing dimensionality of the data [257]. Elastic net regression, which includes the lasso and ridge methods, is particularly useful for high-dimensional data. Both of these models constrain coefficient estimates, which shrinks them towards zero and reduces their variance, resulting in improved prediction accuracy. They also contain a penalty factor that creates bias within the model but decreases variance, resulting in overall increased accuracy [261]. Both ridge and lasso contain a λ tuning parameter that must be determined by cross-validation using a subset of the data. An increasing value of λ in ridge regression reduces the coefficients but keeps all predictors in the model. In lasso, increasing values of λ result in some coefficients becoming zero, which effectively removes the predictors they correspond to from the model. Lasso regression therefore performs variable selection, which decreases the complexity and dimensionality of the data. The elastic net method, which combines both lasso and ridge regression, is useful in that it provides a less flexible approach and prevents overfitting of the model, which can lead to inaccurate predictions.

Both lasso and ridge regression are types of linear classification models that may not adequately model microbiome data, which is usually non-linear in nature [261]. As a result, non-linear regression models are more frequently used, such as support vector machines (SVM), random forests (RF), and neural networks. SVM is a generalization of the maximal margin classifier, and is depicted in Figure 5.1 [257]. It consists of a hyperplane (labeled the maximal margin hyperplane in Figure 5.1) that optimally separates the data points. The maximal margin establishes the best choice of separating hyperplane, since infinite hyperplanes could separate the data. Here, the distance between possible hyperplanes and the observations is measured. The maximal margin hyperplane is the hyperplane with the largest distance between it and the nearest observations. These nearest observations (white points in Figure 5.1) are termed support vectors because they establish and support the maximal margin. The dependence of the margin on a subset of

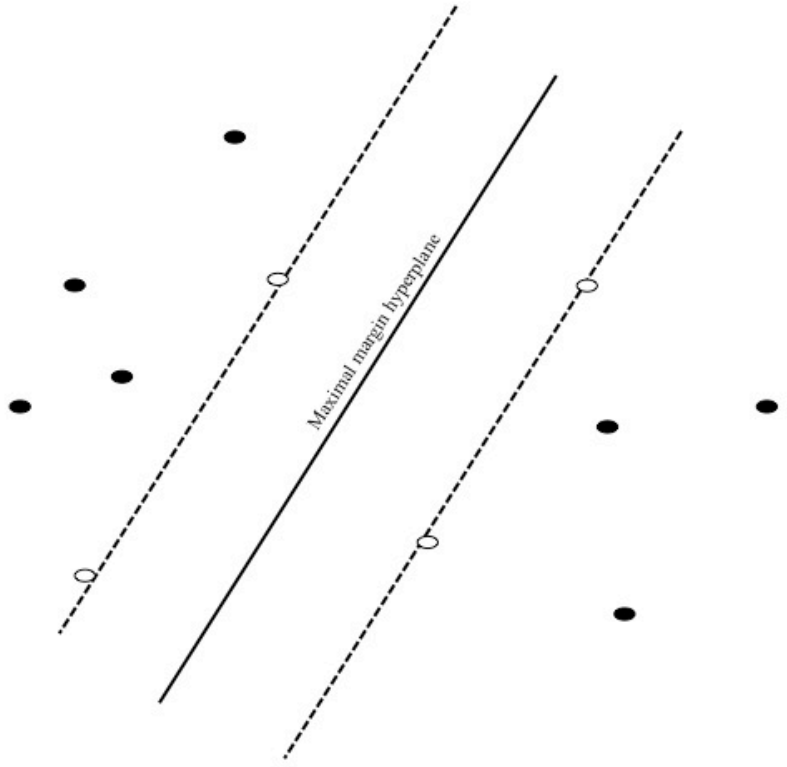

Figure 5.1: The Maximal Margin Classifier. A hyperplane (maximal margin hyperplane; solid line) separates the data. Observations closest to the hyperplane act as support vectors (white data points) that establish the maximal margin (dotted lines).

the data allows for more accurate and consistent classification of incoming data. In many cases, it is not possible to neatly separate data by a linear hyperplane. Further, data may need to be classified into more than two categories. The SVM was developed to address both of these issues. The SVM enlarges feature space using kernels in order to encompass non-linear data. A kernel is a mathematical function that describes the similarity of two data points, allowing use of equations specific to non-linear data. An appropriate kernel function will establish non-linear boundaries for the maximal margin classifier that best fits the data. Incorporation of more than one class is a more difficult challenge and relies on one-versus-one or one-versus-all approaches. In one-versus-one, pairs of classes are compared in order to determine classification. In one-versus-all, each class is compared to all the others to classify observations. Though SVM is useful in categorizing linear data into binary classes, neural networks and RF may be more appropriate choices for most microbiome data.

The development of artificial neural networks was influenced by biological neural networks, in which a single neuron communicates with several other neurons through axons [254]. Though several types of neural networks exist, all function on a similar structural basis, as shown in Figure 5.2. The most used neural network is the multilayer perceptron, which consists of between one and three hidden layers of neurons and an output neuron [254]. Each neuron is connected to every neuron in the layer preceding it but neurons within the same layer are not connected to each other. This is essential to

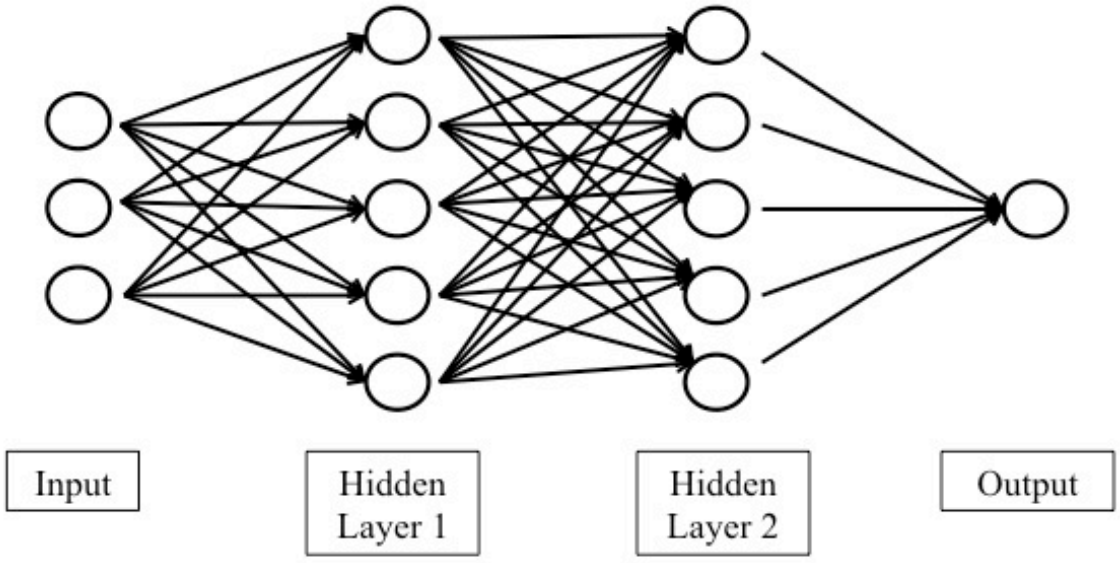

Figure 5.2: Neural Networks. Input neurons receive a weighted sum of data, which is transformed by a mathematical function by interconnected sets of hidden layer neurons. Multiplication by the weights is used to determine classification of the output data.

accurate classification within the model. Each neuron-to-neuron link is associated with a weight. The input values are multiplied by this weight before being subjected to the transfer function in the next layer of neurons. Each of these next neurons also has an associated weight that the resulting values from the previous layer are multiplied by before being subjected to the transfer function again. This process is referred to as feed-forward propagation and results in varying weights of evidence that allow the network to choose the best classification of the input data. The universality theorem states that, with the right number of neurons and the right choice of weights, a neural network model can be used to solve any classification problem accurately [254]. Unfortunately, knowing what these correct choices are is more difficult. The choice of correct weights for a neural network is done through a training process known as backpropagation of error. This is an

iterative method in which a range of weights is assigned to the model and input is propagated through the network. Both the error and each neuron's degree of responsibility for it can be calculated backwards through the network and used to adjust the weights appropriately. Though highly accurate and applicable to nearly any classification problem, neural networks are prone to overfitting and can be computationally expensive, both of which are partly due to the problem of choosing the optimal number of neurons for the network.

The RF algorithm is an extension of decision tree methods that divides the predictor variables in a dataset into homogenous groups that predict the response variable [261].

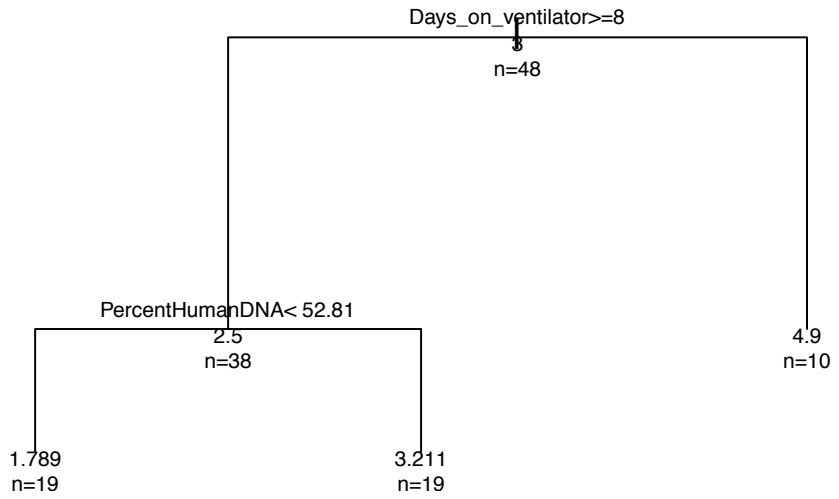

Figure 5.3: Decision Tree. PCA was used to cluster species abundance data for airway microbiota from burn patients and a decision tree was applied to determine which patient variables explain the clustering. The data was split by days spent on the ventilator and percent human DNA in the sample.

The predictors are divided by the process of recursive binary splitting, which is a top-down, greedy approach [257]. It is top-down in that splitting begins with all the predictors present, and it is greedy because the best choice for each specific split is made rather than the split that will generate a better tree for the entire data

set. Figure 5.3 gives an example of a decision tree applied to airway microbiome data. Here, principle components analysis (PCA) was used to cluster species abundance data by similarity. Clusters were assigned numbers and a decision tree was used to predict which patient clinical variables explain the clustering. The labels at each split indicate the variables that best predict the data, the numbers at each leaf represent the mean of the cluster assignments for each group, and the number of samples at each split is indicated by n=x. At the top of the tree, all the data was split based on the number of days the patient spent on the ventilator. Patients with less than eight days (right side of the top branch) were predicted to be in cluster five (mean of cluster assignments = 4.9). Patients with greater than or equal to eight days (left side of the top branch) were further split by the percent of human DNA present in the sample. If the sample had less than 52.81% human DNA (left side of the branch) the sample was predicted to be in cluster two (mean cluster assignment = 1.789). If the same sample had more DNA than that (right side of the branch) it was predicted to be in cluster three (mean cluster assignment = 3.211). To put it another way, if a patient spent less than eight days on the ventilator, their airway microbiota was predicted to be in cluster five. If they spent eight or more days on the ventilator and had less than 52.81% human DNA in their airway sample, their microbiota would be in cluster two. If they spent eight or more days on the ventilator and had 52.81% or more human DNA, their microbiota would be in cluster three. Decision trees are easy to interpret but their simplicity means that they lack the predictive accuracy and power of other classification methods. Further, because they simultaneously consider all the predictors, they can be biased towards variables that dominate the data. Aggregation of trees can greatly improve prediction accuracy and several methods exist to do this, of

which RF is the most powerful. In the RF model, a number of decision trees are built

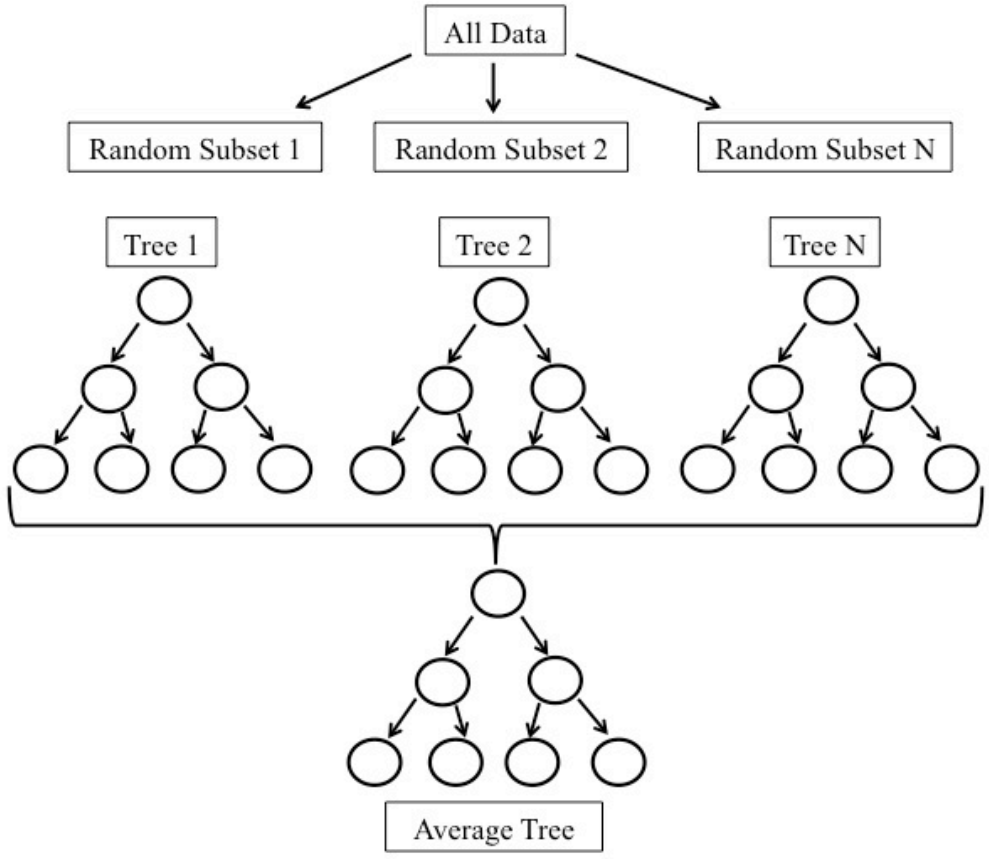

Figure 5.4: Random Forests. The random forest algorithm builds many decision trees based on a subset of the predictors in a dataset. The resulting tree is an average of these decorrelated trees and has improved accuracy and reliability over a single decision tree.

using bootstrapping, which consists of randomly resampling the data and is a powerful way of estimating variance. Unlike the decision tree method, however, RF only selects a subset of the predictors to make each tree. This randomization process eliminates bias that may be present due to dominance of certain predictors among the others. It also decorrelates the trees, which produces less variability and more reliability in the resulting average of the trees. Figure 5.4 illustrates how the RF model works. The other strength of the RF algorithm is its ability to deal with compositional, sparse data. Decorrelation of

the trees accommodates the spurious correlations that arise due to the nature of compositional data, and zeros in the data can be handled due to the method being non-parametric [262]. The RF algorithm is an improvement over decision tree methods and is a powerful predictive tool that can appropriately manage microbiome data.

Several studies have used these algorithms to predict outcomes or groupings based on microbiome data, such as subsistence groups in rural Africa from gut microbial community composition [26] and presence of pneumonia in patients based on upper respiratory tract microbiota [233]. However, the field is lacking validation of machine learning algorithms that accurately predict outcomes from microbiome data. Several studies have addressed this issue by comparing performance of several algorithms on microbiome data. Pasolli *et. al.* used support vector machines (SVM), random forests (RF), and lasso and elastic net regression algorithms to predict disease state among 2,424 publicly available metagenomic datasets [247]. RF produced the best classification accuracy as evaluated by area under the curve (AUC), a metric that summarizes true positive and false positive rates. The study also explored whether a reduced set of features (*e.g.* bacterial species) selected by each of the algorithms could be used to accurately predict disease state. Though feature selection produced good AUC, the best results were obtained with more than 60 bacterial species, indicating the importance of microbiota complexity in predicting outcomes. Statnikov *et. al.* applied 18 machine learning algorithms from seven families, including the four discussed above, to 1,802 human microbiome samples [249]. This work emphasizes the variation present in human microbiome datasets and the challenges they bring to classifying this data, particularly in the case of disease prediction. However, of the methods described here, Statnikov *et. al.*

found that SVM, RF, and ridge regression outperformed neural network classification. Pasolli *et. al.* found RF to be most accurate but used a different dataset, which will vary from Statnikov *et. al.*'s dataset, and did not include neural networks in the models they evaluated. Regardless of the difference in results, these studies confirm the utility of machine learning algorithms in predicting patient outcomes from microbiome data and establish RF, SVM, and ridge regression as appropriate models to use.

5.1.3 Application of Supervised and Unsupervised Methods to Burn Patient Airway Microbiota

We applied both unsupervised and supervised machine learning algorithms to metagenomic data from the airways of patients with burn and inhalation injury. Our goals were to find structure within the data in an unbiased manner using unsupervised methods and predict patient outcomes based on the microbiota with supervised methods. We used PCA, hierarchical clustering, and DAPC to identify clusters within the data. We used the DAPC cluster assignments as the response variable in the RF algorithm to determine which patient variables predict the clustering patterns. RF was also used to identify bacterial taxa that are predictive of development of $PaO_2/FiO_2 \leq 300$. This work demonstrates the utility of machine learning algorithms in predicting patient outcomes from high-dimensional metagenomic data.

5.2 Methods

5.2.1 Patient Samples and Sequencing

Burn patient samples were collected, processed, and sequenced as described in chapters 3 and 4. Briefly, DNA was extracted from samples taken within 72 hours of burn and inhalation injury. 16S rRNA gene amplicon sequencing was performed on the Illumina MiSeq using the molecule tagging (MT) method [215]. Samples were processed and OTU tables generated using the MT-Toolbox pipeline [217].

5.2.2 Unsupervised Clustering

Raw OTU count tables were normalized per patient in Explicet [263] and imported to R for further analysis [227]. The OTU table was called 'abundances' within R. Initial PCA was performed using the R package nsprcomp, which performs constrained PCA for sparse, non-negative data (NSPCA). The following code was used:

```
burn.nspca <- nsprcomp(abundances, nneg=TRUE, scale.=TRUE).
```

DAPC was done using the package adegenet:

```
grp <- find.clusters(abundances_t, max.n.clust=40)
dapc1 <- dapc(abundances_t, grp$grp)
scatter(dapc1)
```

Before continuing with clustering methods, the data was transformed according to the centered log ratio in order to address the compositional nature of the data [106]. For *K*-means clustering, the appropriate number of clusters was determined as follows:

```
wss <- (nrow(abundances)-1)*sum(apply(abundances,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(abundances, centers=i)$withinss)
```

plot(1:15, wss, type="b", xlab="Number of Clusters",ylab="Within groups sum of squares", main="Knee Method for Number of Clusters")

Clustering was then performed according to: fit <- kmeans(abundances, 3)

Hierarchical clustering was done as follows:

hc <- hclust(dist(abundances), method="ward")

The number of clusters for hierarchical clustering was determined after viewing the dendrogram, and the tree was cut as follows: rect.hclust(hc, k=3)

A hierarchical clustering-based heatmap was drawn using Manhattan distance and Ward clustering:

distance <- dist(abundances, method="manhattan")

cluster <- hclust(distance, method="ward.D2")

heatmap.2(abundances, Rowv=as.dendrogram(cluster), Colv=TRUE, scale="column", trace="none", col=redgreen, xlab="Taxa", ylab="Patient ID", margins=c(10,15))

Finally, a *K*-means clustering-based heatmap was drawn:

distance <- dist(abundances, method=”manhattan”)

fit <- kmeans(distance, 3)

heatmap.2(as.matrix(abundances)[order(fit$cluster),], Rowv=NA, Colv=NA, scale="none", trace="none", col=redgreen, xlab="Taxa", ylab="Patient ID", margins=c(10,15))

5.2.3 Supervised Random Forest Predictions

After unsupervised clustering was done, community assignments generated by DAPC were used as response variables in a random forest model to determine which patient variables predict the clustering. The patient’s PaO_2/FiO_2 ratio (≤ 300 or > 300) was used as a response variable in a separate RF model in order to determine which taxa

predict $PaO_2/FiO_2 \leq 300$. For both models the data was randomly split into a training set as follows:

```
choo_train <- abundances_t[sample(1:nrow(abundances_t), 38, replace=FALSE),]
```

The best model parameters were chosen based on the training data:

```
tune.ALI <- tune.randomForest(ALI~., data=choo_train, mtry=c(2.8, 5.6, 11.1),
ntree=c(250,500,1000), na.rm=TRUE)

x<-summary(tune.ALI)
```

The model was then run on the entire data set using the best parameters from the tuning procedure:

```
rf.ALI <- randomForest(ALI~., data=abundances_t, importance=TRUE,
na.action=na.omit, mtry=as.numeric(x$best.parameters[1]),
ntree=as.numeric(x$best.parameters[2]))
```

5.3 Results

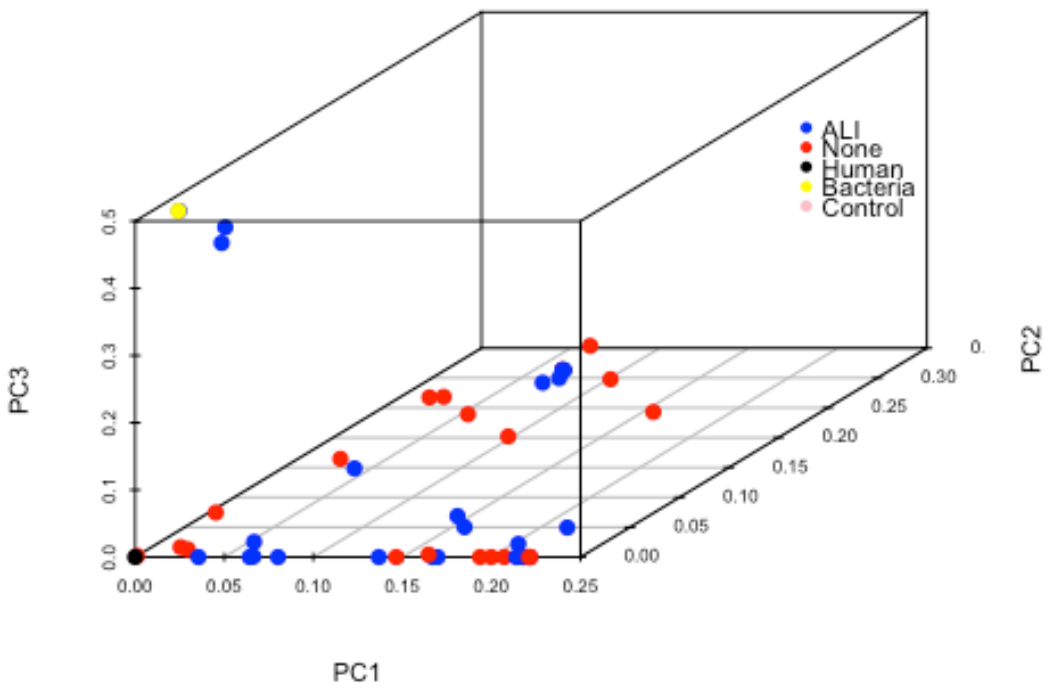

Figure 5.5: Sparse, Non-negative PCA Colored by Patient ALI Status. PCA analysis does not show clear clustering based on development of $PaO_2/FiO_2 \leq 300$ (ALI).

5.3.1 Clustering Trends

The first three components of the sparse, non-negative PCA are shown in Figure 5.5. Each point represents a patient sample and the closer each point is in the three-dimensional plot, the more similar they are. The points were colored by the patient's PaO_2/FiO_2 ratio (≤ 300 = ALI, > 300 = None) in order to determine whether sample similarity was driven by development of $PaO_2/FiO_2 \leq 300$ within 72 hours of injury. Figure 5.5 does not demonstrate clear clustering by $PaO_2/FiO_2 \leq 300$.

To determine *K*, the number of clusters for *K*-means clustering, we used what is referred to as the elbow method. Here, variance within the clusters is graphed against the

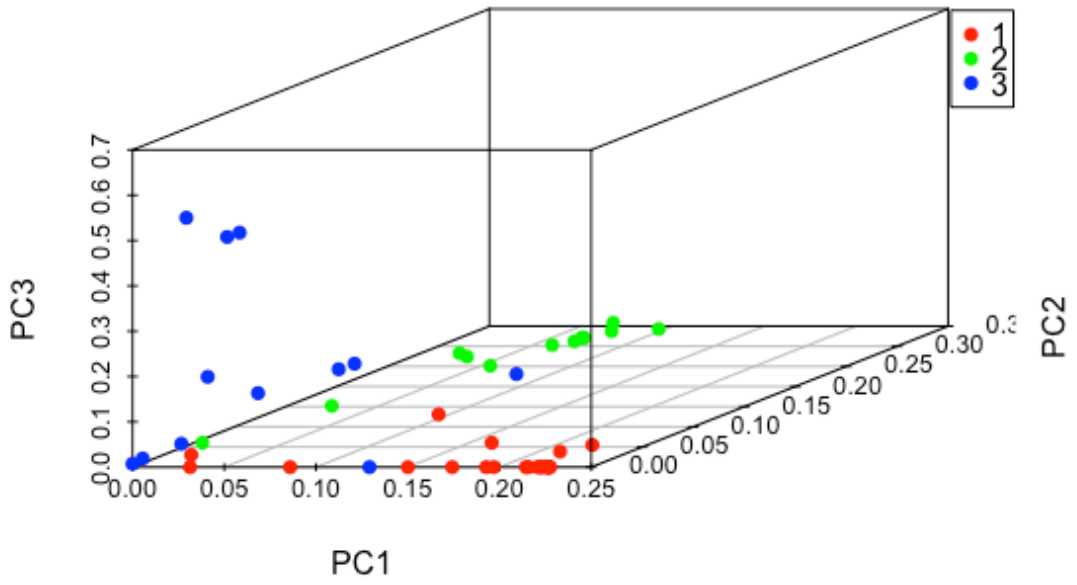

Figure 5.6: Cluster Assignments as Determined by Hierarchical Clustering. The NSPCA plot was colored by clusters identified through hierarchical clustering. 1, 2, and 3 refer to the clusters each sample was assigned to.

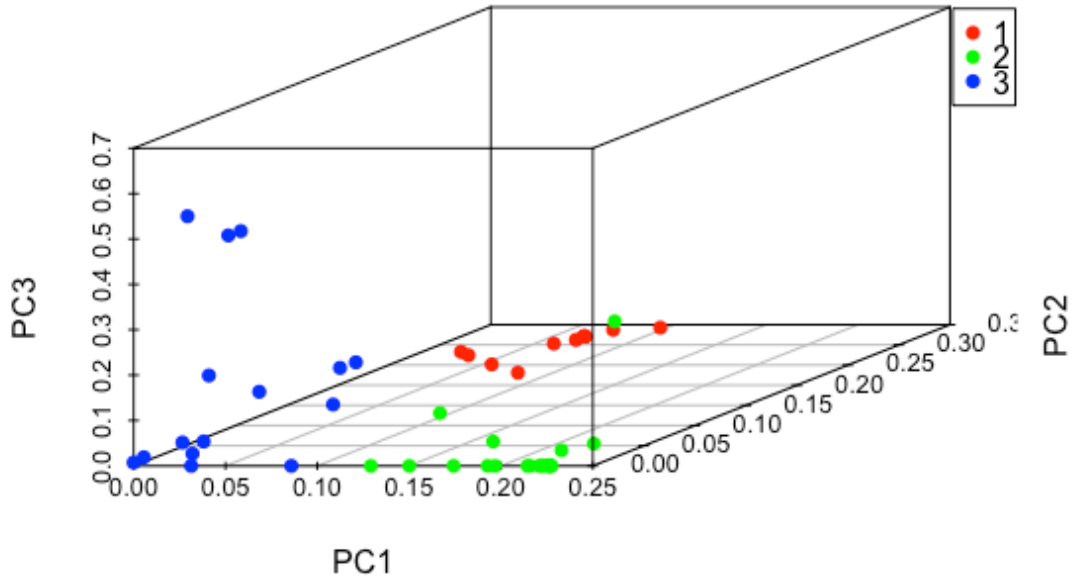

Figure 5.7: Cluster Assignments as Determined by *K*-means Clustering. The NSPCA plot was colored by clusters identified through *K*-means clustering. 1, 2, and 3 refer to the clusters each sample was assigned to.

number of clusters. The first few clusters will explain a large percentage of the variance, and at some point this levels off, giving an "L" or elbow-shaped graph. The bend in this graph is the optimal number of clusters that explains the variance between them. Using this method, we determined that three clusters were appropriate for both *K*-means and hierarchical clustering based on similarity of the samples (Figures 5.6 and 5.7). Though cluster assignment is similar between the two methods, *K*-means is more successful at assigning similar samples to the same cluster than hierarchical clustering is, indicating that *K*-means is a more appropriate method for this data set. The dendrogram produced by hierarchical clustering is shown in Figure 5.8 with the three clusters shown in red boxes. Each number corresponds to a patient sample.

Hierarchical clustering of both the patient samples and taxa abundance was used to produce the heatmap in Figure 5.9. The data has been transformed using the centered

log ratio, which corrects for the compositional nature of the data and any spurious correlations. Taxa abundances are represented by their Z score, or how many standard deviations they are above or below the mean. Taxa that are significantly more abundant than the mean are brighter green, while those that are significantly less than the mean are brighter red. Three clusters can be identified based on differences in abundance within Figure 5.9 and from the dendrogram on the vertical axis. The top-most cluster consists of patients whose microbiota are mostly high abundance, and a few patients with low-abundance microbota. Patients grouped into the middle cluster have a higher number of taxa represented at average abundance, while the last cluster at the bottom consists of patients with low-abundance microbiota. The heatmap in Figure 5.10 displays the same data as Figure 5.9 but here the patients are grouped according to their *K*-means clustering assignments. Though less distinct than the clusters in Figure 5.9, Figure 5.10 displays a similar pattern, with the low abundance group of patients clustered at the top of the heatmap, high abundance in the middle, and average abundance at the bottom. The average abundance group is not as clear as in Figure 5.9 and includes a mixture of some patients with high and low abundance taxa.

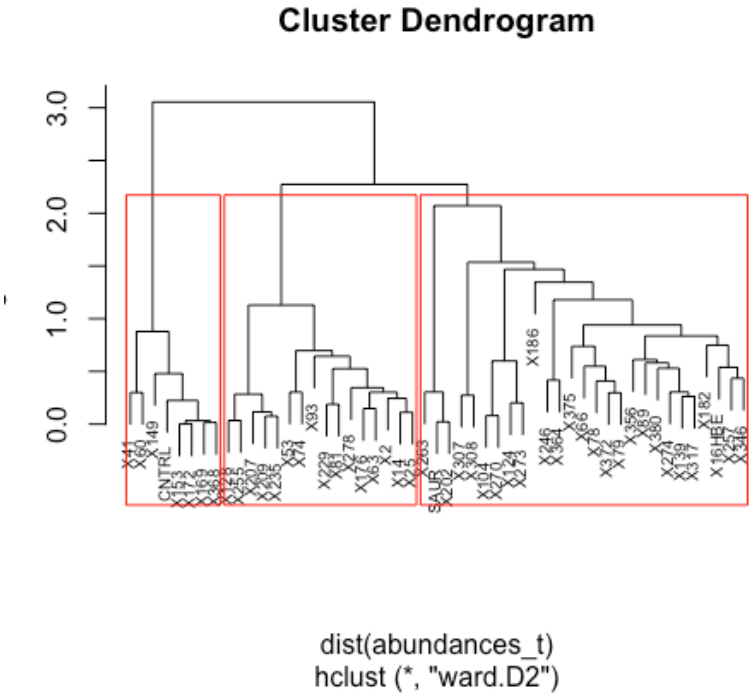

Figure 5.8: Hierarchical Clustering Dendrogram. Three clusters were identified by hierarchical clustering.

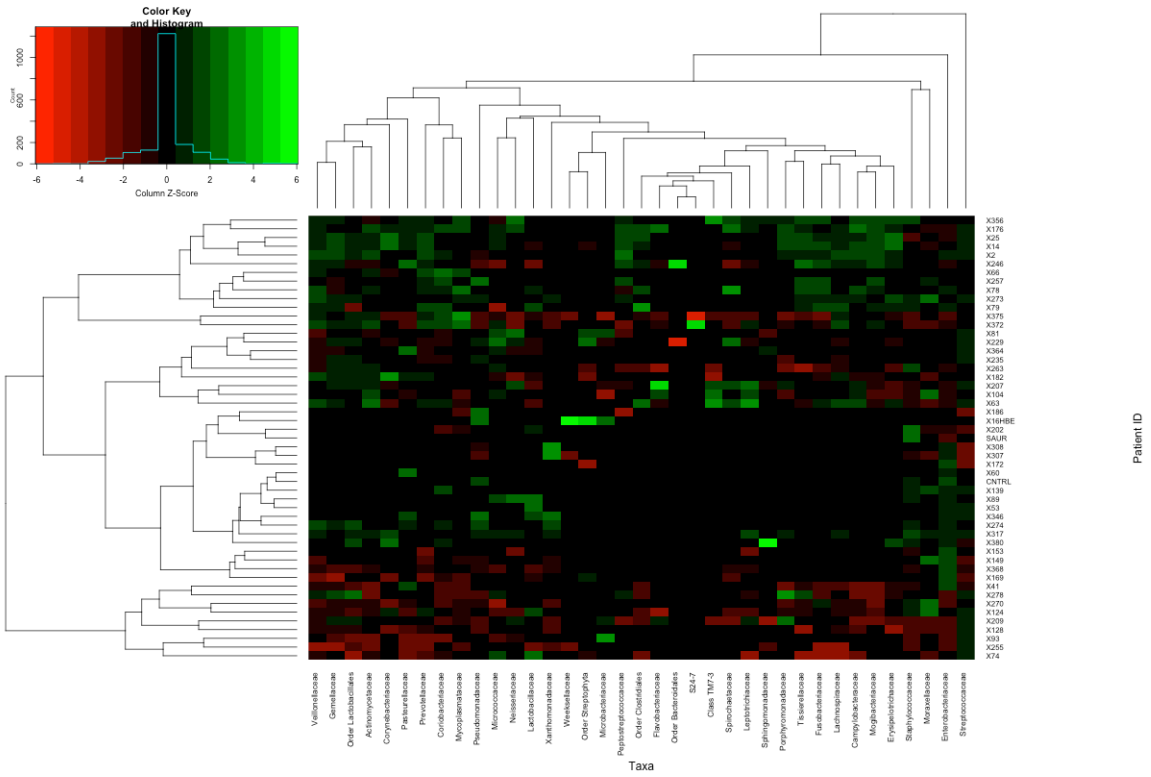

Figure 5.9: Hierarchical Clustering-Based Heatmap of Abundance Data. Data points which are brighter green represent taxa that are significantly more abundant than the mean while points which are brighter red are taxa which are significantly less abundant than the mean.

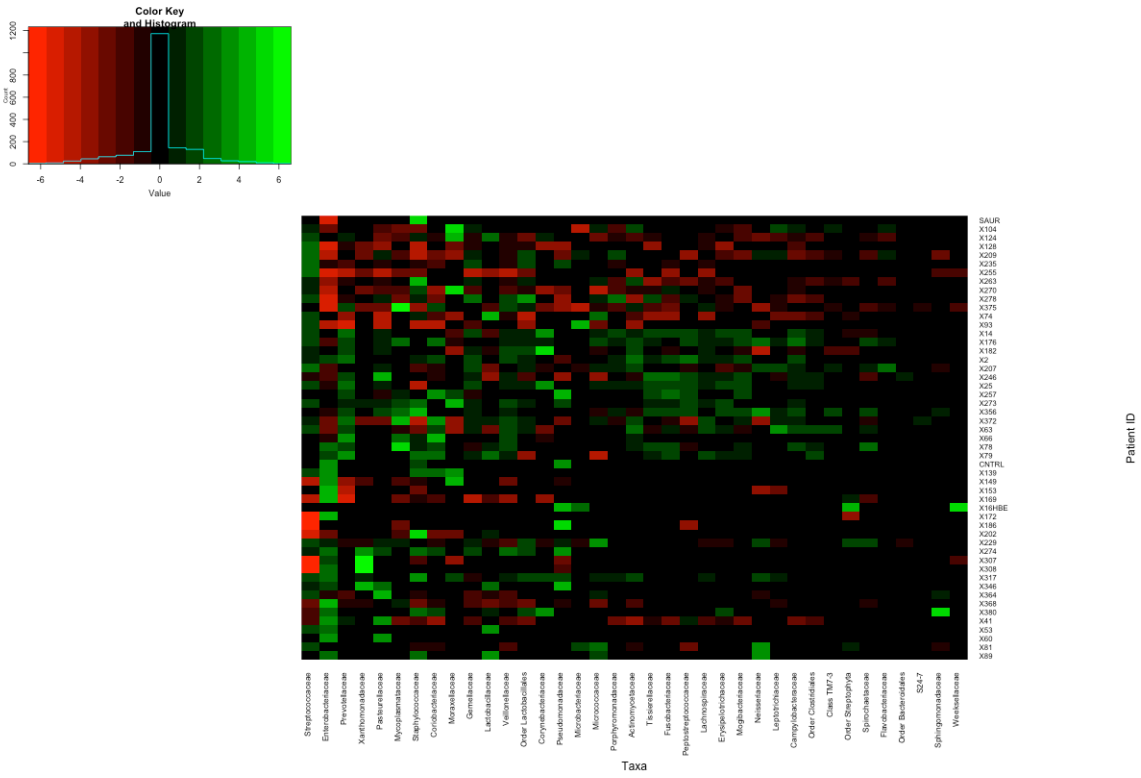

Figure 5.10: *K*-Means Clustering-Based Heatmap of Abundance Data. Data points which are brighter green represent taxa that are significantly more abundant than the mean while points which are brighter red are taxa which are significantly less abundant than the mean.

DAPC clustering, which incorporates *K*-means clustering with PCA and DA, is shown in Figure 5.11. In agreement with *K*-means and hierarchical clustering, DAPC identifies three clusters among the samples. Of the all the unsupervised clustering methods used here, DAPC provides the best grouping of the samples.

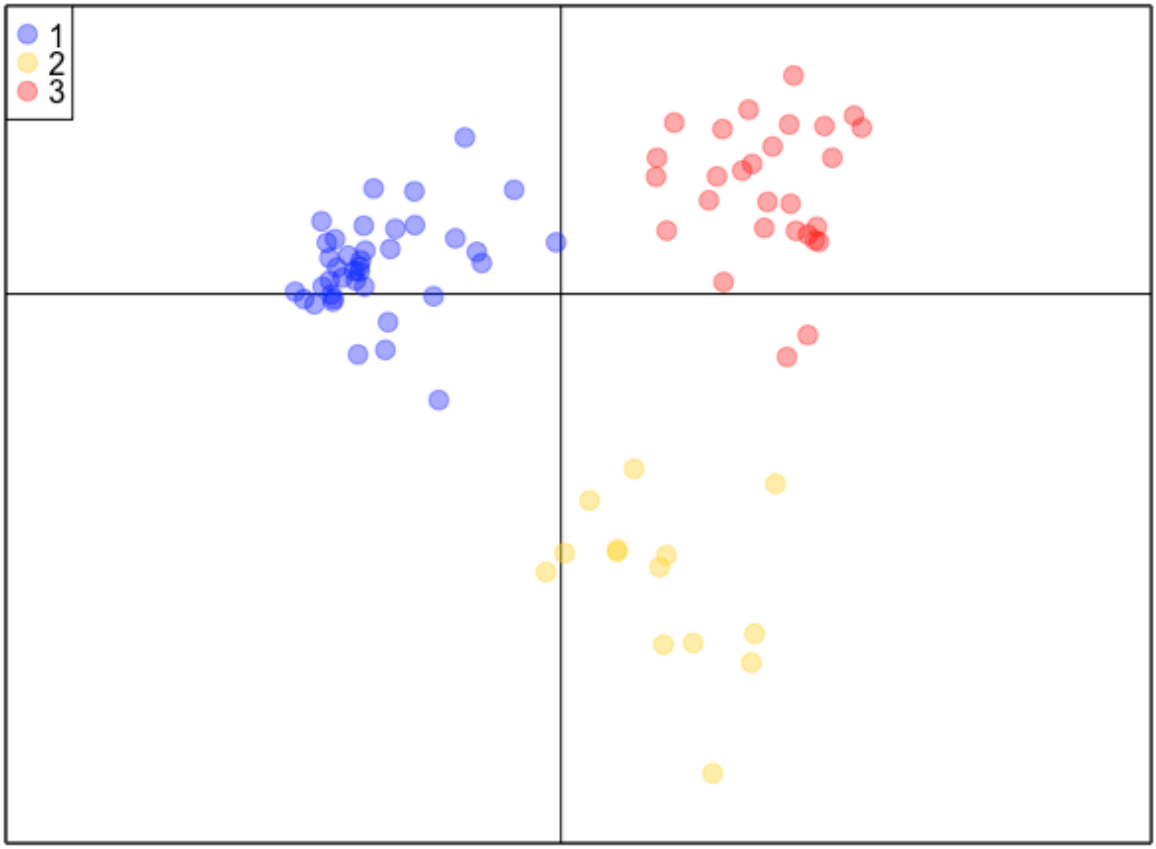

Figure 5.11: DAPC Identifies Three Clusters. The DAPC method found three distinct clusters among the data, in agreement with hierarchical and *K*-means clustering. 1, 2, and 3 refer to the clusters each sample was assigned to.

5.3.2 Random Forest Analysis

Due to the advantages of the DAPC method, cluster assignments generated by it were used as response variables in a RF model to predict which patient variables explain the clustering. Figure 5.12 shows the model output, which consists of ranking of the variables by the mean decrease they produce in the Gini index when removed from the model. Body mass index (BMI) was identified as most important in explaining sample clustering. Although there is not a set cut-off point for variable selection, it is typically done at the first largest change in variable importance. In Figure 5.12, the cut-off could

be after BMI, leaving a single variable for interpretation of the model, after Days on Ventilator, or after Age. While a single variable can make model interpretation too simplistic, including too many can make it overly complicated. Table 5.1 displays the average values for each cluster for the variables using Age as the cut-off point. Although BMI was ranked as the most important variable in determining sample clustering, the average BMI between the three clusters is similar. Cluster 1 contains the lowest average BMI while clusters 2 and 3 have very similar BMI values. Similarity among average BMI values supports our use of additional variables in explaining the model and expanding its complexity. The next most important variable in the model was Days on Ventilator. Patients in cluster 1 spent an average of 37.8 days on the ventilator, while those in cluster 2 spent 35.7 and those in cluster 3 spent 30.5. Cluster 1 also has the youngest patients on average, contained the lowest quantity of IL-12p70, the middle amount of IL-8, had the lowest Baux score, and the highest number of molecule tags (MTs). The average age of patients in cluster 2 was between clusters 1 and 3 but had the highest quantity of IL-12p70, lowest of IL-8 and MTs, and Baux scores between the other two clusters. Finally, cluster 3 had levels of IL-12p70 and MTs between clusters 1 and 2, the highest amount of IL-8, and the highest Baux score.

A RF model was also applied to taxa abundance in order to predict which species are associated with development of $PaO_2/FiO_2 \leq 300$. The model clearly identifies the Streptococcaceae family as most predictive of $PaO_2/FiO_2 \leq 300$ (Figure 5.13).

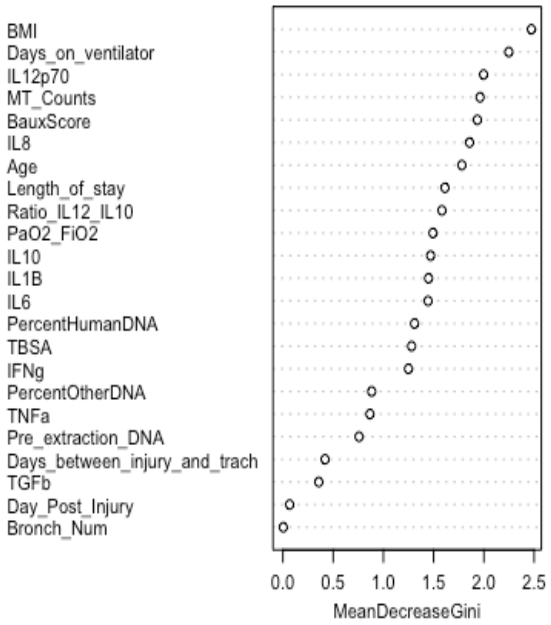

Figure 5.12: RF Analysis Identifies BMI as Most Predictive of Sample Clustering by DAPC. BMI was ranked as most predictive by node purity as quantified by the Gini index.

Cluster	BMI	Days on Ventilator	IL12p70 (pg/ml)	MT Counts	Baux Score	IL8 (pg/ml)	Age	Taxa Abundance
1	25.5	37.8	14.8	72,289.7	53.4	21,866.9	26.1	Low
2	27.7	35.7	23.3	14,775.2	60.0	17,438.9	27.6	Mean + high
3	27.3	30.5	20.4	64,912.7	66.9	43,559.4	48.8	High

Table 5.1: Average Value per Cluster Assignment

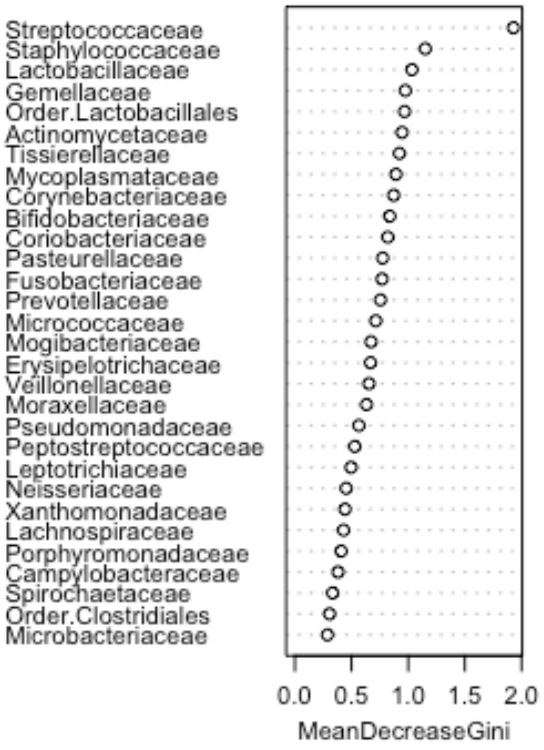

Figure 5.13: RF Analysis Identifies the Streptococcaceae Family as Most Predictive of ALI. Streptococcaceae is ranked as most predictive by the Gini index.

5.4 Discussion

High dimensional metagenomic data present statistical challenges that cannot be addressed with classical methods, requiring the development of sophisticated machine learning methods that can adapt to the needs of the study. Statistical learning methods like clustering and RF provide tools to handle large, complex data sets and allow prediction of output from input of this data [257]. Such methods allow unbiased examination of patterns among the data as well as accurate prediction of patient outcomes based on microbiota composition.

Analysis of burn patient airway microbiota with several different clustering methods revealed grouping based on similarity in species abundance among the patients. Although initial PCA showed loose clusters among the samples, they were not driven by development of $PaO_2/FiO_2 \leq 300$ (Figure 5.5). This implies that similarity among the samples is due to a more complex combination of patient variables, which is unsurprising given the heterogeneous nature of both the injury and the patient group.

We next applied hierarchical and *K*-means clustering without patient variable labels in order to discern the appropriateness of each method to forming unbiased clusters. For burn patient airway microbiota data, the appropriate number of clusters was determined to be three. For hierarchical clustering, pre-determination of the clusters is not necessary due to the "bottom-up" clustering technique used in the algorithm. Hierarchical clustering produced the dendrogram in Figure 5.8. "Cutting" the dendogram at different levels gives different numbers of clusters depending on the branches bisected. Because the elbow method can be applied to any clustering method to quantitatively assess the optimal number of clusters, we cut the dendrogram to produce three clusters. The cluster assignments from *K*-means and hierarchical clustering were then used to label points on the PCA graph. In Figure 5.6, hierarchical clustering misclassifies some of the points with other clusters. For example, a blue point that is clearly more similar to cluster two was assigned to cluster three, while several points that are closer to cluster three were assigned to clusters one and two. Assignment by *K*-means clustering in Figure 5.7 misclassifies only a single green point, which lies within red cluster one. This visual examination reveals that the *K*-means clustering algorithm is more successful in

determining similarity among the data points and is an appropriate method to use in analysis of the data.

Hierarchical clustering is the default method in the creation of heatmaps, which are commonly used to display gene expression changes. The graph clusters the patient samples in the same manner as Figure 5.8 and produces a dendrogram for taxa abundance as well (Figure 5.9). A second heatmap was generated using *K*-means clustering rather than hierarchical due to its appropriateness for this data set (Figure 5.10). In both figures, brighter green data points indicate increased abundance as compared to the mean, while brighter red points indicate decreased abundance. Unlike the PCA plots in Figures 5.6 and 5.7, the superiority of *K*-means over hierarchical clustering is less clear from the heatmaps. Although a high, average, and low abundance group can be found in both Figures 5.9 and 5.10, mixing of some high and low abundance taxa within each group make it difficult to determine which method more effectively clusters the data. From the PCA plot in Figure 5.7, *K*-means appears to be more appropriate than hierarchical clustering (Figure 5.6). However, the hierarchical clustering heatmap in Figure 5.9 produces more distinct clusters by taxa abundance than does the *K*-means heatmap in Figure 5.10. Based on these results, it is not clear which clustering method is more appropriate in determining similarity among the data points.

To address this issue, we re-clustered the data using the DAPC method. DAPC was designed to overcome the limitations of PCA and DA by combining them to produce a method in which dimensionality can be reduced and samples clustered without a loss of information [259]. Although both *K*-means and hierarchical clustering account for between- and within-sample variation, they do not reduce the dimensions of the data as

PCA does. In large metagenomic data sets, dimension reduction can be critical to interpretation of the data. The DAPC method incorporates both *K*-means clustering and PCA for dimensionality reduction followed by application of DA in order to determine which features are most discriminatory. DAPC's incorporation of multiple dimensionality reduction and clustering algorithms makes it more sophisticated than *K*-means or hierarchical clustering alone, giving it an advantage in dealing appropriately with our complex metagenomic data. DAPC sorted the burn patient samples into the three clusters shown in Figure 5.11, in agreement with the number of clusters as determined by the elbow method. However, unlike Figures 5.6 and 5.7, the clusters in Figure 5.11 have no overlap, confirming that DAPC is a better method for unsupervised clustering with this data set.

The DAPC clusters are grouped by maximizing between-group and minimizing within-group variation. This means that the microbiota of the burn patient samples in each of the three clusters is highly similar to others in the same group but maximally different from those assigned to the other groups. This implies an underlying difference in the microbiota that maximally divides them by their differences in three ways. We hypothesized that these groupings may be influenced by specific patient variables and applied a RF model to determine which these may be. RF is a supervised tree-based method used for regression and classification of data [257]. We chose RF over other machine learning methods due to its ability to deal with compositional data, computational efficiency, ability to extract relevant features from the model, and its proven appropriateness for metagenomic data sets [247,249,262,264]. Figure 5.12 displays the decrease in the Gini index, which is a measure of node purity and ranks

variable importance in classification by the model. The most important variable that has the largest decrease when removed is at the top in Figure 5.12. The RF model ranks BMI as the most important variable, which implies that BMI is the primary driver behind the three clusters generated by unsupervised methods. The remaining variables in the model are ranked by decreasing level of importance. Obesity, defined by the National Heart, Lung, and Blood Institute as a BMI over 30, and overweight, defined as a BMI between 25 and 29.9, have been linked to low-grade inflammation as well as a reduction in lung function in healthy individuals [49,265]. None of the average BMIs for each of the DAPC-defined clusters is over 30, but each falls within the overweight category. Based on obesity research, we might expect low-grade inflammation in each of these patient clusters as well as loss of lung function. Reduction in lung function can be indirectly inferred from the number of days spent on a respiratory ventilator, which is associated with increased severity of injury and susceptibility to infection [266]. The number of days spent on the ventilator was ranked as the second most important variable in the RF model, but was highest on average for cluster 1, which had the lowest BMI, and lowest for cluster 3, which had an average BMI between clusters 1 and 2. If an increase in BMI was directly associated with a decrease in lung function, we would expect patients with higher BMIs to spend longer on a ventilator. According to the averages for the clusters and the correlation between BMI and days on the ventilator, this is not true of this data set. Increasing BMI and more days on the ventilator are only weakly correlated (done in R as cor(bmi, vent, method= “spearman”, use= “complete”); correlation is 0.158). This study has a relatively low number of patients, which may be why the correlation is low. Expansion of the number of patients in future studies can be done to confirm whether or

not there is a correlation between BMI and days spent on the ventilator. We would also expect an increase in inflammation among patients with a higher BMI. This is difficult to interpret within burn patients since the burn injury itself induces an acute inflammatory response. However, the cytokines IL-12p70 and IL-8 were both ranked as important variables among the clusters and are both pro-inflammatory mediators. IL-8 expression is induced by a wide range of stimuli, from other cytokines to bacterial and viral products, and recruits neutrophils to the airways [267,268]. IL-12p70 is induced largely by viral and bacterial activation of TLR-4 and TLR-8 on dendritic cells and plays important roles in differentiation of Th1 cells [269,270]. Similar to BMI and the number of days on the ventilator, correlations between the levels of these two cytokines and BMI are small (-0.108 for IL-12p70 and 0.063 for IL-8). Again, inclusion of a larger number of patients in the study will confirm whether these correlations exist. The remaining variables ranked as important are Baux score, MT count, and age. The Baux score is the sum of the patient's age and size of burn injury and increases in this score, along with increases in age, have been associated with increased mortality [175]. Based on the average Baux score per cluster, patients in cluster 3 are predicted to have the worst outcome, followed by those in cluster 2 and then 1. This is the opposite of what might be predicted based on the days the patient spent on the ventilator. The MT count can be viewed as a measure of bacterial 16S rRNA abundance, as these are the sequences used to generate the OTU table. Based on this, cluster 1 has the highest overall abundance of bacteria, followed by cluster 3 and then 2. If we assume the healthy airways maintain very low amounts to no bacteria, we would assume that patients with higher MT counts would have poor outcomes and be at increased risk of infection and pneumonia. However, this does not fit

the pattern of the Baux score or days spent on the ventilator. Cluster 1, with the lowest average BMI, has the highest MT count, and cluster 2, with the highest average BMI, has the lowest MT count. If obesity is associated with an increased risk of lung function reduction and inflammation, generally indicating poorer health, we would hypothesize that patients with higher BMIs would have more bacteria in their airways. However, this is not the case with our data set. The patients within this study are heterogeneous, with a variety of co-morbidities, physical characteristics, and degree of burn and inhalation injury. This makes the data challenging to interpret and necessitates inclusion of a larger number of patients. Despite this limitation, the RF model identifies a subset of patient characteristics that may be important in influencing the airway microbiota after injury, providing future studies with specific patient variables monitor.

Examination of the clusters in Figure 5.10 reveals high abundance of the Streptococcaceae family and low abundance of Enterobacteriaceae in cluster 1, high abundance of Streptococcaceae and a mixture of high and low Enterobacteriaceae in cluster 2, and half the taxa in cluster 3 are present at average abundance with low abundance of Streptococcaceae and high abundance of Enterobacteriaceae. When taken in context with the patient variable data, this information paints a specific picture of both the airway microbiota and the patient characteristics that define these clusters. This information could serve as the basis for a framework to predict patient outcomes based on the clinical characteristics ranked as important as well as the composition of the airway microbiota. Further work is necessary to determine whether this classification holds true for a larger number of patients and if they can be linked to specific patient outcomes.

The RF model in Figure 5.13 specifically addresses the question of which taxa are predictive of $PaO_2/FiO_2 \leq 300$. Here, instead of patient variables, taxa abundance was input into the RF model with patient PaO_2/FiO_2 ratio as the response (≤ 300 or > 300). The Gini index quite clearly indicates that Streptococcaceae is most predictive, given its distance from the next most important variable. Although this result implies that Streptococcaceae is important in predicting $PaO_2/FiO_2 \leq 300$, it does not indicate whether high or low abundance of the family is most predictive. In Figure 5.10, clusters 1 and 2 contain high abundance of Streptococcaceae, while cluster 3 contains low abundance. In cluster 1, 58% of patients have $PaO_2/FiO_2 \leq 300$, and 60% of those in cluster 3 have $PaO_2/FiO_2 \leq 300$. In cluster 2, 38% of patients have $PaO_2/FiO_2 \leq 300$. From this data, it is not clear whether high or low abundance of Streptococcaceae predicts $PaO_2/FiO_2 \leq 300$. A larger sample size is needed to confirm these results and clarify the importance of Streptococcaceae in predicting $PaO_2/FiO_2 \leq 300$.

This study demonstrates the application of unsupervised clustering and supervised machine learning methods to high dimensional metagenomic data from a patient population. We used clustering methods to ascertain patterns in the data in an unbiased method and applied RF models to understand the clinical and biological relevance of the clustering. Although the predictive power of the RF algorithm is limited in this study due to small sample size, it highlights the ability of machine learning to make sense of high dimensional metagenomic data and make clinically relevant predictions. Machine learning algorithms can aid in determining the most biologically relevant dimensions of a study, producing specific questions that can be validated experimentally in the lab. We have shown that airway microbiota from burn patients taken within 72 hours of inhalation

injury produce three distinct clusters. These clusters are most significantly associated with patient BMI and days spent on the ventilator, as well as levels of IL-12p70, IL-8, Baux score, MT count, and age to a lesser degree. Finally, we show through a RF model that the Streptococcaceae family is most predictive of patient development of $PaO_2/FiO_2 \leq 300$ early after injury. This could serve as the basis of a larger study to establish a subset of patient clinical variables that predict airway microbiota composition and link changes in the bacteria to specific patient outcomes.

CHAPTER 6: PREDICTION OF FUNCTIONAL CHANGES AMONG BACTERIAL NETWORKS IN PATIENTS WITH $PAO_2/FIO_2 \leq 300$

6.1 Introduction

High-throughput NGS methods allow characterization of a range of bacteria present in various body and environmental locations as well as observation of how community membership can be altered by a disease or environmental event. Application of high-dimensional data analysis methods, such as machine learning, can sift through the large amount of data produced to identify taxa that may play important roles in a disease process or adaption to a change in environment. While study of individual taxa may lead to discovery of their specific functions, they ultimately exist within a complex community of other bacteria with which they may be interacting [271]. Within ecology, the relationships between organisms in a community can be defined on a spectrum of interactions. These range from mutualism, in which all organisms involved benefit, to commensalism, in which none benefit, to parasitism, where some benefit at the expense of the others [272]. Bacteria participate in each of these interactions, exchanging intermediate compounds with other species in order to make necessary amino acids, eliminating functions they can hijack from the host organism or other bacteria, or exchanging genetic information by lateral gene transfer that may or may not confer a benefit to either organism [271]. To understand whether these interactions are taking place in bacterial communities identified by metagenomic sequencing, methods have

been developed to detect co-occurrence through correlation networks. These methods largely depend on Spearman's or Pearson's correlations, with the strength of the correlation indicating the degree of association of the microbes and negative correlations indicating competition among them [248]. However, these methods are not well suited to metagenomic data, as they do not take the compositional nature of the data nor its sparseness into account [124]. Further, co-occurrence analysis only provides an observation of the frequency with which specific bacteria appear together in a community and does not directly indicate whether they are interacting on a functional level. A systems-based approach, incorporating WGS or 16S rRNA gene amplicon sequencing with other 'omics' techniques, such as metabolomics and transcriptomics, provides a more distinct picture of how the bacteria are interacting with each other, the host, and their environment [273]. Though the cost of conducting these studies continues to fall and computational methods for analysis continue to improve, large-scale meta-omics are not always feasible to employ.

To address these issues, we applied the community clustering algorithm SparCC [128] to abundance data from airway communities in patients with burn and inhalation injury. We then used a computational approach to predict bacterial functions from 16S rRNA gene sequencing data in order to understand potential functional changes within these communities based on whether or not the patient had hypoxia as indicated by $PaO_2/FiO_2 \leq 300$. SparCC, which stands for Sparse Correlations for Compositional Data, was developed to handle the sparseness present in all microbiome data as well as its compositional nature. SparCC addresses the compositionality of the data using a log transformation and then estimates the linear Pearson correlations [128]. When compared

to the Pearson's and Spearman's methods, SparCC produces fewer spurious and more accurate correlations.

To predict community functions we used phylogenetic investigation of communities by reconstruction of unobserved states (PICRUSt) [170]. Langille *et. al.* developed this computational tool to address the lack of functional information 16S rRNA gene sequencing provides. PICRUSt employs an algorithm that reconstructs gene families present based on available reference genomes that correspond to the taxa predicted by the 16S rRNA gene. Testing of this method against WGS data revealed its ability to accurately predict bacterial functions. This method provides a reliable technique to predict functions from 16S rRNA gene amplicon data, which can then be confirmed with further experiments.

We combined SparCC and PICRUSt in a community-wide, systems-based approach to functional changes in the airway microbiota in patients with $PaO_2/FiO_2 \leq 300$ after burn and inhalation injury. We used SparCC to predict bacterial interactions among patients with and without hypoxia based on taxa abundance and presence, applied PICRUSt to predict functional changes among these communities, and then employed the machine learning algorithm random forests (RF) to identify the functions most important to the community clustering. We find that, although there is not a clear difference based on abundance alone, each community has a distinct pattern of predicted functions. Further, RF analysis reveals that predicted functions most important in determining the community clustering in patients with and without hypoxia are distinct. Our work demonstrates that a community-wide approach can identify predicted functional changes among the communities as a whole and that application of machine learning algorithms

can be used to narrow this information down to identify important functions that can be pursued further in an experimental setting. Our approach focuses on overall changes among the community, rather than a few enriched individuals, providing a more complete picture of how these bacteria are interacting and influencing each other as well as host outcomes. Understanding community dynamics is crucial to designing effective therapeutic strategies to manipulate the microbiota and improve patient and environmental outcomes.

6.2 Methods

6.2.1 Patient Samples

Bronchial washings were collected from burn victims hospitalized at the North Carolina Jaycee Burn Center at the University of North Carolina Hospital. Within 24 hours of hospitalization, patients with suspected inhalation injury (II) underwent therapeutic bronchoscopy in order to flush soot and other debris from the airways. After clinical use of the sample, what remained was stored frozen in a repository. The repository was approved by the UNC Institutional Review Board (IRB) under study #10-0959. Consent for retaining samples was obtained from the patient or their legally authorized representative. After bronchoscopy, samples were placed on ice and processed within 72 hours. The washing was spun down and the pelleted cell fraction was stored separately from the supernatant at -80°C. Special permission from the IRB was obtained to use the pelleted portion of the sample to extract bacterial DNA (IRB #12-2475).

Samples were collected over a three-year period and de-identified before storage in the repository. Patient clinical and demographic data was also collected and stored in

the electronic Red Cap database. Information such as patient gender, race, comorbidities, clinical bacterial cultures, and measurement of inflammatory cytokines was collected and stored.

6.2.2 DNA Extraction and Sequencing

Burn patient samples were collected, processed, and sequenced as described in chapters 3 and 4. Briefly, DNA was extracted from samples taken with five days of burn and inhalation injury. 16S rRNA gene amplicon sequencing was performed on the Illumina MiSeq using the molecule tagging (MT) method [215]. Samples were processed and OTU tables generated using the MT-Toolbox pipeline [217].

6.2.3 Application of SparCC to Burn Patient Microbiome Data

Raw counts from the OTU table were normalized and log-transformed according to the SparCC method [128]. Two networks were made: one for OTU abundances from patients with $PaO_2/FiO_2 \leq 300$ and one for those from patients with $PaO_2/FiO_2 > 300$. SparCC returns positive and negative pairwise correlations for each of the 372 OTUs within each network ($PaO_2/FiO_2 \leq 300$ or $PaO_2/FiO_2 > 300$). We considered only positive correlations, limiting our analysis to mutually beneficial or neutral microbial interactions. The sparseness of the OTU table produces many edges, which we reduced by introducing a threshold that produces networks with one large component, and results in stability in the cluster assignments as well as the number of clusters. For a range of thresholds, we compared similarities in community (**Z)** assignments for adjacent thresholds (t-1 and t), (**Z**_t-1, **Z**_t). To compare them, we computed the normalized mutual information (NMI).

So, NMI($\mathbf{Z}_{t-1}$, $\mathbf{Z}_t$). We would like to choose a threshold where the NMI between adjacent community assignments is close to 1 (meaning it is not changing as we vary the threshold). The NMI for the $PaO_2/FiO_2 \leq 300$ network is shown in Figure 6.1 and that for the $PaO_2/FiO_2 > 300$ network in Figure 6.2.

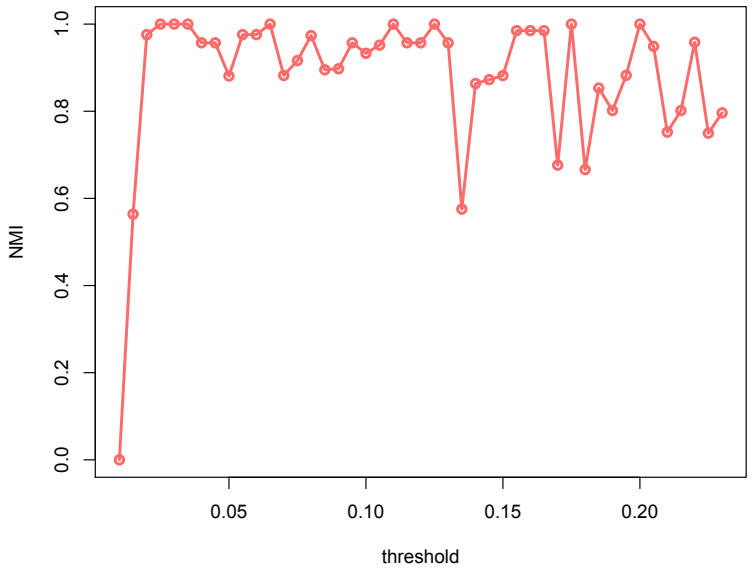

Figure 6.1. NMI Between Adjacent Threshold Points in the $PaO_2/FiO_2 \leq 300$ Network. The NMI was plotted for each threshold value. It was close to 1 between thresholds of 0.01 and 0.14.

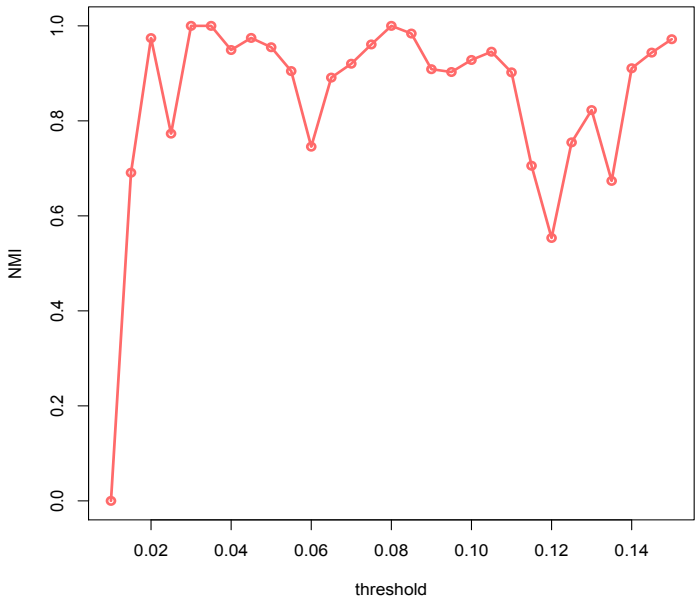

Figure 6.2. NMI Between Adjacent Threshold Points in the $PaO_2/FiO_2 > 300$ Network. The NMI was plotted for each threshold value. The NMI is less stable than for the $PaO_2/FiO_2 \leq 300$ network but is close to 1 between thresholds of 0.01 and 0.10.

In order to ensure stability of the number of clusters found by SparCC, we also plotted the number of communities as a function of threshold. These are shown in Figures 6.3 and 6.4. For both the $PaO_2/FiO_2 \leq 300$ and $PaO_2/FiO_2 > 300$ networks, a threshold of 0.14 produced stable community assignment with an NMI close to 1. Therefore, we used a threshold of 0.14 for final community assignment. Abundance heatmaps were made using R for each network.

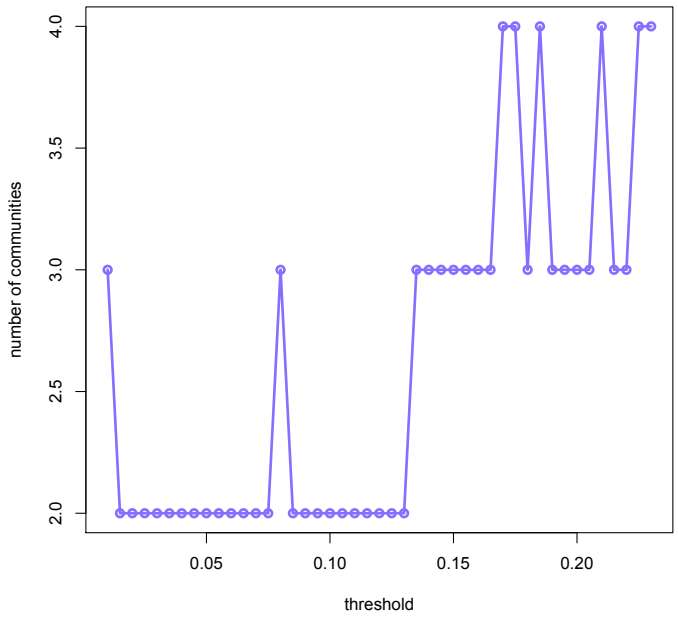

Figure 6.3: Number of Communities vs. Threshold for the $PaO_2/FiO_2 \leq 300$ Network. The number of communities was plotted per threshold value. A threshold of 0.14 produces a stable number of communities and the NMI is near 1 (Figure 6.1). Therefore, we used a threshold of 0.14 for the $PaO_2/FiO_2 \leq 300$ network.

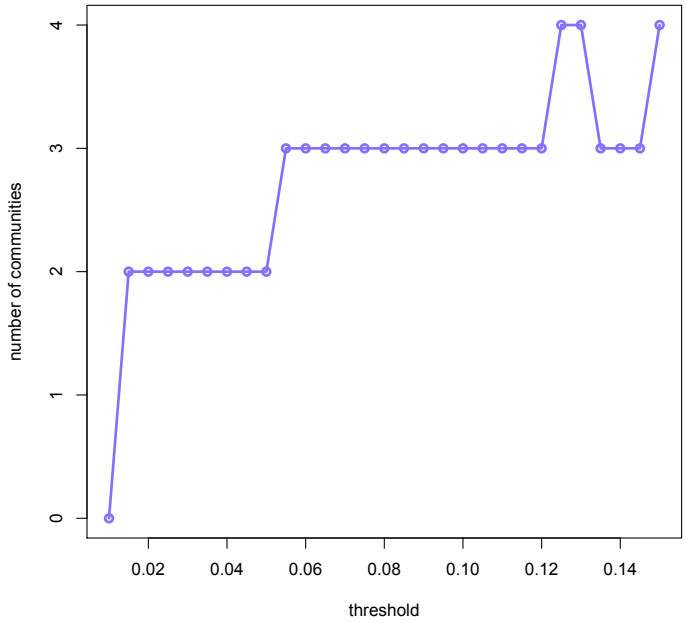

Figure 6.4: Number of Communities vs. Threshold for the $PaO_2/FiO_2 > 300$ Network. The number of communities was plotted per threshold value. A threshold of 0.14 produces a stable number of communities here, as for the $PaO_2/FiO_2 \leq 300$ network. At this threshold the NMI for the $PaO_2/FiO_2 > 300$ network is also near 1 (Figure 6.2). Therefore, we used a threshold of 0.14 for both the $PaO_2/FiO_2 \leq 300$ and $PaO_2/FiO_2 > 300$ networks.

6.2.4 Predicted Functional Gene Content of the Airway Microbiome

We implemented PICRUSt as detailed by Langille *et. al.* [170]. Briefly, raw OTU counts were normalized and OTU numbers matched to those found within the Greengenes database [219]. OTUs with matches were kept while those that did not match were removed from the table. Matching OTUs were corrected for 16S rRNA copy number differences and a predicted gene table was produced by multiplying the normalized OTU table by the gene content predictions. This table was imported into R [227], where predicted functions for each OTU were matched to their community assignment based on the SparCC networks. Heatmaps were produced using R to identify predicted differences in functions for each of the four communities identified by SparCC within patients with and without $PaO_2/FiO_2 \leq 300$. Manhattan distance and Ward clustering were used to create the heatmap dendrograms.

6.2.5 Use of Machine Learning to Predict Functions Associated with Networks

A random forest algorithm was used to identify which predicted functions were most important in the SparCC community assignments. The package used was modified by Liaw and Wiener for implementation in R [274]. Before running the model, variable selection of the predicted functions was necessary as they outnumbered the patient samples. There were 328 functions with variability greater than 0.05 for communities in patients with and without $PaO_2/FiO_2 \leq 300$. These were chosen for use in the RF model. The model was trained on a random subset of half the data and tested on the entire data set. This was done separately for communities in patients with and without $PaO_2/FiO_2 \leq 300$. The code was written as follows:

Training set:

```
set.seed(10)
choo_train.3 <- ALI.rf[sample(1:nrow(ALI.rf), nrow(ALI.rf)*0.5, replace=FALSE),]
tune <- tune.randomForest(ALI.communities~., data=choo_train.3, mtry=c(2.8, 5.6,
11.1), ntree=c(250,500), na.rm=TRUE)
x.3<-summary(tune)
```

Testing set:

```
rf.ALI.funct <- randomForest(ALI.communities~., data=ALI.rf, importance=TRUE,
na.action=na.omit, mtry=as.numeric(x.3$best.parameters[1]),
ntree=as.numeric(x.3$best.parameters[2]))
```

6.3 Results

6.3.1 OTU Networks Among Patients

Before applying the SparCC method to our data, we needed to determine an appropriate threshold. All metagenomic data tends towards sparsity, as a large number of OTUs may be identified but they may have many zero counts per sample [107]. SparCC creates additional edges in data sets with increasing sparseness, resulting in a dense community network that is difficult to interpret. To reduce these edges, we chose a threshold at which the NMI was close to one, indicating stability. NMI is a measure of dependence between two variables, or how much information about one variable can be obtained through its dependence on the other [107,261]. We plotted the NMI against a range of thresholds for networks among patients with and without $PaO_2/FiO_2 \leq 300$ (Figures 6.1 and 6.2). We also wanted to choose a threshold that produced stability in the number of communities, so we plotted this against each threshold as well (Figures 6.3 and 6.4). Setting a threshold of 0.14 for communities within patients with and without $PaO_2/FiO_2 \leq 300$ resulted in four community clusters for each network. $PaO_2/FiO_2 \leq 300$

clusters are shown in Figure 6.5 and $PaO_2/FiO_2 > 300$ clusters in Figure 6.6. Each node in both networks indicate individual OTUs while the length of the edge between them indicates the strength of the positive correlation, which could be interpreted as a mutually beneficial interaction. The nodes in Figure 6.5 are colored according to the community cluster they were assigned to. We maintained these colors in Figure 6.6 in order to show the change in OTU correlations in patients without hypoxia. The networks have dissimilar overall shapes, indicating differences in positively correlated relationships among OTUs in patients with and without $PaO_2/FiO_2 \leq 300$.

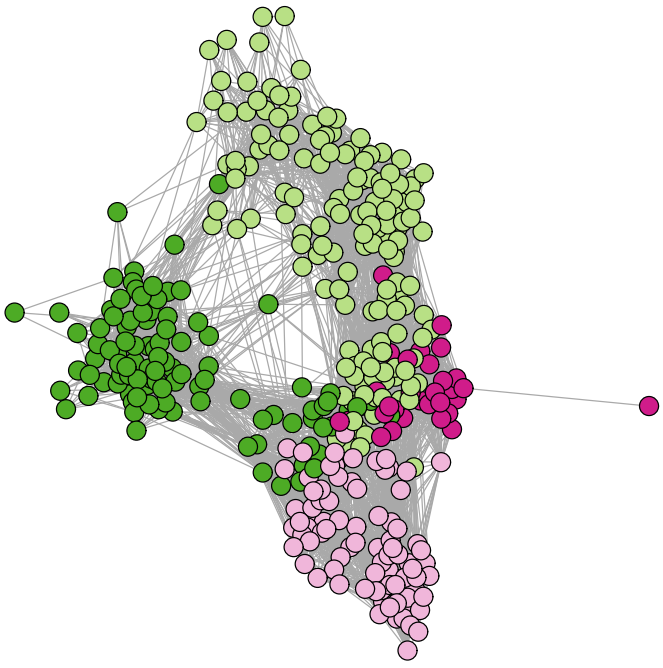

Figure 6.5: $PaO_2/FiO_2 \leq 300$ Community Clusters Identified by SparCC. A threshold of 0.14 identifies four community clusters among microbiota of patients with $PaO_2/FiO_2 \leq 300$ (hypoxia).

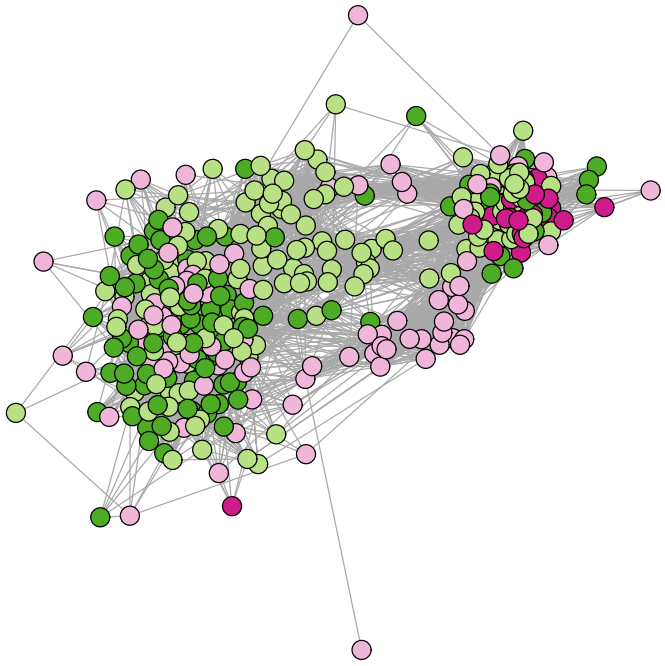

Figure 6.6: PaO_2/FiO_2 > 300 Community Clusters Identified by SparCC. A threshold of 0.14 identifies four community clusters among microbiota of patients with PaO_2/FiO_2 > 300 (no hypoxia).

Figures 6.7 and 6.8 show abundance of the 372 OTUs per patient. The OTUs are ordered based on their SparCC community assignment. Figure 6.7 shows no clear patterns in OTU abundance by community in patients with $PaO_2/FiO_2 \leq 300$. Figure 6.8 indicates that community A in patients with $PaO_2/FiO_2 > 300$ contains OTUs with mostly average abundance, but there is no clear pattern among the other communities.

Table 6.1 is a contingency table showing the number of OTUs that overlap between the community networks. This indicates OTUs that are shared between the communities within the two networks. $PaO_2/FiO_2 \leq 300$ community 1 has the most overlap with $PaO_2/FiO_2 > 300$ community C, but the least of all with $PaO_2/FiO_2 > 300$ communities A and B. $PaO_2/FiO_2 \leq 300$ 2 has high overlap with $PaO_2/FiO_2 > 300$ A and D and less with $PaO_2/FiO_2 > 300$ B and C. $PaO_2/FiO_2 \leq 300$ 3 has the greatest total overlap of all the communities, sharing the highest number of OTUs with $PaO_2/FiO_2 >$

300 A, followed by B and C. It has low overlap with $PaO_2/FiO_2 > 300$ D. Finally, $PaO_2/FiO_2 \leq 300$ 4 and $PaO_2/FiO_2 > 300$ A have the highest number of OTUs in common. $PaO_2/FiO_2 \leq 300$ 4 has lower overlap with $PaO_2/FiO_2 > 300$ communities B through D.

	P/F > 300 A	P/F > 300 B	P/F > 300 C	P/F > 300 D
P/F ≤ 300 1	1	1	27	2
P/F ≤ 300 2	50	5	9	22
P/F ≤ 300 3	66	46	37	5
P/F ≤ 300 4	77	5	15	4

Table 6.1: OTU Overlap within $PaO_2/FiO_2 \leq 300$ (P/F ≤ 300) and $PaO_2/FiO_2 > 300$ (P/F > 300) Communities.

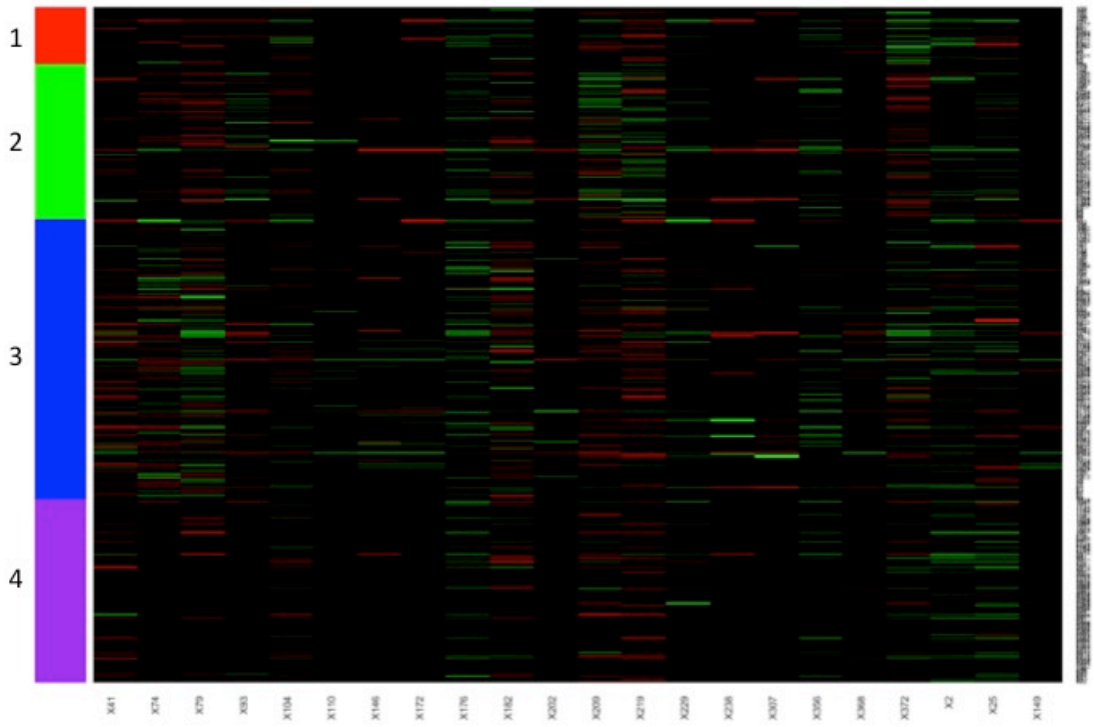

Figure 6.7: $PaO_2/FiO_2 \leq 300$ Communities Abundance Heatmap. A heatmap of OTU abundance per community was created for the $PaO_2/FiO_2 \leq 300$ networks. OTUs are listed on the vertical axis and patient numbers across the horizontal axis. The colored bar on the left side of the graph indicates the community SparCC assigned each OTU to. Abundances are displayed as standard deviation from the mean, with brighter green indicating higher abundance than the mean and brighter red indicating lower abundance than the mean. $n = 22$ patients.

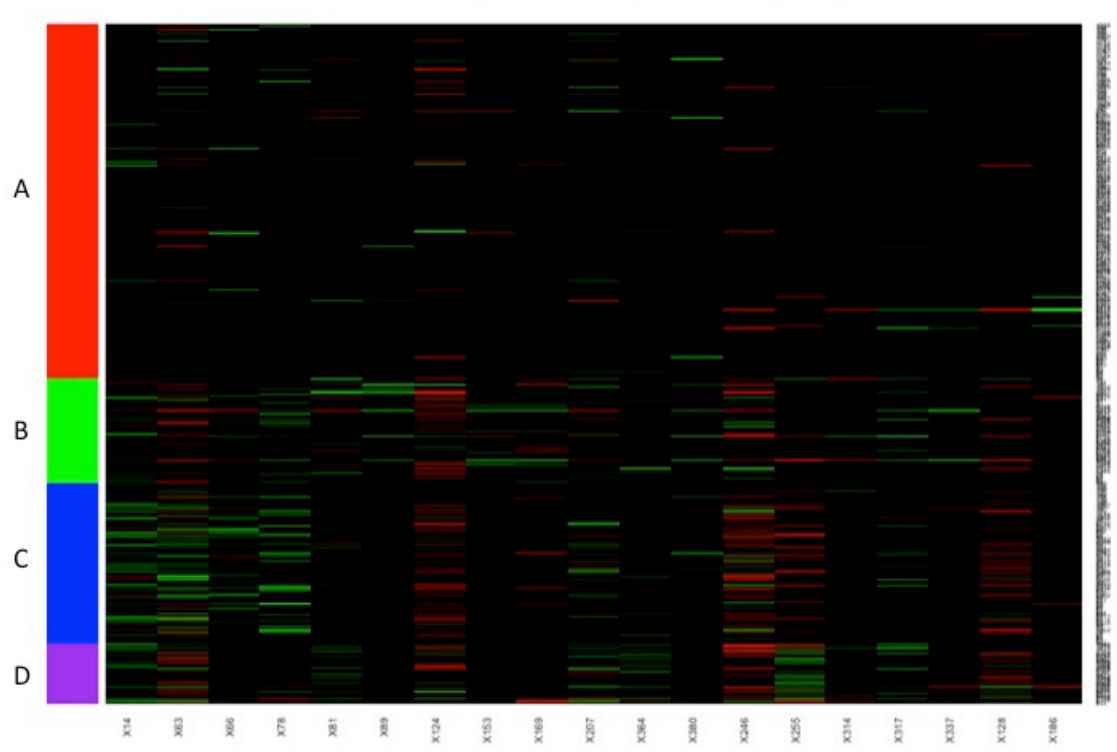

Figure 6.8: PaO_2/FiO_2 > 300 Communities Abundance Heatmap. A heatmap of OTU abundance per community was created for the PaO_2/FiO_2 > 300 networks. OTUs are listed on the vertical axis and patient numbers across the horizontal axis. The colored bar on the left side of the graph indicates the community SparCC assigned each OTU to. Abundances are displayed as standard deviation from the mean, with brighter green indicating higher abundance than the mean and brighter red indicating lower abundance than the mean. n = 19 patients.

6.3.2 Highly Represented Predicted Gene Functions

Heatmaps were created to visualize changes in predicted functions for each community. Figures 6.9 – 6.12 represent changes in predicted functions of the communities in the $PaO_2/FiO_2 \leq 300$ network. Hierarchical clustering was used to group the OTUs (on the vertical axis) and the predicted functions (on the horizontal axis). While the dendrograms for the OTUs demonstrate different clustering per community with distinguishable clusters, clustering by predicted function is not as clear and is similar among all communities. Overall, patterns of over- and under-abundant predicted

functions are different for each community. $PaO_2/FiO_2 \leq 300$ community 1 appears to contain more predicted functions with lower than average abundance (Figure 6.9). This community also contains a larger group of predicted functions with higher than average abundance than the other $PaO_2/FiO_2 \leq 300$ communities. $PaO_2/FiO_2 \leq 300$ communities 2 through 4 contain groups of functions with both higher and lower than average abundance but not to the degree that $PaO_2/FiO_2 \leq 300$ community 1 does. This is also the community with a single OTU in common with $PaO_2/FiO_2 > 300$ communities A and B (Figures 6.13 and 6.14), two with $PaO_2/FiO_2 > 300$ D (Figure 6.16), but 27 with $PaO_2/FiO_2 > 300$ C (Figure 6.15). Comparison of the heatmaps of these predicted functions show distinct patterns of abundance despite the similarities or differences in OTUs. Figures 6.13 – 6.16 represent changes in predicted functions of the communities in the $PaO_2/FiO_2 > 300$ network. As with the $PaO_2/FiO_2 \leq 300$ networks, the $PaO_2/FiO_2 > 300$ predicted functions were ordered by hierarchical clustering in Figures 6.13 – 6.16. Clustering of the predicted functions across the horizontal axis is similar among the $PaO_2/FiO_2 > 300$ communities but distinct clusters are difficult to distinguish. Clustering of the OTUs across the vertical axis reveals different patterns of clustering per community with clearer groupings than among the predicted functions. Similar to the $PaO_2/FiO_2 \leq 300$ communities, the patterns of predicted function abundance are different per community. $PaO_2/FiO_2 > 300$ community A contains predicted functions with mostly average abundance, while communities B and C have some groups of both over- and under-abundance predicted functions. $PaO_2/FiO_2 > 300$ community D has an interesting pattern of a few functions with higher than average predicted abundance and a few others with lower than average predicted abundance but the majority with average predicted

abundance. All of the communities in the $PaO_2/FiO_2 \leq 300$ and $PaO_2/FiO_2 > 300$ networks display distinct patterns of predicted function abundance.

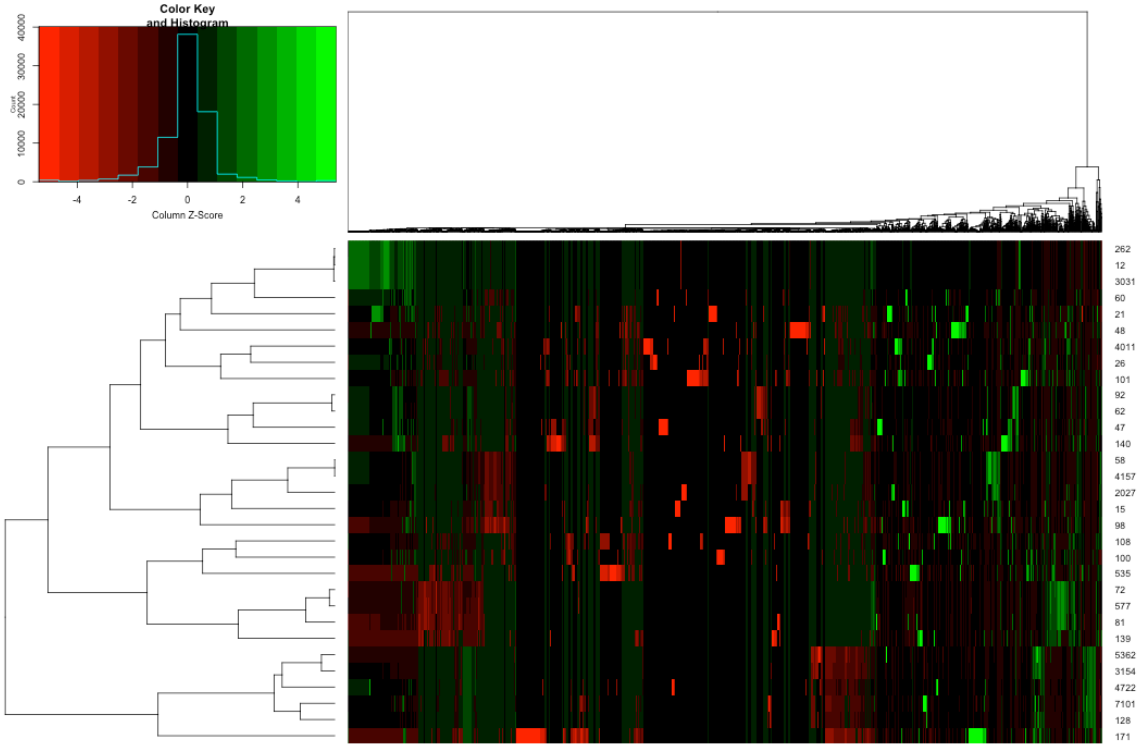

Figure 6.9: Community 1 Predicted Functions. A heatmap was made to show changes in predicted functions among OTUs assigned to community 1 in the $PaO_2/FiO_2 \leq 300$ network. OTU numbers are listed across the vertical axis and functions across the horizontal axis. Brighter green indicates functions with greater prevalence than the mean and brighter red indicates functions with less prevalence than the mean.

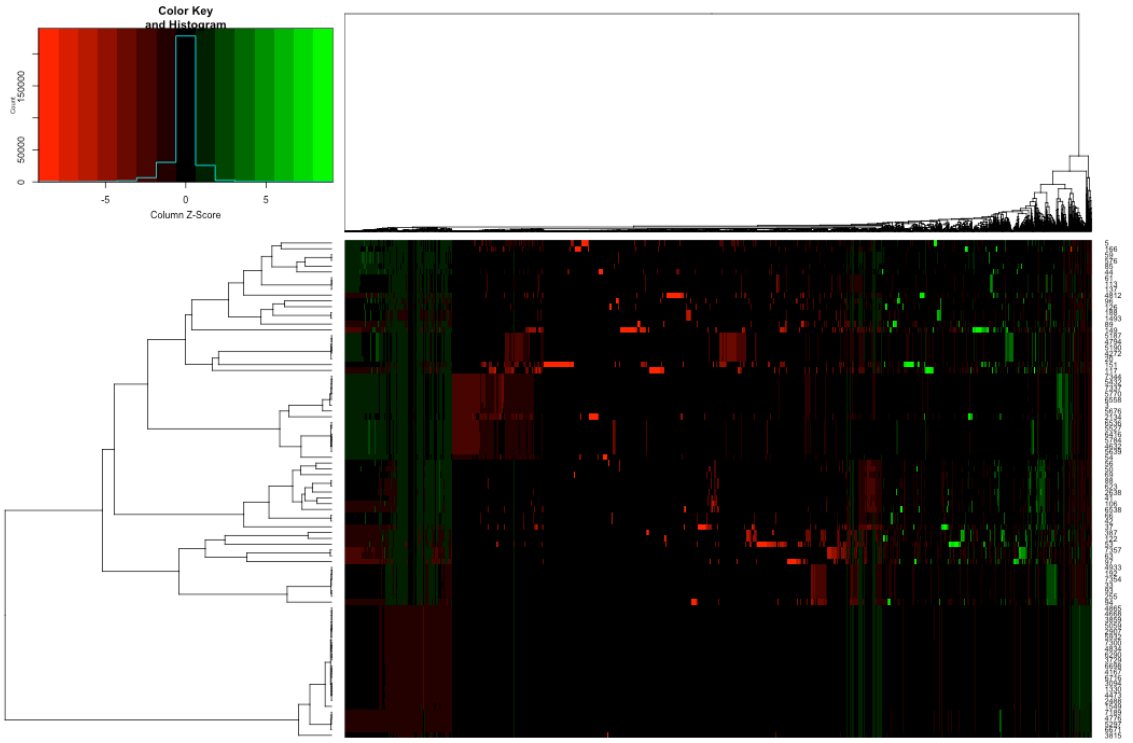

Figure 6.10: Community 2 Predicted Functions. A heatmap was made to show changes in predicted functions among OTUs assigned to community 2 in the $PaO_2/FiO_2 \leq 300$ network. OTU numbers are listed across the vertical axis and functions across the horizontal axis. Brighter green indicates functions with greater prevalence than the mean and brighter red indicates functions with less prevalence than the mean.

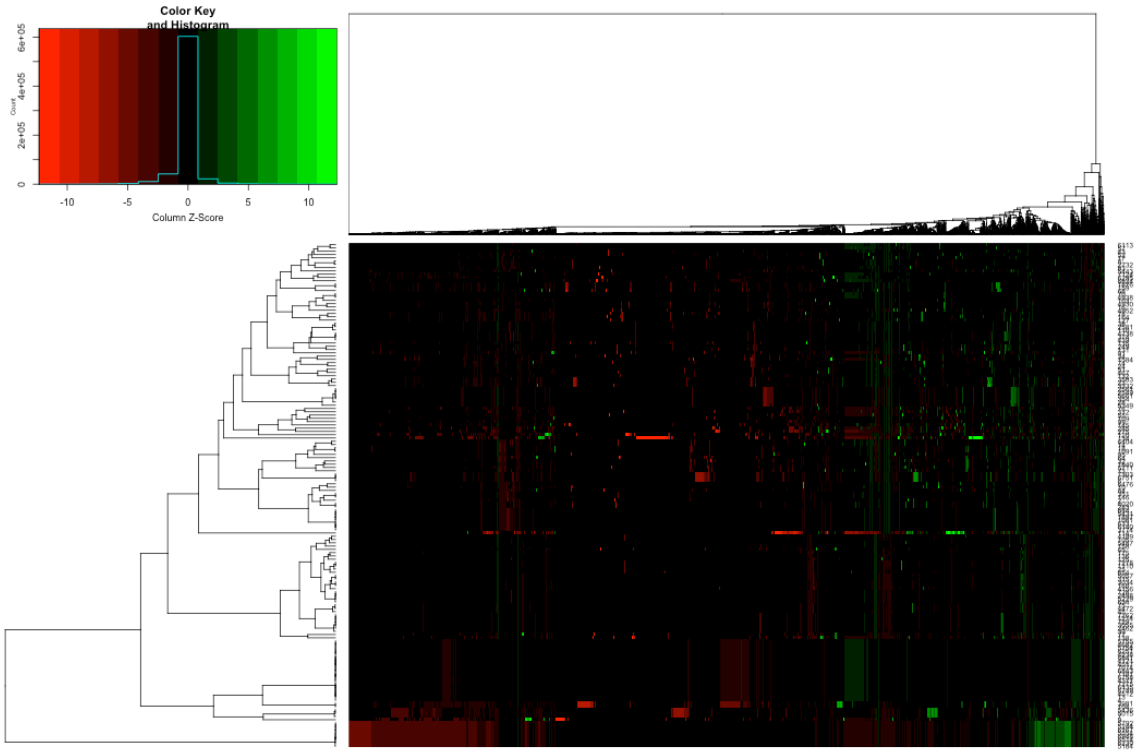

Figure 6.11: Community 3 Predicted Functions. A heatmap was made to show changes in predicted functions among OTUs assigned to community 3 in the $PaO_2/FiO_2 \leq 300$ network. OTU numbers are listed across the vertical axis and functions across the horizontal axis. Brighter green indicates functions with greater prevalence than the mean and brighter red indicates functions with less prevalence than the mean.

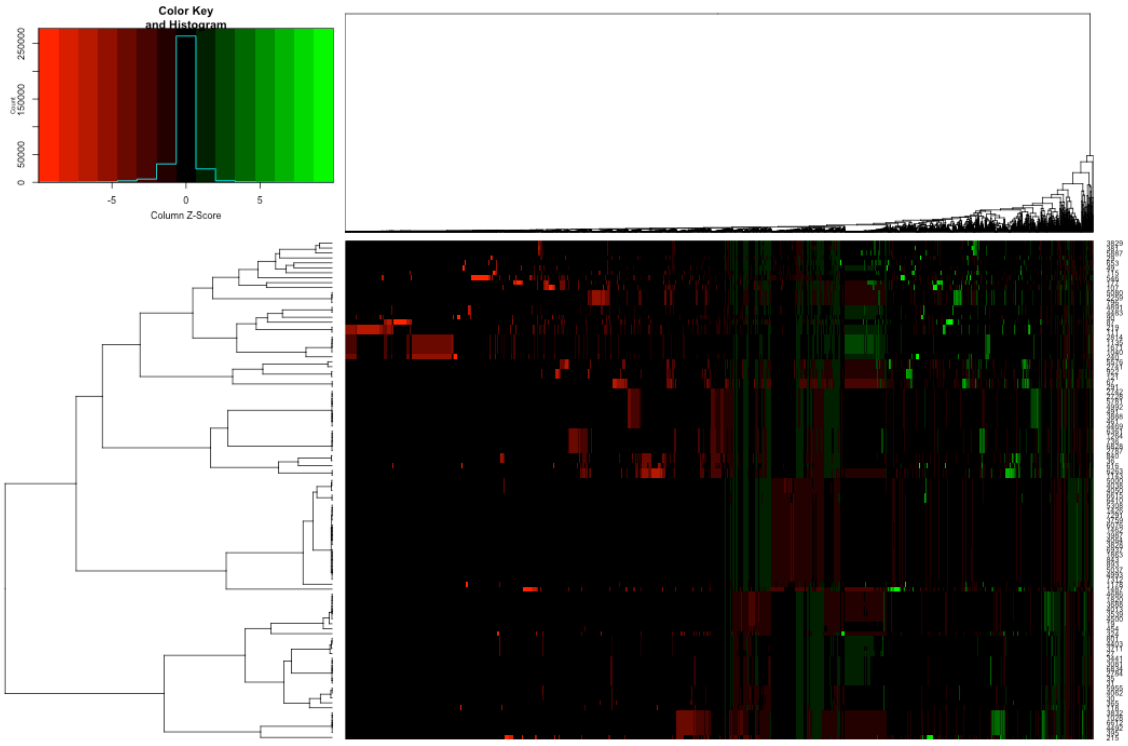

Figure 6.12: Community 4 Predicted Functions. A heatmap was made to show changes in predicted functions among OTUs assigned to community 4 in the $PaO_2/FiO_2 \leq 300$ network. OTU numbers are listed across the vertical axis and functions across the horizontal axis. Brighter green indicates functions with greater prevalence than the mean and brighter red indicates functions with less prevalence than the mean.

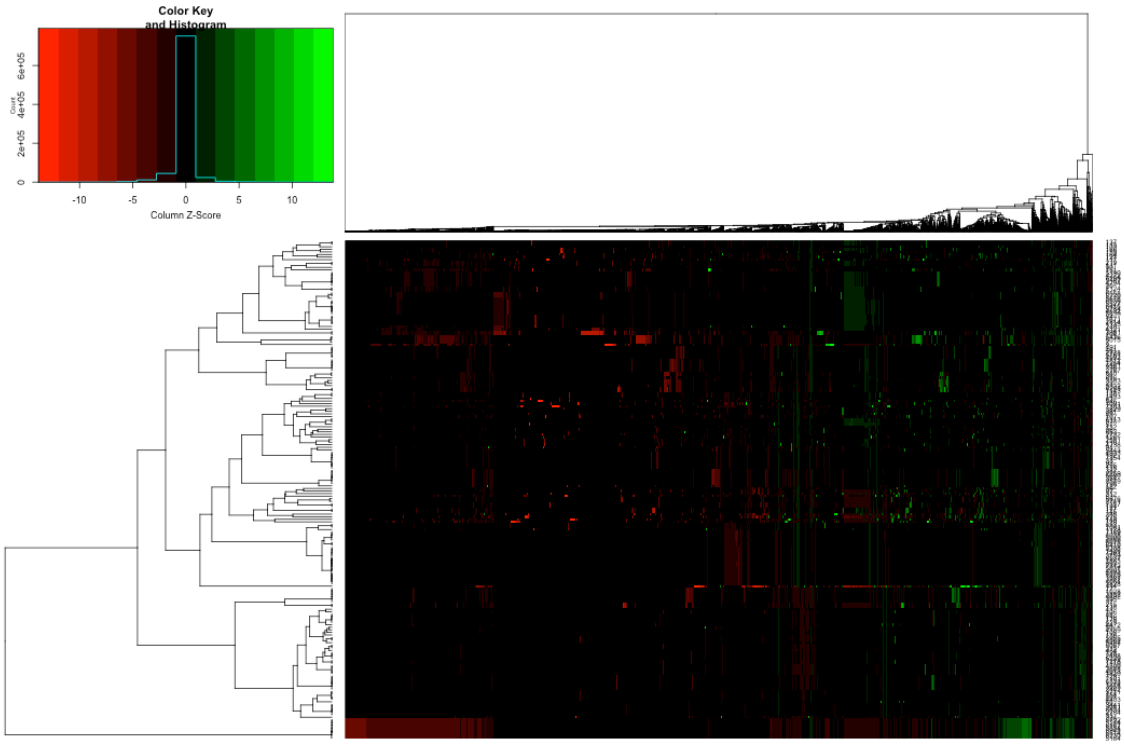

Figure 6.13: Community A Predicted Functions. A heatmap was made to show changes in predicted functions among OTUs assigned to community A in the PaO_2/FiO_2 > 300 network. OTU numbers are listed across the vertical axis and functions across the horizontal axis. Brighter green indicates functions with greater prevalence than the mean and brighter red indicates functions with less prevalence than the mean.

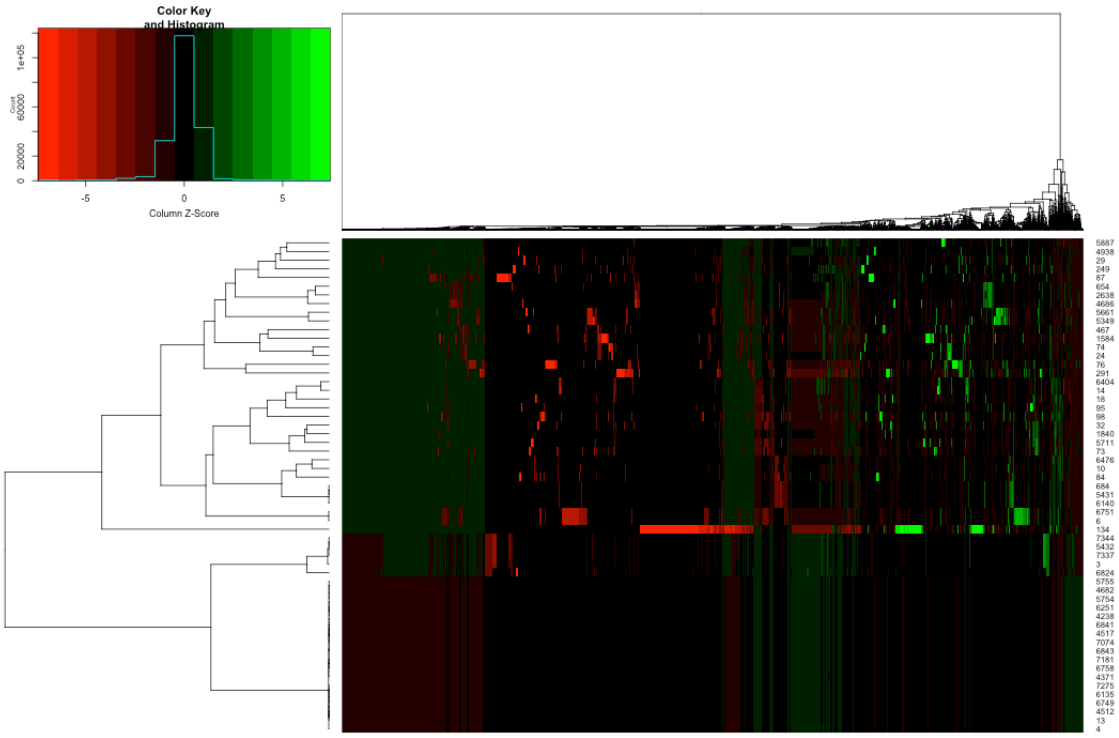

Figure 6.14: Community B Predicted Functions. A heatmap was made to show changes in predicted functions among OTUs assigned to community B in the PaO_2/FiO_2 > 300 network. OTU numbers are listed across the vertical axis and functions across the horizontal axis. Brighter green indicates functions with greater prevalence than the mean and brighter red indicates functions with less prevalence than the mean.

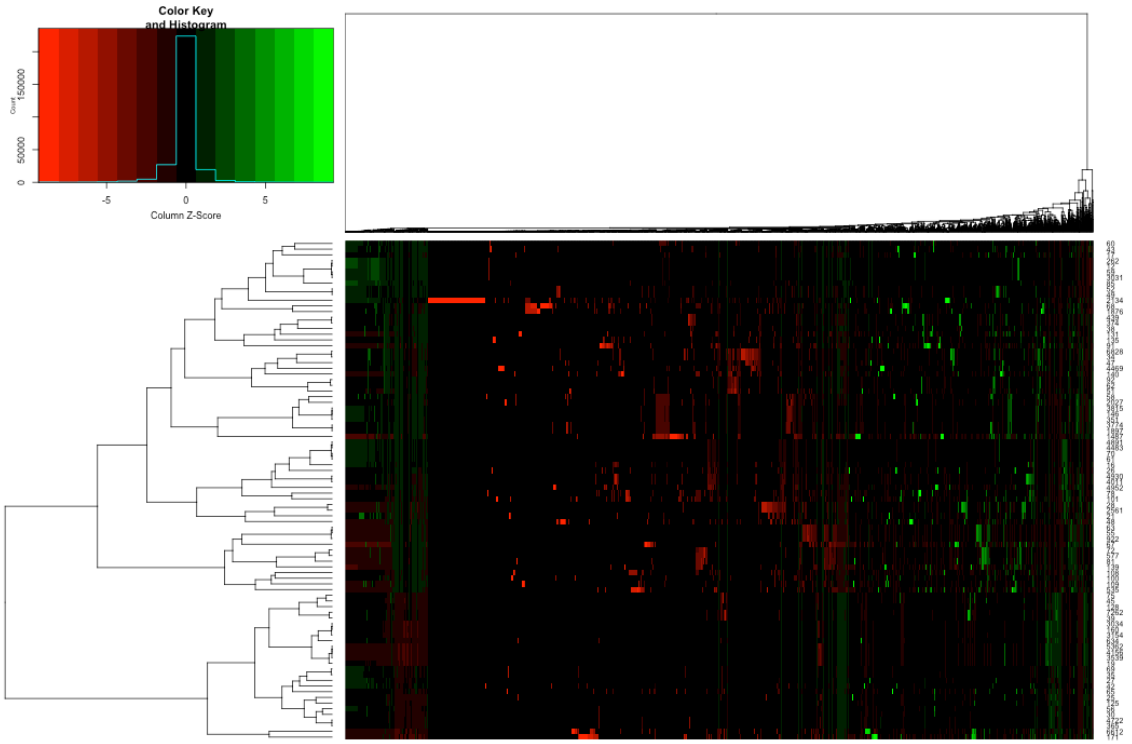

Figure 6.15: Community C Predicted Functions. A heatmap was made to show changes in predicted functions among OTUs assigned to community C in the PaO_2/FiO_2 > 300 network. OTU numbers are listed across the vertical axis and functions across the horizontal axis. Brighter green indicates functions with greater prevalence than the mean and brighter red indicates functions with less prevalence than the mean.

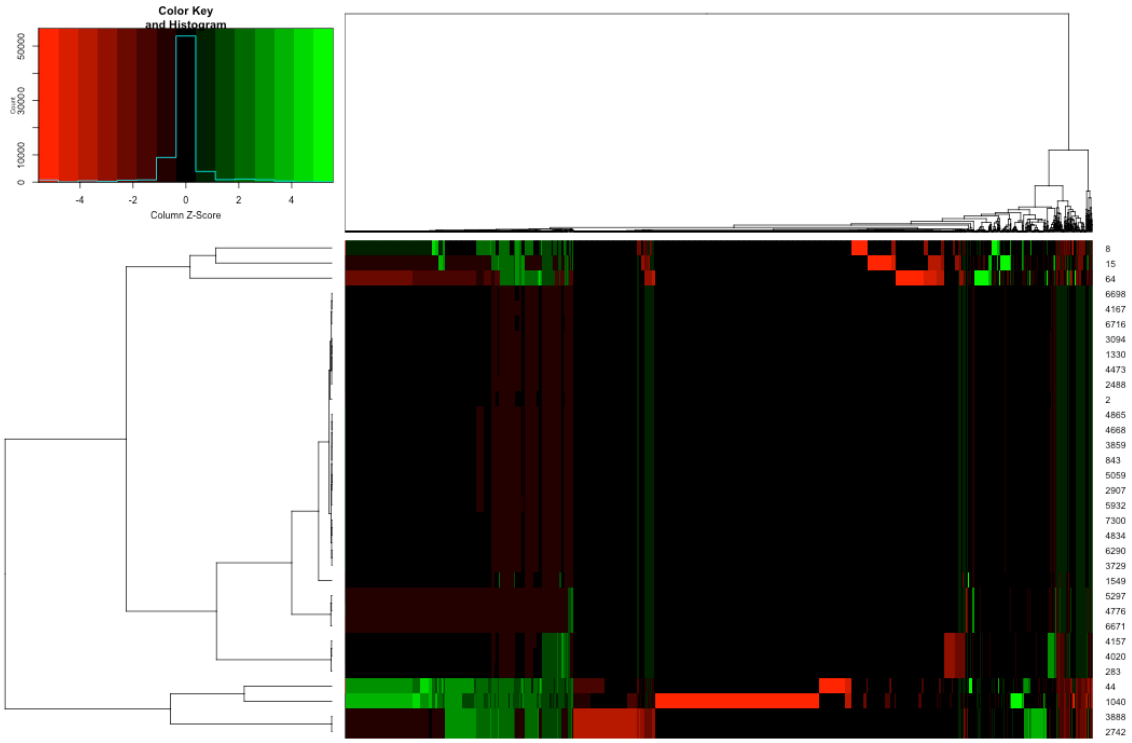

Figure 6.16: Community D Predicted Functions. A heatmap was made to show changes in predicted functions among OTUs assigned to community D in the PaO_2/FiO_2 > 300 network. OTU numbers are listed across the vertical axis and functions across the horizontal axis. Brighter green indicates functions with greater prevalence than the mean and brighter red indicates functions with less prevalence than the mean.

Tables 6.2 and 6.3 summarize the differences in predicted functions per $PaO_2/FiO_2 \leq 300$ and $PaO_2/FiO_2 > 300$ communities. For each community within each network, we identified the single predicted function with maximum abundance and the single predicted function with maximum variance. Each community had multiple predicted functions with the same minimum values and minimum variance. The tables contain the number of functions per community per network with the same minimum values when multiple were present. The name of the function is listed if there was a

single minimum value. Overall, predicted functions with maximum abundance and maximum variance were similar among all the communities. The number of predicted functions with minimum abundance and variance contained more variability per community. Functions with minimum abundance in ALI communities 2 and 4 were identical.

Community	**Maximum Abundance**	**Minimum Abundance***	**Maximum Variance**	**Minimum Variance**
1	Probable multidrug resistance ABC transporter ATP-binding permease protein	757	Probable multidrug resistance ABC transporter ATP-binding permease protein	71
2	Fumarate reductase flavoprotein subunit	771[§]	Iron complex outer membrane receptor protein	19
3	Iron complex outer membrane receptor protein	920	Iron complex outer membrane receptor protein	Transposase for insertion sequence elements
4	Probable multidrug resistance ABC transporter ATP-binding permease protein	771[§]	RNA polymerase sigma-70 factor	17

Table 6.2: $PaO_2/FiO_2 \leq 300$ Communities Predicted Function Summary. *Number of functions with the same minimum value [§]Identical functions

Community	Maximum Abundance	Minimum Abundance*	Maximum Variance	Minimum Variance
A	Iron complex outer membrane receptor protein	920	Methyl-accepting chemotaxis protein	4
B	Probable multidrug resistance ABC transporter ATP-binding permease protein	1108	Iron complex outer membrane receptor protein	19 (identical to minimum functions of ALI community 2)
C	Probable multidrug resistance ABC transporter ATP-binding permease protein	771 (identical to minimum functions of ALI community 4)	Iron complex outer membrane receptor protein	Transposase for insertion sequence elements
D	Uncharacterized gene/protein	808	RNA polymerase sigma-70 factor	17 (identical to minimum functions of ALI community 4)

Table 6.3: $PaO_2/FiO_2 > 300$ Communities Predicted Function Summary. *Number of functions with the same minimum value

6.3.3 Random Forest Prediction of Functions Representative of Network Communities

We used the RF model to determine the importance of the predicted functions in classifying the OTUs by their SparCC-assigned communities. We fit one model to the

$PaO_2/FiO_2 \leq 300$ data alone and another to the $PaO_2/FiO_2 > 300$ data alone. Here, our response variable was the SparCC community assignment and our features consisted of the predicted functions per OTU. Before selection of the training data, we performed variable selection to decrease *p*. A limitation of the RF algorithm is that it cannot handle *p* larger than *N* in an *N x p* data matrix, where *N* is the number of samples and *p* the features. Our *N* was 372 for both data sets, as 372 OTUs were identified among the data. Our *p* was initially 6,911, the number of predicted functions for each OTU from the PICRUSt algorithm. Elimination of predicted functions for all OTUs that totaled zero brought this down to 4,621. Further variable selection was necessary to make *p* equal to or less than our *N* of 372. To do so, we selected functions which had a small ratio of within-class variance to between-class variance. This left us with predicted functions with the highest variance among a single function but low variance between functions. Setting a threshold of greater than or equal to 0.05 brought *p* down to 328 for both the $PaO_2/FiO_2 \leq 300$ and $PaO_2/FiO_2 > 300$ predicted functions. With this data set, we first chose appropriate model parameters by randomly selecting half of the data for use as a training set. We then fit the model to the entire data set using the pre-determined parameters. This was done independently for the $PaO_2/FiO_2 \leq 300$ and $PaO_2/FiO_2 > 300$ data sets. These are output with the most important predicted function at the top for patients with $PaO_2/FiO_2 \leq 300$ in Figure 6.17 and for patients with $PaO_2/FiO_2 > 300$ in Figure 6.18. Importance of the variables, which is determined using the Gini index, decreases in order down the list for both figures. The Gini index measures the purity of the node when splitting the trees by each predictor. The purer the node, the more accurately the chosen variable explains the categorization of the response variables, and the higher the Gini

index. The KEGG ID numbers from PICRUSt were replaced with shortened versions of the function name in both figures. Importantly, there was no overlap among the functions in Figures 6.17 and 6.18, indicating that the functions of most importance in determining the communities are distinct among patients with and without $PaO_2/FiO_2 \leq 300$. For patients with $PaO_2/FiO_2 \leq 300$, the antibiotic transport system permease protein was ranked as most important in assigning the OTUs to the communities within the $PaO_2/FiO_2 \leq 300$ network in Figure 6.5. For patients with $PaO_2/FiO_2 > 300$, glyceraldehyde phosphate dehydrogenase was ranked as most important in determining the communities in Figure 6.6.

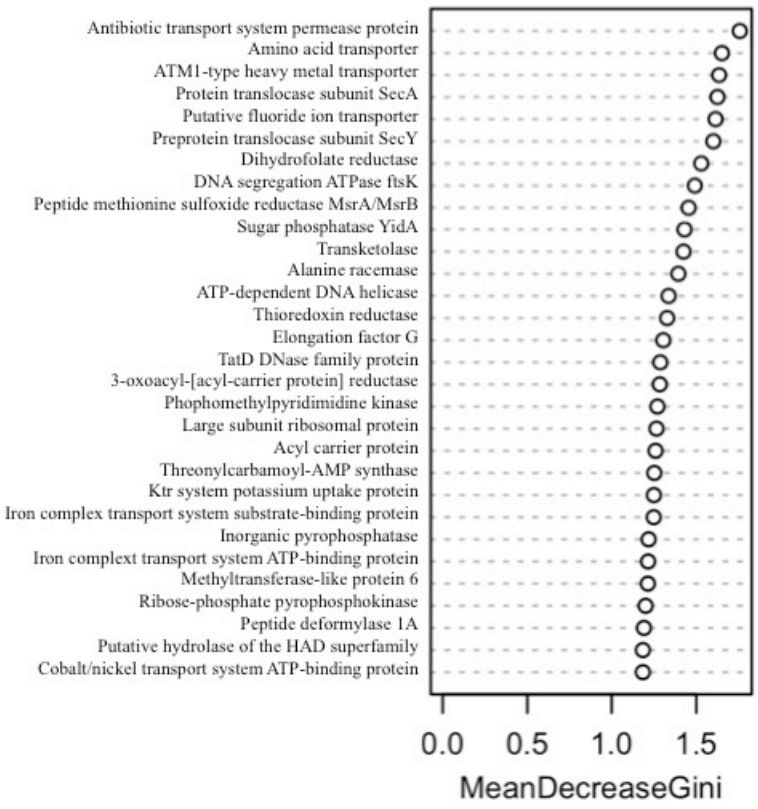

Figure 6.17: Predicted Functions Ranked by Importance in Determining the SparCC Community Assignments in the $PaO_2/FiO_2 \leq 300$ Network. A RF classification model was run to determine which predicted functions are most important in determining the SparCC community assignments. Functions listed at the top of the figure have the highest importance and decrease from there.

Community	Number of OTUs Containing the Predicted Function	Name of OTU with Highest Abundance of the Function
1	25	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Actinomycetales; f__Actinomycetaceae; g__Actinomyces; s__
2	54	
3	138	
4	93	

Table 6.4: OTUs Containing the Most Important Predicted Function Among $PaO_2/FiO_2 \leq 300$ Communities.

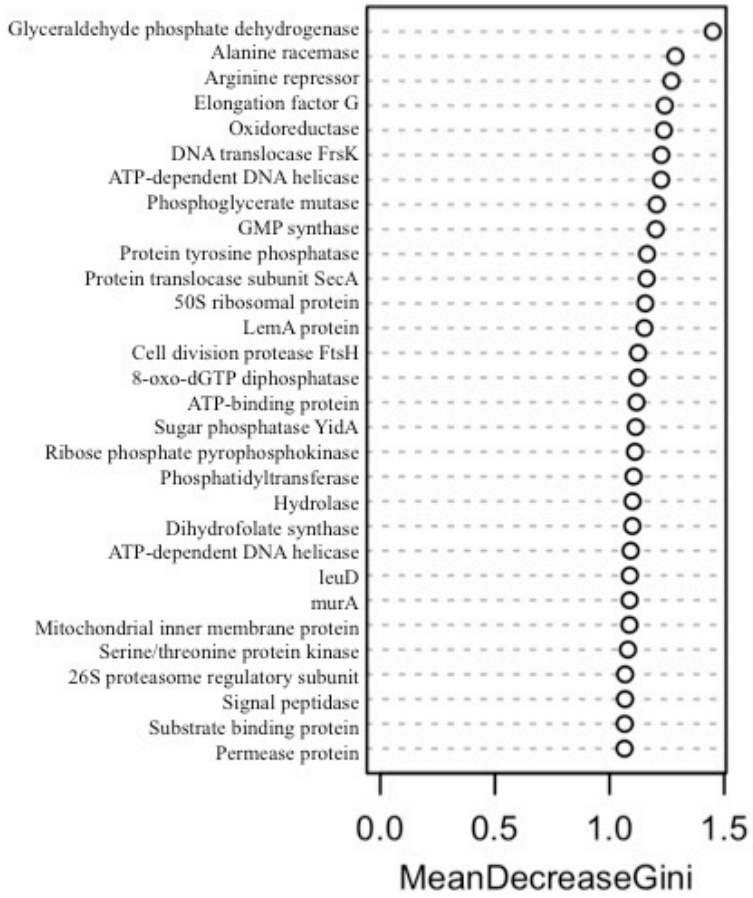

Figure 6.18: Predicted Functions Ranked by Importance in Determining the SparCC Community Assignments in the $PaO_2/FiO_2 > 300$ Network. A RF classification model was run to determine which predicted functions are most important in determining the SparCC community assignments. Functions listed at the top of the figure have the highest importance and decrease from there.

Community	Number of OTUs Containing the Predicted Function	Name of OTU with Highest Abundance of the Function
A	193	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Actinomycetales; f__Micrococcaceae; g__Rothia; s__mucilaginosa
B	57	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae; g__; s__
C	88	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Veillonellaceae; g__Veillonella; s__dispar
D	33	k__Bacteria; p__Firmicutes; c__Bacilli; o__Gemellales; f__Gemellaceae; g__; s__

Table 6.5: OTUs Containing the Most Important Predicted Function Among PaO_2/FiO_2 > 300 Communities.

Tables 6.4 and 6.5 indicate OTUs that contain the predicted functions ranked as most important by the RF models in Figures 6.17 and 6.18 among the $PaO_2/FiO_2 \leq 300$ and $PaO_2/FiO_2 > 300$ communities, respectively. The second column of both tables shows the total number of OTUs per community that contain the predicted function while the third column lists the name of the OTU with the highest predicted abundance of this function. In Table 6.4, each $PaO_2/FiO_2 \leq 300$ community contains a variable number of OTUs that are predicted to have the antibiotic transport system permease protein. P/F ≤ 300 community 3 contains the most OTUs with this function. For every $PaO_2/FiO_2 \leq 300$ community, the OTU with the highest abundance of the predicted function is identified as a species of *Actinomyces*. Similarly, each $PaO_2/FiO_2 > 300$ community in Table 6.5 shows varying numbers of OTUs that contain glyceraldehyde phosphate dehydrogenase, with P/F > 300 A containing the highest number. Unlike Table 6.5, the identities of the

OTUs with highest abundance of this predicted function are different for each community.

6.4 Discussion

Understanding microbial community dynamics is critical to predicting functional changes that may have a significant impact on patient or environmental outcomes. Current sequencing technology effectively identifies the bacteria comprising a community of interest but elucidation of their functions remains a challenge. In theory, WGS captures both microbial identity and function. However, the short reads and inadequate coverage make assembly difficult [275]. 16S rRNA gene amplicon sequencing is more commonly used due to its lower computational complexity and increased identification accuracy but its reliance on a common ribosomal gene sequence means that it cannot directly provide functional information on the identified bacteria. To infer possible microbial interactions based on co-occurence, correlation methods have been applied to metagenomic abundance data, with positive correlations interpreted as beneficial interactions and negative as competitive [124]. Ideally, metgenomics sequencing studies would be carried out in parallel with other 'omics' techniques, such as metabolomics and transcriptomics, which would give more in-depth information on bacterial relationships and community functions. However, this is often cost-prohibitive, necessitating the development of computational methods that can identify bacterial interactions and predict functions with high accuracy. Community network algorithms continue to be developed and improved upon, while other algorithms exist to predict functions from WGS [275] data as well as 16S rRNA gene amplicon data [170]. Use of these algorithms alongside other 'omics' techniques has demonstrated their accuracy in

predicting community functions and their power in identifying important therapeutic targets from the community as a whole [126]. However, few studies have examined interactive networks of bacteria within the entire community and their corresponding functional changes. We have used a community network algorithm appropriate for microbiome data to identify changes in interactive communities among the airway microbiota following burn and inhalation injury. We applied the PICRUSt algorithm to these communities to predict bacterial functions and fit a RF model to determine which predicted functions are most important in classifying the community assignments. Our results identified distinct predicted functional differences among the SparCC communities within patients with and without $PaO_2/FiO_2 \leq 300$. Application of the RF model allows prediction of the most important predicted function determining the SparCC communities, providing a specific hypothesis that can be validated experimentally. This work employs a systems biology approach, using computational methods to examine the community as a whole and identify specific interactions that may play significant roles in patient outcomes. Such an approach allows further pursuit of focused hypotheses with maximum relevancy to the overall community, resulting in discovery of therapeutic targets most likely to improve patient outcomes.

Our previous work in chapter 4 identified significant enrichment of the OTU *Prevotella melaninogenica* in patients with $PaO_2/FiO_2 \leq 300$ 72 hours after injury. This implies that *P. melaninogenica* may play an important role in $PaO_2/FiO_2 \leq 300$ after burn and inhalation injury, but we cannot discern the specifics of its role without further study. Selection of this specific organism does not take into account its interaction with others in the community that may contribute to its enrichment through beneficial sharing of

metabolic factors or inhibit its growth through competition. To take these interactions into consideration, we employed methods that allow study of the range of bacterial relationships within the community. Previous studies have used Pearson's and Spearman's correlations, which are inappropriate for metagenomic data as they do not take into account its compositional nature nor the sparseness of the data [128]. Instead, we applied SparCC, which performs a log transformation of the data that allows appropriate application of the Pearson correlation method. Here, we implemented a threshold in order to account for the sparseness in the data set. For both the NMI and number of communities, a threshold of 0.14 produced the best stability for both networks. This threshold was applied to produce the $PaO_2/FiO_2 \leq 300$ network in Figure 6.5 and the $PaO_2/FiO_2 > 300$ network in Figure 6.6. The OTUs in Figure 6.5 are colored according to community assignment and these colors are maintained in Figure 6.6 to show the change in OTU correlations among patients with and without $PaO_2/FiO_2 \leq 300$. The networks have overall shapes that are different from each other, indicating differences in positive correlations among OTUs in patients with and without $PaO_2/FiO_2 \leq 300$. This could imply one of two things; (1) that development of $PaO_2/FiO_2 \leq 300$ drives changes in mutually beneficial interactions among bacteria or (2) that changes in mutually beneficial relationships precede development of $PaO_2/FiO_2 \leq 300$. Further experiments need to be done to confirm which of these is correct. Although the network graphs visually display differences among the OTU correlations, we quantified OTU overlap in the contingency table in Table 6.1. Overlap of OTUs within the communities of both networks varies. Some overlap at a maximum of 77 OTUs and some have an overlap of only one OTU. Higher overlap indicates similarity in membership among the communities, which may

also indicate similarity in overall community interactions and functions. Conversely, less overlap may imply greater differences among interactions and functions and highlight OTUs that play important roles in $PaO_2/FiO_2 \leq 300$ through their interactions with other OTUs. The communities that display the least overlap are $P/F \leq 300$ community 1 and $P/F > 300$ communities A and B, which share a single OTU. Further investigation into the differences between OTUs among these three communities as well as their predicted functions could reveal relationships that distinguish $P/F \leq 300$ and $P/F > 300$ communities, possibly providing biomarker species for $PaO_2/FiO_2 \leq 300$ and/or other taxa that could be important therapeutic targets.

Figures 6.7 and 6.8 are OTU abundance heatmaps ordered by the community to which each OTU was assigned. Although there are differences in abundance among the OTUs, there are no clear patterns by community assignment. In Figure 6.7, abundance among the $PaO_2/FiO_2 \leq 300$ communities appears mostly random. This is similar for the $PaO_2/FiO_2 > 300$ communities in Figure 6.8, except that $P/F > 300$ community A contains more OTUs with average abundance. This demonstrates that abundance alone does not explain differences among the communities detected by the SparCC algorithm. By itself, SparCC can identify possible beneficial or competitive interactions based on co-occurrence. We took this a step further by applying PICRUSt to predict functions for the OTUs within these communities and using machine learning to explore significant differences between them. The heatmaps in Figures 6.9 through 6.12 show differences in abundance of predicted functions for the $PaO_2/FiO_2 \leq 300$ communities while Figures 6.13 through 6.16 show them for $PaO_2/FiO_2 > 300$ communities. Each heatmap displays a different pattern of predicted function over- or under-abundance as compared to the

mean, indicating that our results predict that each SparCC-identified community is doing something different. Through visual examination of the heatmaps, community 1 (Figure 6.9) shows the greatest variation in predicted functions among the $PaO_2/FiO_2 \leq 300$ communities (Figures 6.9 through 6.12). This predicted variation in function abundance could indicate that this community is more active than the other $PaO_2/FiO_2 \leq 300$ communities, whose patterns of over- and under-abundant predicted functions is less variable (Figures 6.10 – 6.12). $P/F \leq 300$ community 1 is also the community that has only a single OTU in common with $P/F > 300$ communities A (Figure 6.13) and B (Figure 6.14). This community is possibly the most active among the $PaO_2/FiO_2 \leq 300$ communities and the most different. These results indicate that further investigation of the OTUs present within $PaO_2/FiO_2 \leq 300$ community 1 could reveal bacterial interactions and functions that play important roles in disease pathogenesis. The greatest variation among the $P/F > 300$ communities (Figures 6.13 through 6.16) is not as clear from visual inspection. None of these communities appear to contain as much variation as $P/F \leq 300$ community 1, except possibly $P/F > 300$ community 4. This could imply that the $P/F > 300$ communities are not as functionally active as the $P/F \leq 300$ communities, especially $P/F \leq 300$ community 1. This comparison again implies that further investigation of the OTUs within $P/F \leq 300$ community 1 and their predicted functions could reveal interactions relevant to hypoxia as indicated by the PaO_2/FiO_2 ratio.

In Tables 6.2 and 6.3 we summarized predicted functions with maximum and minimum abundance and maximum and minimum variance among the communities. This quantifies the differences in expression implied by the heatmaps in Figures 6.9 – 6.16. While each community contained a single predicted function with maximum

abundance and variance, there were multiple predicted functions with the same minimum abundance and variance for each community. For P/F ≤ 300 and P/F > 300 communities, the maximum function for two was a multidrug resistance ABC transporter ATP-binding permease protein and one was iron complex outer membrane receptor protein (Tables 6.2 and 6.3). P/F ≤ 300 community 2 was predicted to have fumarate reductase flavoprotein subunit as its maximally abundant function while for P/F > 300 community 4 it was an uncharacterized protein. The number of functions with the same value of minimum abundance varied between all the communities. PF > 300 community 3 contained identical minimum functions as P/F ≤ 300 community 4. Differences among the communities in maximum and minimum variance showed similar patterns. Two P/F ≤ 300 and two P/F > 300 communities displayed maximum variance in the iron complex outer membrane receptor protein and one community within each network had maximum variance in RNA polymerase sigma-70 factor. P/F ≤ 300 community 1 contained maximum variance in the permease protein and P/F > 300 community 1 in a methyl-accepting chemotaxis protein. Two P/F ≤ 300 communities had minimum variance in predicted functions identical to those with minimum variance within two P/F > 300 communities. Despite these similarities, the variation across all the maximum and minimum values are distinct per community, in agreement with the observational data from the heatmaps. This quantitative summary is limited in scope as compared to the heatmaps, which makes the heatmaps more useful in identifying differences in overall predicted function expression patterns. Though subtle, these differences imply that these communities are doing different things, possibly as a consequence of the interaction of their different OTUs.

The heatmaps and quantitative summaries of the predicted functions per community indicate that they are doing different things, but this does not explain which of the functions are most important among them and may play roles in hypoxia. To identify predicted functions of importance to the SparCC-identified communities, we fit a RF model to the data. As explained in chapter 5, this model is appropriate for metagenomic data sets due to its computational efficiency, its ability to handle compositional data, and its selection of relevant features from the data. The importance of the predicted functions according to the Gini index is shown for the P/F $\leq$ 300 data set in Figure 6.17 and for the P/F > 300 data set in Figure 6.18. Output from the PICRUSt algorithm uses KEGG Orthology identification numbers from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database to identify predicted genes from 16S rRNA gene sequencing data. We used the KEGG [276] and Universal Protein Resource (UniProt) [277] databases to identify the names of the predicted genes and their functions from the KEGG orthology numbers. The names of the predicted functions deemed important by the RF algorithm are listed in Figures 6.17 and 6.18. Overall, there was no overlap in functions that the RF model determined most important in the P/F $\leq$ 300 and P/F > 300 communities, implying that predicted functions that define the communities in both networks are distinct. In agreement with the SparCC networks in Figures 6.5 and 6.6, this implies that either (1) development of hypoxia within three days of burn and inhalation injury induces distinct communities of bacteria with distinct functions within the airways or (2) development of distinct communities of bacteria with distinct functions early after injury drives development of hypoxia. Although further experiments need to be done to confirm this, our work reveals the development of distinctly correlated bacterial

communities with distinct predicted functions dependent on patient PaO_2/FiO_2 ratio. Further analysis of these communities could provide marker species to detect hypoxia early as well as other co-occuring species that induce other problems later on.

Besides identifying distinct predicted functions among the P/F ≤ 300 and P/F > 300 communities, the RF model ranked these functions in order of their importance for each. Among patients with hypoxia, the antibiotic transport system permease protein was ranked as most important in determining the SparCC-identified communities, and glyceraldehyde phosphate dehydrogenase was identified as most important in determining the communities among patients without hypoxia. If these functions can be experimentally validated, this may indicate that the communities among patients with $PaO_2/FiO_2 \leq 300$ are better able to transport antibiotics out of the cell, making them more resistant to treatment. Glyceraldehyde phosphate dehydrogenase has been shown to enhance adhesion of *Neisseria meninigitidis* to host cells independently of the presence of a capsule [278], implying a role in pathogenesis and infection for this protein. Based on these results, antibiotic transport may be more important for the $PaO_2/FiO_2 \leq 300$ bacterial communities and their subsequent roles in patient outcomes than their ability to adhere to host cells and cause infection. These results are interesting in light of our previous work in which we found enrichment of *Prevotella melaninogenica* in patients with $PaO_2/FiO_2 \leq 300$, and that this was not affected by antibiotic treatment. Future work could isolate *P. melaninogenica* from burn patient airways and determine whether it contains this antibiotic transport system permease protein and if this gives the bacteria an advantage in resisting antibiotic treatment and persisting in the airways of patients with hypoxia. Other strains could be isolated from burn patients without hypoxia and tested for

glyceraldehyde phosphate dehydrogenase in order to confirm its presence and perform future studies to elucidate its specific role. The RF model effectively identified predicted functions both distinct and important to microbiota communities among patients with and without $PaO_2/FiO_2 \leq 300$.

After identifying the predicted functions most important in determining the P/F ≤ 300 and P/F > 300 communities, we wanted to know which OTUs contained these functions and which communities they belonged to. This is listed in Table 6.4 for the P/F ≤ 300 communities and Table 6.5 for the P/F > 300 communities. Among the P/F ≤ 300 communities, community 3 had the most OTUs predicted to express the antibiotic transport system permease protein. P/F ≤ 300 community 1 contained the fewest OTUs with this predicted function. This is also the community with the highest variability in predicted function expression. The OTUs with the highest abundance of the permease protein were all identified as a member of the *Actinomyces* species for each community. In the P/F > 300 communities, community A contained more OTUs predicted to express glyceraldehyde phosphate dehydrogenase. Unlike the P/F ≤ 300 communities, OTUs with the highest expression of this function were all from different bacterial families. In patients with hypoxia, if *Actinomyces* has the highest expression of the permease protein, which the RF model predicted to be most important in determining the P/F ≤ 300 communities, it is likely that these bacteria play a significant role in the disease. Unfortunately, it is impossible to determine what this role is without specific experiments with *Actinomyces*, perhaps in a mouse model of inhalation injury. Highest expression of the most important function among the same taxa in P/F ≤ 300 communities but among different taxa in each P/F > 300 community suggests similarity induced among the

communities and their functions in patients with hypoxia. Longitudinal studies could reveal whether the P/F ≤ 300 communities continue to become functionally, if not taxonomically, similar the worse hypoxia becomes, and whether the P/F > 300 communities continue to diversify by function and/or taxa. If this is so, functional similarity among SparCC-identified communities could be a possible early indicator of hypoxia.

This work is clearly hypothesis generating and specific hypotheses will need to be validated in the lab. However, it provides a set of tools to identify areas and/or interactions of interest in a large dataset that would be difficult to interpret otherwise. A limitation to our method is that it is a prediction based on pre-existing databases of information. It cannot detect the exchange of genetic information between bacteria in the community through lateral gene transfer or development of antibiotic resistance. Additional studies will be necessary to elucidate this and to confirm functional predictions. The method's ability to examine the community as a whole and identify specific and relevant hypotheses for future, focused studies out weigh this limitation. Other limitations include the inability of the RF algorithm to handle *p* larger than *N*, necessitating either variable selection or extension of the number of samples.

In conclusion, we have employed a set of computational methods in a novel way to make predictions about alterations in disease-driven OTU interactions and functional changes. Rather than identifying differences in OTU abundance as a result of hypoxia, as we did in our previous work, this method allows identification of OTU interactions and how this alters their predicted functions. This systems-level approach takes the entire community into account, allowing determination of specific interactions that may play

significant roles in the development and/or progression of disease. These interactions can then be replicated experimentally to pursue mechanistic studies that could lead to new and more effective therapeutic targets.

CHAPTER 7: ADDITIONAL STUDIES: ALTERATION OF BRONCHIAL EPITHELIAL CELL RESPONSE TO WOOD SMOKE PARTICLES BY BACTERIA

7.1 Introduction

Application of NGS to burn patient bronchial washings allows investigation into broad changes in airway microbiota following inhalation injury and the development of hypoxia, as well as correlation with other patient factors and outcomes. Data analysis with machine learning algorithms can identify important elements of the microbiota that may play clinically relevant roles. However, NGS methods alone are observational and follow-up with mechanistic methods is necessary to confirm results and identify specific therapeutic targets. Both *in vivo* animal and *in vitro* cell culture models can be used to confirm the importance of specific bacterial taxa identified in NGS studies. Our previous work using 16S rRNA gene amplicon sequencing to characterize the airway microbiota after burn and inhalation injury revealed specific changes associated with the development of $PaO_2/FiO_2 \leq 300$. However, this does not answer how or why these changes occur, which is important in identifying effective therapeutic targets. Therefore, we sought to develop a cell culture model using primary airway bronchial epithelial cells in order to determine changes in their response to smoke and bacteria. We used cells from healthy human volunteers grown at air-liquid interface (ALIF) and introduced wood smoke particles (WSP) alone and with bacteria and assessed epithelial integrity and inflammatory response. We found that WSP induces an oxidative stress response in

human bronchial epithelial cells (HBEC) that is attenuated when *Klebsiella pneumoniae* is introduced. Further, doses of greater than $1x10^4$ colony forming units (CFU)/ml of *K. pneumoniae* disrupted epithelial integrity in fully differentiated HBECs.

7.2 Methods

7.2.1 Primary Human Bronchial Epithelial Cells

HBECs were obtained from healthy volunteers using a protocol previously approved by the University of North Carolina at Chapel Hill Institutional Review Board [279]. Briefly, healthy volunteers underwent bronchoscopy after consent was received. Cells were extracted from bronchial brushings and expanded in bronchial epithelial growth medium (BEGM, Clonetics, San Diego, CA, USA).

7.2.2 Air-Liquid Interface and Exposures

Isolated cells were plated on a transwell insert at $1x10^5$ cells/insert. Cells were grown submerged in a 1:1 mixture of BEGM and Dulbecco's Modified Eagles Medium (DMEM) with high glucose, growth supplements, bovine pituitary extract, bovine serum albumin, and nystatin until they reached confluence. Apical media was replaced with 0.5ml fresh media and basolateral media with 1ml fresh media every 48 hours. Once the cells reached confluence, retinoic acid was added to the media and the apical media was removed to bring the cells to the air-liquid interface. Cells were maintained for 21 days with replacement of basolateral media every 48 hours to allow for differentiation into ciliated and mucous-producing cells. At day 14 at ALIF, basolateral media was replaced with antibiotic-free media in preparation for WSP and bacteria exposures. Media

continued to be replaced every 48 hours. On day 21, WSP alone, bacteria alone, or WSP and bacteria were introduced to the cells for 24 hours.

7.2.3 Bacterial Strains and Culture Conditions

Klebsiella pneumoniae and *Staphylococcus aureus* were obtained from Carolina Biological Supply Company (Burlington, NC, USA). Prior to experiments, bacteria were grown separately in LB growth media overnight at 37°C in a shaking incubator. Appropriate doses were calculated from growth curves done for both species and the optical density taken at a wavelength of 600nm with a spectrophotometer. Bacteria were then centrifuged for two minutes at top speed, media was aspirated off, and cells were suspended in phosphate buffered saline (PBS, Thermo Fisher Scientific, Pittsburgh, PA, USA). Two more washes were done in PBS before resuspension to the appropriate concentration and introduction to the apical side of HBECs grown on a transwell insert. HBECs were grown at ALIF for 22 days and all exposures were done on day 23.

7.2.4 Wood Smoke Particle Generation and Composition

WSP were generated as described by Ghio *et. al.* [280]. Briefly, red oak wood was heated in a Quadrafile 3100 woodstove (Colville, Washington, USA). A teflon filter was used to collect smoke and particles were extracted in 1N HCl. Metals present in the WSP were determined using inductively coupled plasma optical emission spectroscopy (Perkin Elmer, Norwalk, Connecticut, USA). For cell exposures, WSP were collected from the stainless steel chimney above the woodstove and sonicated (Thermo Fisher

Scientific, Pittsburgh, PA, USA) in water to disaggregate the particles. Particles were resuspended in PBS prior to exposure to HBECs.

7.2.5 Cytotoxicity Assay

Cytotoxicity of WSP and bacteria exposures was measured following 24 hours of exposure using the CytoTox 96 Non-Radioactive Cytotoxicity Assay (Promega, Madison, WI, USA). The assay was performed per the manufacturer's instructions. HBECs alone as well as HBECs with bacteria served as positive lysis controls.

7.2.6 Transepithelial Eletrical Resistance

Transepithelial electrical resistance (TEER) was used as a measure of epithelial integrity. TEER measurements were performed using the EVOM2 (World Precision Instruments, Sarasota, FL, USA), which applies alternating current and measures resistance through the membrane of the transwell insert with a two-pronged electrode. Measurements were made in triplicate every 48 hours while the cells were differentiating for 21 days as well as immediately before and 1, 3, and 24 hours after WSP and bacteria exposures. During TEER measurements, 0.5ml of PBS was placed on the apical surface and 1ml of media in the basolateral side of the transwell in order to submerge each end of the electrode. A transwell without cells but with 1ml basolateral media and 0.5ml apical PBS was always used as a blank and its averaged triplicate value was subtracted from subsequent cell measurements.

7.2.7 Oxidative Stress Response and Pro-Inflammatory Gene Expression

Total cellular RNA was extracted at 1, 3, and 24 hours post WSP and bacteria exposure. Complementary DNA was made and quantitative PCR was performed using TaqMan primer/probe sets targeted to the HO-1 and IL-8 genes (Applied Biosystems, Pittsburgh, PA, USA). Gene expression was normalized to the human beta-actin gene and analyzed using the Pfaffl method [281].

7.3 Results

7.3.1 Post-Exposure Cytotoxicity

Release of lactate dehydrogenase as measured by the CytoTox kit was used as an indication of cytotoxicity. Cytotoxicity was measured after 24 hours of $1x10^2$ CFUs/ml *K. pneumoniae*, a dose which does not disrupt epithelial integrity, as well as increasing doses of WSP, and WSP and *K. pneumoniae* together. Additional exposures with WSP and *S. aureus* together, as well as *K. pneumoniae*, and *S. aureus* co-exposures with WSP, have been planned but not completed. Figure 7.1 demonstrates cytotoxicity of *K. pneumoniae* after 24 hours as compared to cells alone and PBS-only controls. Figure 7.2 demonstrates WSP toxicity from 0.3μg/cm^2 up to 530.53μg/cm^2 as compared to cells with PBS controls. Finally, Figure 7.3 shows cytotoxicity with doses of WSP alone from 1μg/cm^2 and up to 50μg/cm^2 as well as these doses with concurrent exposure of $1x10^2$ CFUs/ml *K. pneumoniae.*

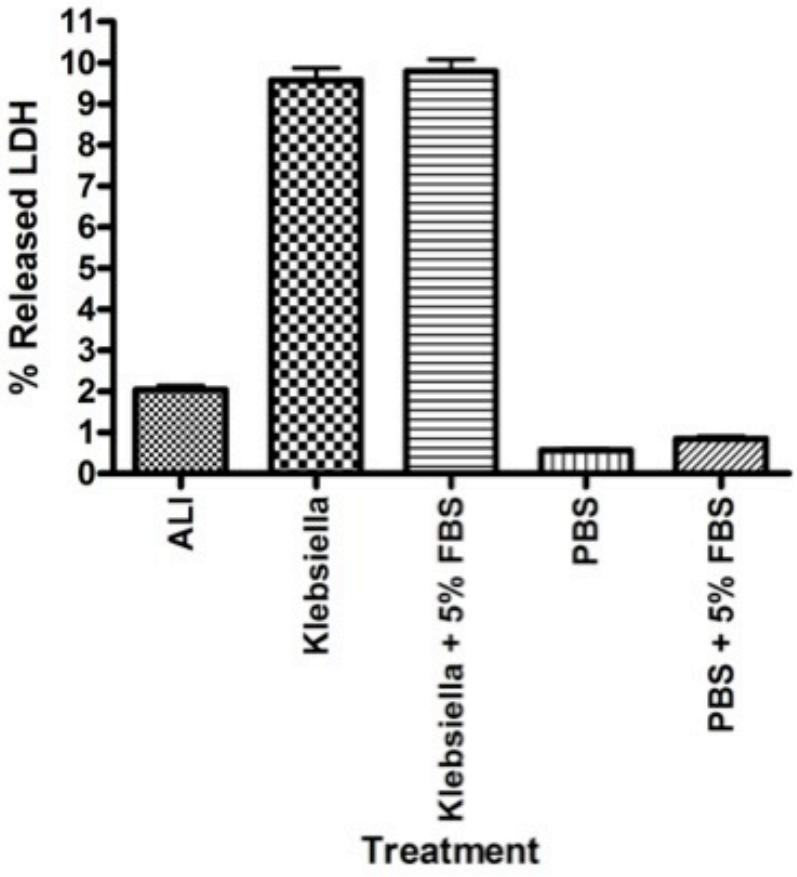

Figure 7.1: *K. pneumoniae*-Induced Cytotoxicity. HBECs grown at ALIF for 23 days were exposed to $1x10^2$ CFUs/ml *K. pneumoniae* suspended in PBS for 24 hours and LDH release in the apical compartment was measured using the CytoTox kit.

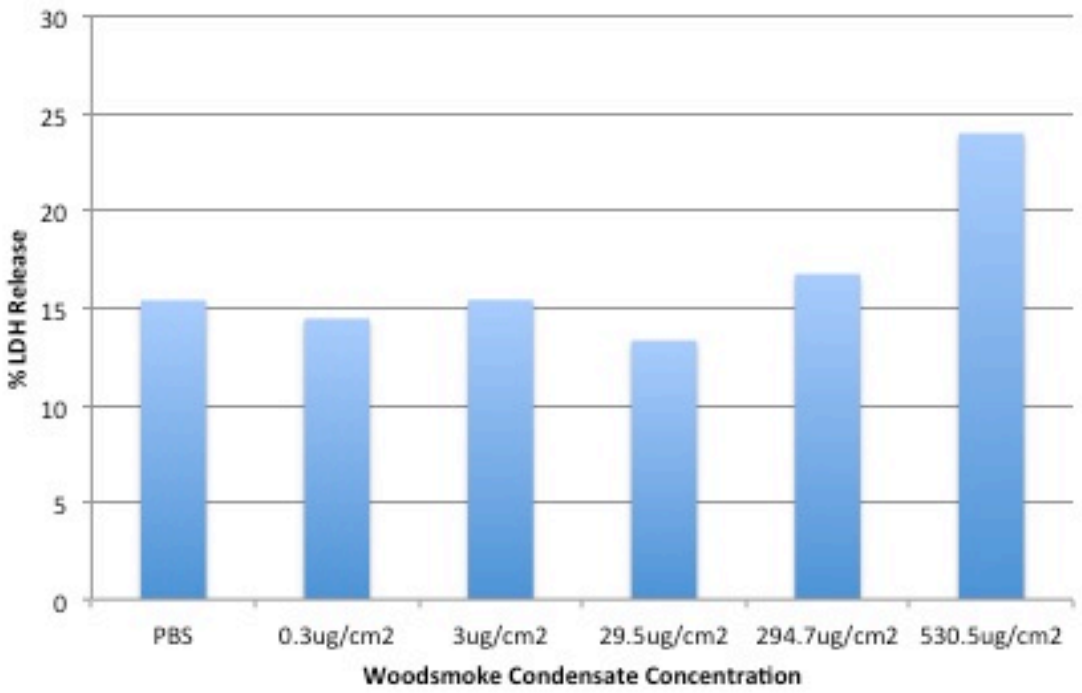

Figure 7.2: WSP-Induced Cytotoxicity. HBECs grown at ALIF for 23 days were exposed to the indicated concentrations of WSP suspended in PBS for 24 hours and LDH release in the apical compartment was measured using the CytoTox kit. Woodsmoke condensate = WSP. $N = 6$

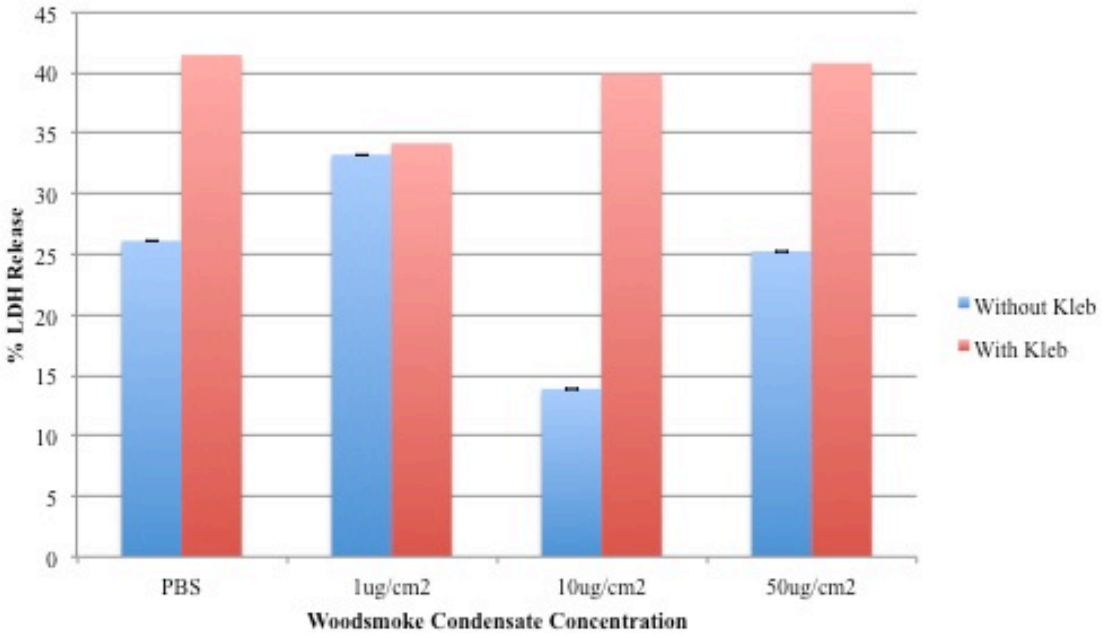

Figure 7.3: WSP Alone and WSP with *K. pneumoniae* Cytotoxicity. HBECs grown at ALIF for 23 days were exposed to the indicated concentrations of WSP suspended in PBS alone for 24 hours of the indicated concentration and $1x10^2$ CFUs/ml *K. pneumonia*. LDH release in the apical compartment was measured using the CytoTox kit. Woodsmoke condensate = WSP. $N = 6$.

7.3.2 TEER During Cellular Differentiation

Prior to WSP and bacteria exposures, TEER measurements were taken every 48 hours over the 21 days of HBEC differentiation in order to determine the appropriate time period for exposure. All measurements were normalized to a blank control well and done in triplicate. Figure 7.4 shows a representative graph of changes in TEER over 21 days at ALIF. TEER gradually increases from ALIF day 0, peaks around day 10, and decreases over the next 3 – 5 days to plateau between days 17 and 22. This pattern was seen in cells from all donors. Bacterial exposures were done on day 23, once TEER plateaued. Antibiotic-free media was added at day 14 in order to avoid killing bacteria.

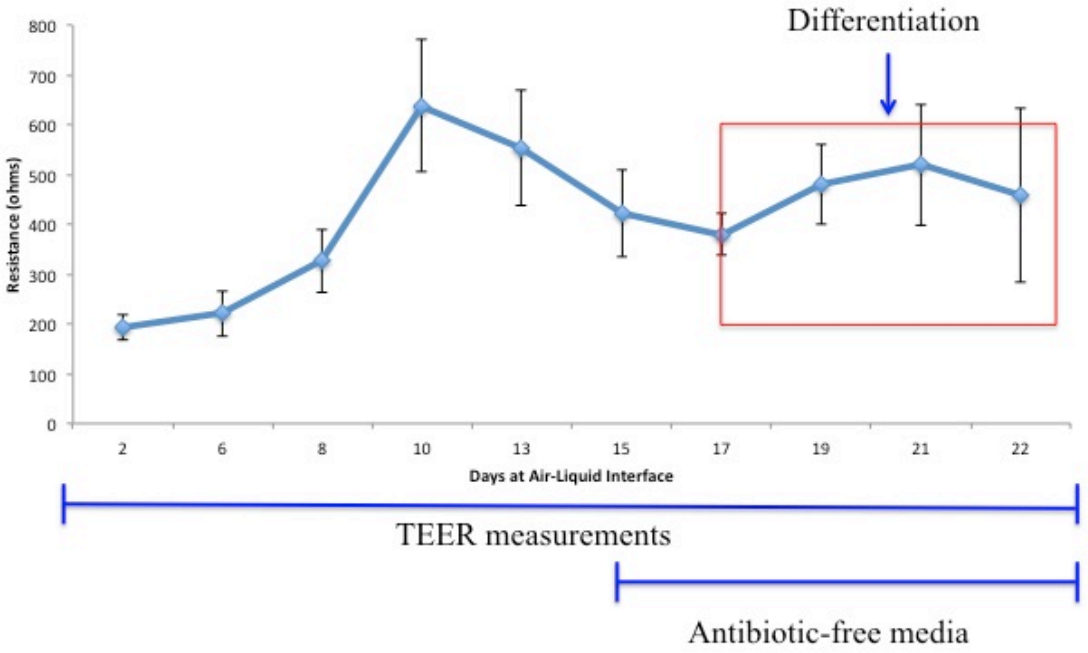

Figure 7.4: Changes in TEER During HBEC Differentiation. TEER measurements were taken every 48 hours over the 21 days of cellular differentiation in order to determine the appropriate time to begin exposures. TEER levels plateau between days 17 and 22, so exposures were done on day 23. Cells were placed in antibiotic-free media 7 days prior to exposure.

TEER measurements were taken after 24 hours of exposure of increasing doses of *K. pneumoniae* to HBECs that had been at ALIF for 21 days in order to determine an appropriate dose that did not disrupt epithelial integrity (Figure 7.5). Staurosporine, a non-specific protein kinase inhibitor that induces apoptosis [282], was used as a positive control. Growth of *K. penumoniae* on agar plates from basolateral media post exposure was used to confirm disruption of epithelial integrity. Of the four doses used in Figure 7.5, no bacterial growth was seen only at $1x10^2$ CFUs/ml *K. pneumonia.* This dose was selected for future exposures as well as combined WSP and *K. pneumoniae* exposures.

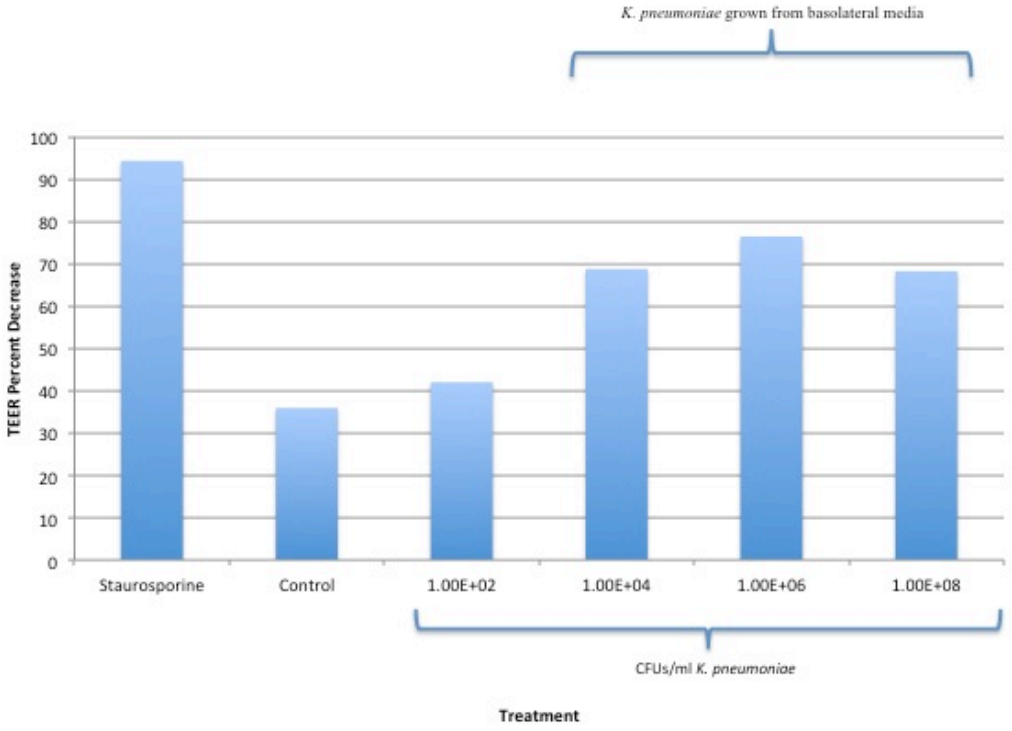

Figure 7.5: Disruption of Epithelial Integrity by *K. pneumoniae.* HBECs grown at ALIF for 23 days were exposed to the indicated doses of *K. pneumoniae* for 24 hours. Bacterial were placed on the apical side of the transwell insert. TEER was measured before and after exposure and basolateral media was incubated on agar plates after exposure. A dose of $1x10^2$ CFUs/ml resulted in the lowest decrease in TEER and no growth from basolateral media.

7.3.3 Induction and Attenuation of Oxidative Stress Response by WSP and Bacteria

RNA was harvested from cells after 1, 3, and 24 hours of exposure to WSP alone or WSP with *K. pneumoniae*. RNA was used as a template to make cDNA, which was then used in a qPCR reaction to quantify expression of the oxidative stress gene heme oxygenase-1 (HO-1). Additional exposures with WSP and *S. aureus* together as well as *K. pneumoniae* and *S. aureus* co-exposure with WSP have been planned but not completed. Figure 7.6 shows expression of HO-1 24 hours after exposure of HBECs to

increasing concentrations of WSP alone or WSP and *K. pneumoniae*. HBECs were grown for 23 days at ALIF and exposed on day 23.

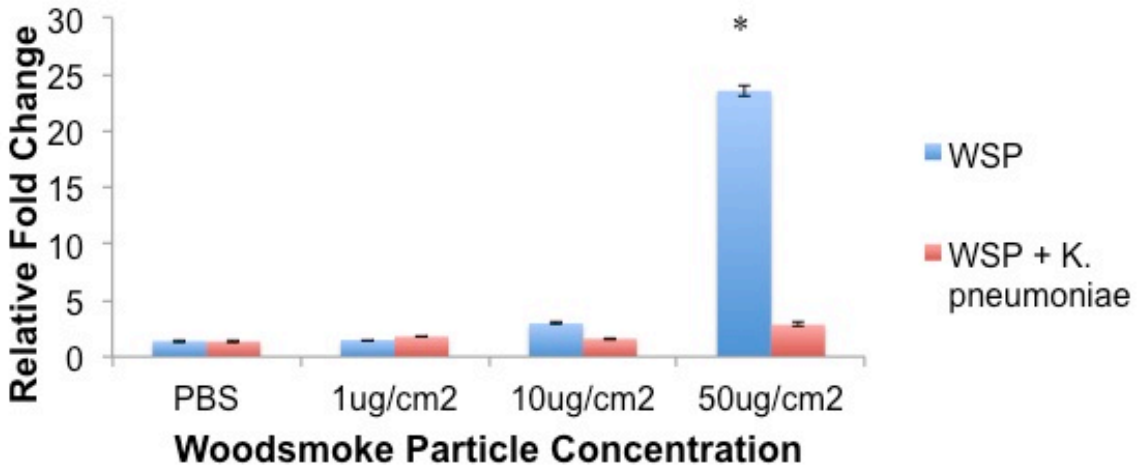

Figure 7.6: *K. pneumoniae* Attenuates WSP-Induced HO-1 Expression. HBECs were exposed to increasing concentrations of WSP alone or WSP with *K. pneumoniae* for 24 hours on day 23 at ALIF. The highest dose of WSP induced a significant increase in HO-1 expression that was attenuated by addition of *K. pneumoniae*. $n = 4$.

7.3.4 Induction of Inflammatory Response by WSP and Bacteria

The cDNA made from the RNA above was also used in a qPCR reaction to quantify expression of the pro-inflammatory cytokine interleukin-8 (IL-8) after WSP and *K. pneumoniae* exposure. Figure 7.7 shows IL-8 expression after 24 hours of exposure. IL-8 was induced at much lower levels by both exposures as compared to HO-1. WSP and *K. pneumoniae* together induced significantly increased IL-8 expression over WSP alone.

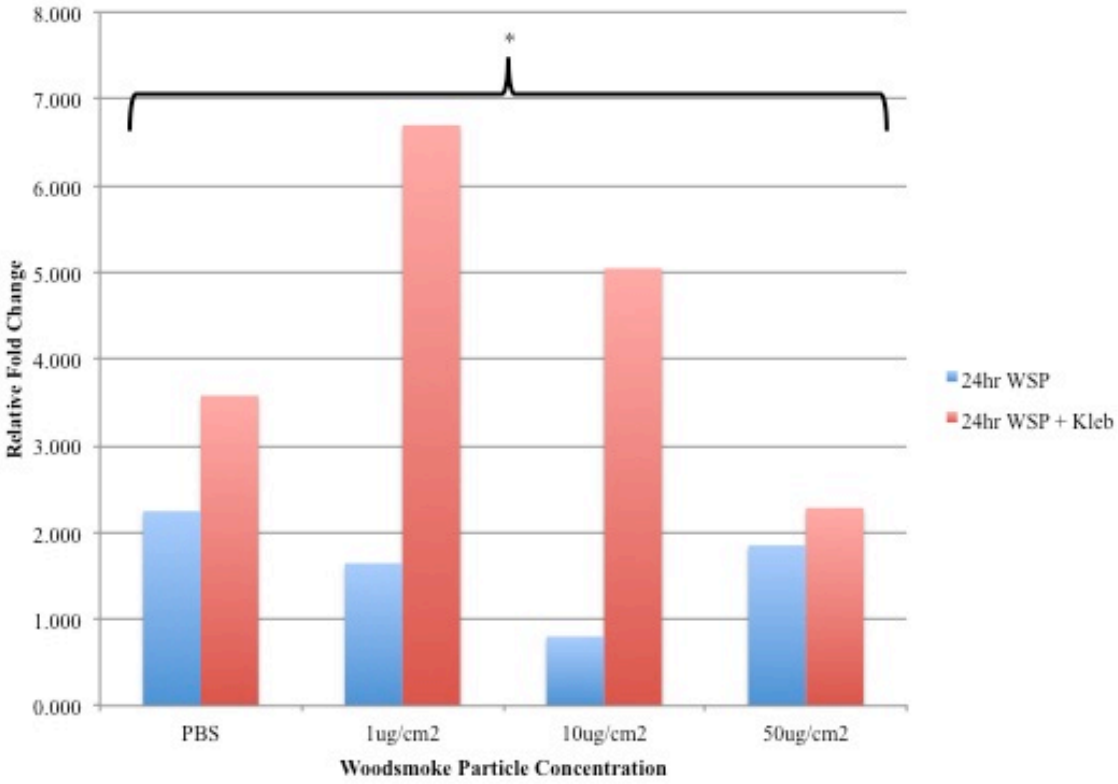

Figure 7.7: *K. pneumoniae* Increases IL-8 Gene Expression Over WSP. HBECs were exposed to increasing concentrations of WSP alone or WSP with *K. pneumoniae* for 24 hours on day 23 at ALIF. *K. pneumoniae* induced significantly increased IL-8 expression as compared to every dose of WSP. $n = 4$.

7.4 Discussion

Advances in NGS have made possible high dimensional metagenomic studies in which the populations of bacteria present in a sample can be identified and examined simultaneously. The study of bacterial communities rather than individual species or strains allows insight into population dynamics and host-community interactions that may be crucial to identifying promising new therapeutic targets. However, these large studies require the development of advanced computational methods, along with sufficient computing power and speed, in order to understand and interpret this complex data.

Though complex and sophisticated, these methods are largely observational and can be used to identify taxa that may play important roles in various disease processes or environments. Hypothesis-driven, mechanism-based microbiological methods are critical to exploring and confirming the interactions revealed by metagenomic studies.

Our previous metagenomic work on changes in the airway microbiota following burn and inhalation injury suggested that specific taxa might play important roles in the development of ALI early after injury. In-depth computational work predicted differences in functions of the communities in patients with and without ALI and how these taxa may be interacting with each other. The need to experimentally validate this work led to the development of a model in which we exposed HBECs to WSP and burn patient-relevant bacteria in order to understand how these exposures changed the response of healthy airway epithelial cells.

We chose to work with HBECs due to their physiological relevance to the airways and their lack of genetic abnormalities that are frequently seen in cell lines [283,284]. When grown at air-liquid interface, primary human bronchial epithelial cells undergo mucociliary differentiation into ciliated and mucous-producing cells, which mimics the cell types present *in vivo* [283]. This pseudo-stratified epithelium better represents physiologic airway responses but genetic variation from donor to donor can make reproducibility difficult, necessitating the use of more cells than would be needed with a cell line [285]. Immortalized cell lines, however, contain chromosomal abnormalities that can alter cell behavior and their ability to differentiate normally, bringing in to question their accuracy in modeling physiologic responses [284,286]. We monitored cellular differentiation of HBECs over 21 days through visual detection of mucous production

and cilia movement. Further, we measured changes in TEER over the course of differentiation and observed an early and rapid increase in resistance near day 10 at ALIF that plateaued between days 17 and 22 (Figure 7.4). We speculate that this may be due to initial formation of tight junctions followed by differentiation into a pseudo-stratified epithelium. Immunofluorescence to detect the presence of tight junction proteins has been used in other ALIF models to confirm the association of TEER with epithelial integrity. We began optimizing immunofluorescence methods to detect tight junction proteins such as zonula occludens-1 (ZO-1) and those in the claudin family. Immunofluorescence has been used to detect reduction of ZO-1 and ZO-2 in association with disruption of epithelial integrity by cigarette smoke extract [287]. Further, immunofluorescence has demonstrated reduced ZO-1 present in the airways of patients with atopic asthma [288]. These proteins have been associated with increased TEER values, but recent studies suggest observed changes in TEER involve other tight junction proteins as well. ZO-1 was observed to be consistently present regardless of TEER value in primary bronchial cells at ALIF from horses as well as humans, indicating that this single tight junction protein does not fully explain the observed changes in TEER. To determine whether changes in tight junction proteins are responsible for the variation in TEER we observed in Figure 7.4, ZO-1, occludins, claudins, and other tight junction proteins need to be detected, visualized, and quantified using immunofluorescence and Western blot. Measurement of TEER and visualization of tight junction formation also do not confirm cellular differentiation. Immunofluorescent staining for mucin and cilia proteins as well as haemotoxylin and eosin staining could be used to visualize and confirm differentiation.

Based on the TEER values in Figure 7.4, we chose to introduce bacteria and WSP exposures on ALIF day 23, once TEER plateaus. We used WSP generated by a wood burning stove in order to increase reproducibility in our results. Replication of the smoke that the burn patients were exposed to is not possible due to its heterogeneous composition and it would introduce a high level of variability into our experiments. To avoid these issues, we generated smoke of a consistently reproducible composition using the wood burning stove. Using whole smoke could have increased the physiological relevance of our model, but techniques that currently exist to do this do not reliably and consistently expose cells to reproducible doses. Use of WSP extracted from the wood stove allowed us to consistently expose the cells to the same dose of particles of the same composition, increasing reproducibility in our system. Our choice of bacteria was based on organisms commonly cultured in the Burn Center. *K. pneumoniae*, a gram negative organism, and *S. aureus*, which is gram positive, are both frequent causes of pneumonia in the Burn Center. Further, *K. pneumoniae* is part of the Enterobacteriaceae family, which dominated the patient bacterial communities in our 16S rRNA gene amplicon sequencing results (Chapter 4). A more appropriate choice for the gram positive organism would have been a member of the Streptococcaceae family, since this taxa was also highly abundant. The long-term plan for these experiments is to introduce *K. pneumoniae* and *S. aureus* in a co-exposure with WSP in order to replicate bacterial community interactions at a very simplistic level. The community network results from Chapter 6 could be used to guide these models and select relevant co-occuring taxa that are predicted to interact with one another.

We gave the cells antibiotic-free media for 7 days prior to the exposures in order to avoid toxic effects on the bacteria that could skew our results. In Figure 7.5, we introduced four doses of *K. pneumoniae* at day 23 ALIF ranging from $1x10^2$ CFUs/ml up to $1x10^8$ CFUs/ml for 24 hours. We measured TEER before and after introduction of bacteria as well as a staurosporine-treated positive control well and a negative control well containing PBS on the apical side. From this, we calculated the percent decrease in TEER, which is displayed in Figure 7.5. Bacteria were only introduced to the apical side of the transwell, so that we did not expect to find them in the basolateral side if epithelial integrity remained intact. After exposure, we incubated the basolateral media overnight on LB agar plates to determine which dose of *K. pneumoniae* disrupted epithelial integrity, allowing the bacteria to cross through the porous transwell membrane into the bottom of the well. As shown in Figure 7.5, the three highest doses disrupted epithelial integrity. We chose to continue our experiments with a dose of $1x10^2$ CFUs/ml *K. pneumoniae*. If these experiments are continued with *S. aureus*, this experiment will need to be conducted again to determine an appropriate dose of this or another new organism.

Although our chosen dose of *K. pneumoniae* did not disrupt epithelial integrity, we did not know whether this dose was toxic to the cells. We used the CytoTox kit to measure release of LDH, an enzyme that catalyzes the conversion of pyruvate to lactate found in all living cells. This is an indirect measure of cytotoxicity, as it is assumed that increased release of LDH can be attributed to rupture of cell membranes from increased toxicity. In Figure 7.1, we measured LDH release from unexposed cells, untreated cells at ALIF (ALI), cells treated with *K. pneumoniae*, cells treated with *K. pneumoniae* and supplemented with fetal bovine serum (FBS), and control wells treated with PBS alone

and PBS supplemented with FBS. LDH release was higher in cells treated with *K. pneumoniae*, and there was not a significant difference in toxicity depending on FBS supplementation. Supplementation of *K. pneumoniae* with FBS was done to determine whether this altered the bacteria's cytotoxicity to the cells as compared to bacteria suspended in PBS. Since FBS did not alter cytotoxicity significantly, we continued our exposures with bacteria suspended in PBS. Although $1x10^2$ CFUs/ml of *K. pneumoniae* was more toxic than PBS or no treatment at all, it only induced an increase in LDH release of 10%. In Figure 7.2, we introduced HBECs to a wide range of WSP doses in order to assess cytotoxicity. LDH release is not significantly different from the PBS control except for the highest dose. Based on these results, we chose doses of $1\mu g/cm^2$, $10\mu g/cm^2$, and $50\mu g/cm^2$ to assess cytotoxicity in combination with $1x10^2$CFUs/ml of *K. pneumoniae.* We exposed HBECs at ALIF on day 23 to each of these WSP doses alone or simultaneously with *K. pneumoniae* for 24 hours (Figure 7.3). Each of these doses alone did not significantly increase LDH release as compared to the PBS control, but the addition of *K. pneumoniae* increased release by 15% for all but one of the doses.

For each of these exposures we performed qPCR to quantify expression of the oxidative stress response gene HO-1 (Figure 7.6) and the pro-inflammatory gene IL-8 (Figure 7.7). There was no significant difference in HO-1 expression between WSP alone and WSP with *K. pneumoniae* until the highest dose of WSP. At $50\mu g/cm^2$ WSP, HO-1 expression increased 25-fold. Addition of *K. pneumoniae* attenuated this response, which could suggest that *K. pneumoniae* inhibits HBEC oxidative stress response to WSP. In Figure 7.3, addition of *K. pneumoniae* to each dose of WSP increased LDH release, implying increased cytotoxicity. Increased cytotoxicity could explain attenuation of HO-1

expression at 50μg/cm^2 WSP with *K. pneumoniae* in Figure 7.6. However, we do not see this pattern at the other two doses of WSP, which suggests that an increase in cytotoxicity does not completely explain attenuation of HO-1 expression. Clearly, more experiments need to be done to confirm this effect.

In Figure 7.7, addition of *K. pneumoniae* increased expression of IL-8 at each dose of WSP, including the PBS control. Interestingly, the increase in IL-8 expression was highest with the lowest dose of WSP and decreased for increasing doses of WSP. Although the overall fold increase in IL-8 is low, addition of *K. pneumoniae* increases it significantly at each dose (one-way ANOVA performed in R: anova(lm(Fold_Change~Treatment, data=IL8))). For 50μg/cm^2 WSP, the dose at which *K. pneumoniae* attenuates HO-1 expression, IL-8 is significantly increased with *K. pneumoniae* but not to the extent that it is at the lower doses of WSP.

Both IL-8 and HO-1 are expressed when the transcription factor nuclear factor erythroid 2-related factor 2 (Nrf2) binds to the antioxidant response element (ARE) [289,290]. Nrf2 is normally bound to the repressor protein Kelch ECH associated protein 1 (Keap1) in the cytosol, inhibiting it until the application of stress to the cell changes the conformation of Nrf2 and releases it. It then crosses into the nucleus, where it binds to ARE, inducing transcription of HO-1 and IL-8 [267,290]. Despite their similarity of location, it has been shown that transcription of IL-8 and HO-1 are independent of each other [291]. Co-culture of nasal epithelial cells from patients with non-allergic chronic rhinosinusitis with *Streptococcus pneumoniae* increases expression of IL-8 and neutrophil adherence to endothelial cells [292]. Other studies have demonstrated the ability of *Pseudomonas aeruginosa* to induce increased IL-8 expression and recruit

neutrophils to the airways [293]. Induction of sepsis by intratracheal instillation of *Klebsiella pneumoniae* in mice and inhibition of HO-1 expression showed increased recruitment of neutrophils to bronchoalveolar spaces, decreased bacterial load, decreased alveolar collapse, and increased survival rate [294]. Together, these studies suggest that increased HO-1 expression interferes with IL-8 expression and its ability to recruit neutrophils to the site of injury. At doses of 1 and 10μg/cm^2 WSP with *K. pneumoniae*, we see increased levels of IL-8 with decreased levels of HO-1, which agrees with previous studies. Based on this, at 50μg/cm^2 WSP with *K. pneumoniae*, where we see attenuation of HO-1 expression, we would expect increased levels of IL-8. Instead, we see the lowest level of IL-8 expression with both WSP and bacteria (Figure 7.7). *K. pneumoniae* mutants lacking a polysaccharide capsule have been shown to induce signaling through toll-like receptor-4 (TLR4), and inhibition of TLR4 is associated with decreased HO-1 expression and increased iron levels [295,296]. The availability of iron regulates the ability of *K. pneumoniae* to synthesize capsular polysccacharide, which in turn upregulates expression of TLR4 and TLR2 on airway epithelial cells [297,298]. Signaling induced by lipopolysaccharide in bacterial cell walls may depend on both TLR2 and TLR4 [299]. This suggests that application of *K. pneumoniae* lacking a capsule in combination with WSP will activate TLR4 and TLR2, leading to increased expression of HO-1, decreased expression of IL-8, and decreased iron levels that would inhibit capsule polysaccharide synthesis. Instead, we see attenuation of HO-1 expression and a negligible increase in IL-8, which implies that TLR4 signaling is somehow inhibited. Without further experiments and detection of activation of the signaling molecules involved in expression of IL-8 and HO-1 it is impossible to determine how *K.*

pneumoniae is altering these genes. However, attenuation of HO-1 expression without an increase in IL-8 may indicate a novel pathway response to WSP and *K. pneumoniae* together, which could give mechanistic insight into bacteria and HBEC interaction after WSP exposure with further investigation.

This is an additional, incomplete study and, as such, has many limitations. The use of HBECs alone, rather than in co-culture with neutrophils or other innate immune cells, only represents the response of the airway epithelium to bacteria and WSP. Inclusion of innate immune cells could help elucidate the interaction between the epithelium and the innate immune system to regulate responses to bacteria. Further, exposure with a single bacterial species does not replicate the complexity of the airway microbiota as elucidated in chapter 4. Bacterial interactions are important in determining their impact on host cells, and this model fails to capture this accurately. A mouse model would be better suited to capture the spectrum of bacterial as well as host cell interactions and their response to whole wood smoke. Methods exist to expose both mice and cells at ALIF to whole wood smoke, but reproducibility of smoke concentration remains a challenge for both systems. Cell culture is superior to mouse models in regard to reproducibility, as the same concentration of WSP can be added to the same numbers of cells and bacteria each time. With mice, a consistent dose of WSP could be instilled intratracheally, but consistency in the number of host and bacterial cells exposed is more difficult to control. The cell culture model used here needs to be expanded to include *S. aureus* as well as more complex mixtures of bacteria. Immunofluorescence methods should be used to visualize formation of tight junction proteins and Western blot to quantify them. Western blot should also be used to detect activation of signaling

molecules involved in expression of IL-8 and HO-1 as well as to quantify these proteins themselves.

In conclusion, we have developed an *in vitro* model to understand the mechanisms behind changes in the airway microbiota following burn and inhalation injury. Metagenomics studies serve as a hypothesis-generating step, while application of computational methods allows selection of important relationships to explore further experimentally. Our results suggest that the gram negative bacteria *Klebsiella pneumoniae* inhibits expression of HO-1 induced by 50μg/cm^2 WSP after 24 hours of exposure. If this result can be confirmed, it suggests that *K. pneumoniae* can play a potentially beneficial role in burn patient airways by inhibiting oxidative stress responses that could damage the airway epithelium. Such results could alter thinking about what we consider a pathogen in the airways and how it might be used to improve patient outcomes. It will be important in future studies to examine how *K. pneumoniae*'s interaction with HBECs changes when other bacteria are introduced to the system, as bacterial interactions can shift their functions in beneficial, negative, or neutral directions. Our additional studies provide a model with which many avenues of microbiota-host interaction can be explored in a focused, mechanistic way that has the potential to reveal important therapeutic targets in the burn patient population.

CHAPTER 8: CONCLUSIONS AND FUTURE DIRECTIONS

8.1 Summary

The human microbiome holds promising potential for the future of preventative and precision healthcare. Research has revealed the role of the microbiota in various states of health and disease, including obesity [300], Crohn's disease [301], asthma [302], pregnancy [41], chronic rhinosinusitis [16], colorectal cancer [232], and COPD and smoking [144], among others. These studies have captured trends in changes of microbiota among disease states and identified specific taxa with important roles that could serve as therapeutic targets. The use of predictive machine learning algorithms alongside experimental data has demonstrated the ability of these computational methods to identify taxa that predict these disease states with high accuracy, providing tools to detect specific taxa for early and preventative treatment [26,247,249,264]. Although this work has demonstrated the importance of specific taxa in various disease states, the Human Microbiome Project has been instrumental in showing that the composition of healthy human microbiota can vary widely from individual to individual [8]. Significantly, the functions of these healthy communities remain similar despite changing bacterial community composition. More recent studies have demonstrated that these communities have a degree of fluidity in their composition that can be altered by diet [11,26,28,65,125,303] as well as xenobiotics, and particularly antibiotics [29,244,245,304]. The ability to modify these communities has led to the development of

strategies in which specific microbes are introduced to shift the community towards a healthier state. The most successful example of this strategy is in fecal microbiota transplantation (FMT), in which transfer of a healthy gut community to patients infected with nosocomial *Clostridium difficile* has been used to successfully treat the infection [38,305]. Although this is a promising treatment strategy, lack of understanding of the complex interactions among microbiota have hindered the success of FMT in treating other gut disorders, such as ulcerative colitis [306]. Such challenges indicate the importance of understanding individual microbial community dynamics in designing effective personalized treatment strategies.

Work on the airway microbiota is relatively recent and has focused on changes in bacterial communities during airway disease. Though few studies have examined the impact of airway injury on microbial communities, they have suggested that alteration of airway homeostasis leads to conditions conducive to bacterial colonization and growth, resulting in disease-specific alterations [141]. Our work has focused on changes in airway microbiota in the context of burn and inhalation injury and the development of airway disease as a result of these insults. We have applied amplicon sequencing techniques to bronchial washings from patients hospitalized for burn and inhalation injury. We detected a broad range of bacteria in the lower airways as soon as 24 hours after injury and observed significant differences in taxa abundance among patients with and without acute lung injury (ALI). These differences led us to apply machine learning techniques to our data to discover pre-existing patterns, what patient variables drive these patterns, and which taxa are predictive of ALI. Functional differences are increasingly being recognized as more important than community composition, which encouraged us to use

a community detection algorithm in combination with prediction of bacterial functions and machine learning to discern differences in community functions in patients with and without ALI. We recognize that these computational methods do not capture within-community changes, such as the development of antibiotic resistance and lateral gene transfer. To address this limitation, we developed an *in vitro* model examining the interaction of bacteria and human bronchial epithelial cells (HBECs) after exposure to wood smoke particles. Our work incorporates observational metagenomic studies with predictive computational methods and validation by experiment, providing a framework that allows identification of important microbial interactions and predicted functions that can be explored experimentally for identification of effective therapeutic targets.

Prior to 16S rRNA gene amplicon sequencing, bacterial DNA extraction methods had to be optimized for the bronchial washing samples. These samples were contaminated with mucous, soot, blood, and human airway cells, all of which could interfere with efficient DNA extraction from bacterial cells. A combination of physical, chemical, and enzymatic lysis methods followed by a commercial spin-column-based kit gave the best quality and highest quantity DNA. To account for the low quantity of bacterial DNA present in airway samples as compared to gut samples, we utilized a molecule tagging method to increase the accuracy of our sequencing results. Together, these methods ensured the best sequencing accuracy for our low-quantity airway samples.

Statistical analysis of our sequencing results revealed differences in bacterial community composition among patients with and without ALI early after injury. The OTUs *Streptococcus* and Enterobacteriaceae were dominant among all patients, but several low-abundance taxa were enriched among patients with ALI. Linear discriminant

analysis effect size ranked the OTU identified as *Prevotella melaninogenica* as most significantly enriched among patients with ALI, which was not impacted by antibiotic treatment. These results suggest that *Prevotella melaninogenica* is associated with ALI early after burn and inhalation injury but we cannot discern whether ALI drives enrichment of this OTU or enrichment of this OTU induces development of ALI. Further studies in a mouse model of inhalation injury are necessary to elucidate the relationship between this bacteria and ALI following burn and inhalation injury.

Application of high dimensional data analysis methods revealed clustering among the bacterial communities that was driven by patient body max index (BMI). Further investigation into this relationship could reveal specific patient outcomes predicted by each cluster. A random forest model showed that, among the bacterial families present within the patient samples, the family Streptococcaceae is predictive of ALI status. As with enrichment of *Prevotella melaninogenica*, these results require experimental confirmation.

Use of whole genome sequencing (WGS) can reveal limited information on bacterial functions, while parallel use of 'omics' techniques can provide a broader range of transcriptional and metabolomics changes among communities. The computational complexity of WGS data and the expense and sample requirements of additional 'omics' methods often make these approaches unfeasible. To address this issue, computational methods have been developed to predict bacterial functions from 16S rRNA gene amplicon sequencing data and the method is surprisingly accurate as compared to WGS [170]. Our previous work with 16S rRNA gene amplicon sequencing and machine learning methods revealed which bacteria were present among burn patients with

inhalation injury and which were predicted to be most important in association with ALI. However, since bacterial function is likely more important in patient outcomes than community composition, we applied the PICRUSt algorithm to our data to predict the abundance of functions among the OTUs in our samples. To understand how the microbes may be interacting with each other, we used the SparCC algorithm to detect four distinct communities of bacteria among patients with and without ALI. PICRUSt predicted distinct functions for each of these communities, and application of a random forest model identified different functions as most important to the ALI and No ALI communities. Among patients with ALI, an antibiotic transport system permease protein was ranked as most important in determining the interactive communities, while glyceraldehyde phosphate dehydrogenase was ranked as most important in determining the communities among patients without ALI. This may indicate that OTUs among patients with ALI express proteins to resist antibiotics at a higher level than OTUs among patients without ALI, giving them an advantage in resisting antibiotic treatment. Since OTUs among patients without ALI do not express this function at a high level, they may be more susceptible to antibiotic treatment. Glyceraldehyde phosphate dehydrogenase plays an important role in adherence of *Neisseria meningitidis* to nasal epithelial cells, suggesting that its expression may play a role in bacterial pathogenicity. This implies that bacterial communities in both groups of patients may be pathogenic, but are so in different ways according to the presence of ALI. If these functions can be experimentally validated, they may guide identification of specific therapeutic targets to prevent infection in patients both with and without ALI.

In an effort to confirm our computational work, we developed a method to mechanistically evaluate the impact of wood smoke particles on the interaction between bacteria and HBECs. Our additional studies suggest that bacteria inhibit cellular oxidative stress responses, which could be protective for cell survival. Future studies will be done to confirm and expand this work.

Our work has several limitations. Collection of burn patient bronchial washings was not originally intended for a microbiome study, which limited our ability to control for patient clinical treatment, alter collection conditions, and expand sample collection to encompass a greater number of patients. This also prevented us from collecting additional material to perform parallel transcriptomics and metabolomics studies in order to confirm the computationally predicted functions from the 16S rRNA sequencing results. The samples were taken and used first and foremost for patient care. Whatever was left after clinical testing was done was stored frozen in the sample repository. When this study was done, the samples had been stored for at least one year, prohibiting quantification and identification of bacterial species through traditional selective culture. The computationally predicted functions require experimental validation, which can be done in future studies. The *in vitro* experimental model is simplistic in its representation of the microbiota. We were limited to use of aerobic bacterial species when facultative anaerobic bacteria were identified as significantly enriched in the burn patients, and we used a single species when the microbiota is clearly more complex than this. Future studies will need to expand the complexity of the introduced community.

8.2 Future Directions

The nature of this data set leaves a variety of future directions open, from further observational studies with longitudinal data, to application of other computational methods, to mouse and continuation of *in vitro* cell and bacterial models.

8.2.1 Continuing Studies

Although we only performed a cross-sectional study with this data set, multiple samples per patient taken throughout their stay at the Burn Center were sequenced. These samples could be incorporated into a longitudinal study examining how the airway microbiota change throughout recovery. Patient clinical information exists in the RedCap database for each of these data points, allowing correlation of changes in cytokine production, PaO_2/FiO_2 ratio, carboxyhemoglobin levels, and others with the composition of the microbiota at each time point. Since this is clinical data, however, samples were not taken at consistent time intervals and fewer samples exist for patients with longer hospital stays. A longitudinal study will be biased towards patients with poorer outcomes since these patients were required to stay in the Burn Center for longer periods of time. The Burn Center implemented a routine standard of care for all incoming patients prior to collection of the samples used in our work. Plans exist to start collecting new samples from incoming patients according to this new protocol, which will increase consistency among sample collection and patient treatment. A future study could take advantage of this and repeat our cross-sectional study along with performing a longitudinal study. Other samples, such as oral and nasal swabs and fecal samples, could be collected

alongside the bronchial washings, expanding the study from the lower airway microbiota to upper airways and gut.

Additional work with our predicted bacterial functions could be done to associate various patient variables with changes in predicted functions. For example, we found that patient BMI drives similarity of the airway microbiota among all patients. Functional predictions for these communities could be associated with the respective patient BMIs and overall outcomes to predict the influence of BMI on microbiota interactions. Longitudinal analysis could show how interactions among the bacteria change over time along with their predicted functions, perhaps revealing dependencies among bacteria or competitive relationships that could be manipulated to improve patient outcomes. Clearly, these functional predictions require experimental validation. This could be done through isolation of bacterial strains with the functions of interest, and then employing bacterial genetic manipulation to mutate the gene of interest. The mutant bacterial strains could then be introduced to a mouse model of inhalation injury and metagenomic sequencing as well as transcriptomics and metabolomics could be performed to evaluate the impact of this function and its removal on inhalation injury. If collection of patient samples is re-started, bacterial RNA and protein could be isolated for specific detection of the functions of interest. This could be expanded to samples from additional body areas in order to compare functional changes across body sites.

The additional studies we initiated using HBECs, bacteria, and wood smoke particles could be expanded in future studies. As mentioned above, a limitation of this model is its simplicity in replicating the microbiota. Study of individual bacterial species will lead to an understanding of how each interacts with HBECs following smoke

exposure, which is valuable and may aid in understanding specific community interactions. However, the presence of other bacterial species may alter the functions of these individual species and their interactions with HBECs, rendering results using individual species meaningless. Future studies should employ a complex mixture of bacteria identified in burn patient airways following burn and inhalation injury. Further tailoring of these communities could be done by modeling them after species identified in patients with specific diseases, such as ALI and pneumonia. Our experiments with this model focused on the response of HBECs, but future studies should isolate RNA and protein from bacteria in order to evaluate their response as well. Finally, the biological relevance of this model could be improved through use of whole wood smoke. Methods exist to expose cells to whole cigarette smoke, and other methods are under development to expose cells to whole wood smoke. At the time we conducted these studies, none of these methods were able to provide a consist dose of smoke to the cells. However, future studies should be able to overcome this limitation, and may be able to use customized mixtures of smoke replicating the conditions burn patients have been exposed to.

8.2.2 Mouse Models

Mice are a commonly used model in biomedical research in general as well as within the microbiome field [42,123,125,307]. Advances in genetic manipulation along with the creation of germ-free and gnotobiotic mice have allowed discovery of the importance and function of the microbiome in ways that are not possible to do with humans [66]. Although we have begun to incorporate cell culture models, our study would benefit from use of a mouse model. Inhalation injury could be induced in both

germ-free and conventionalized mice in order to examine changes in airway microbiota over time in a more controlled manner than is possible in the burn patients. The germ-free mice could also be colonized with customized communities of bacteria that reflect those found in the airways of burn patients, allowing controlled study of their function and interaction. Gnotobiotic mice could be created by introducing specific bacteria of interest to germ-free mice, such as *Prevotella melaninogenica*. RNA and protein could be isolated to examine changes in gene expression and function as a result of these bacteria, and bacterial species lacking or containing specific functions could be studied as well. A mouse model ties together the observational metagenomic and predictive computational portions of our work with the more focused, mechanistic cell culture model. Mice allow specific manipulation of both the host and the microbiota, which can lead to confirmation of the human studies and closer examination of interactions that cannot be replicated in the cell culture model.

8.2.3 Predictive Modeling

Besides future experimental studies, additional computational methods could be incorporated into this work in order to model changes in community composition and function over time and predict patient outcomes. Changes in microbiota associated with development of a specific disease, such as pneumonia, could be traced over time. This would allow identification of changes in bacterial interactions and functions at specific time points that lead to development of pneumonia, possibly providing windows of time in which to most effectively target the responsible bacteria. Due to the heterogeneity in both patient clinical data and microbiota composition, such a model would require a large

number of patients and longitudinal samples in order to achieve accuracy. Replication of such a study would be necessary in a mouse model, where specific communities could be assembled and monitored in order to confirm their impact on outcome. If such a study could be done, it would provide additional clinical guidelines that physicians could use to test for specific bacteria at specific time points in accordance with specific clinical indications in order to assess the likelihood of development of certain diseases. This could allow early and targeted treatment that may improve burn patient outcomes following inhalation injury.

8.3 Conclusion

In conclusion, this study begins with an observational metagenomic study, incorporates advanced computational methods to predict bacterial interactions and functions, and suggests specific hypotheses relevant to these predictions that can be experimentally validated. Our work leaves several avenues open for future study, including incorporation of mouse models to confirm and further explore community interactions, as well as predictive modeling of longitudinal data. Mechanistic, hypothesis-driven work is crucial to identifying specific, effective therapeutic targets within the microbiome, but large-scale, high-dimensional studies are necessary to understanding the importance of these hypotheses to the overall community dynamics. Incorporation of both of these approaches takes into consideration community interactions as a whole to identify specific functions that can be validated with mechanistic studies. Our work attempts to unite high-dimensional, large-scale data analysis with more traditional mechanistic work in order to improve the overall efficiency and effectiveness of

metagenomic studies and their application to human health. It provides a starting point for the expansion of future studies to investigate and confirm specific host-microbe and microbe-microbe interactions, which may provide effective targets for therapeutic intervention within the burn patient population.

APPENDIX 1: DNA EXTRACTION PROTOCOL

Extraction of DNA Using Lysis Buffer Containing Lysozyme and Qiagen UCP Pathogen Mini Kit

This protocol gives high quality and high yield *S. aureus* DNA.
Materials Needed:

- Lysis buffer: 200mM NaCl, 100mM Tris pH 8 (filtered), 20mM EDTA (filtered), 20 mg/ml lysozyme
- RNase A (20ug/ml)
- 10% SDS (filtered)
- Qiagen UCP Pathogen Mini Kit
- Qiagen Pathogen Lysis tubes

Protocol:

1. Dissolve lysozyme in lysis buffer at 37C.
    a. Lysozyme hydrolyzes β(1→4) linkages between *N*-acetylmuramic acid and *N*-acetyl-D-glucosamine residues in peptidoglycan; this step dissolves lysozyme into the lysis buffer solution, which will enhance its activity
2. Resuspend sample in 400µl of lysis buffer; transfer to lysing tubes.
    a. Sample is in lysis buffer but lysozyme is not activated yet
3. Vortex and incubate at 37C for 30 min.
    a. Activation of lysozyme; degradation of gram positive cell wall, allowing cell membrane to swell and lyse in next step
4. Vortex for 10 min at full speed
    a. This will mechanically break up the cell wall in preparation for cell membrane lysis
5. Add 45ul of SDS and vortex.
    a. SDS is an ionic detergent and will disarticulate the cell membrane, allowing the cell to lyse open and release RNA, DNA, proteins
6. Add 1ul of RNase A and vortex. Let sit at room temperature for 30 min.
    a. Breakdown of RNA into pieces; RNA will contribute to overall nucleic acid concentration and I only want DNA; this does not remove the RNA nucleic acids, just breaks down the RNA
7. Spin down briefly and transfer 400uL to clean microcentrifuge tubes
8. Continue with step 1 of Qiagen QIAamp UCP Pathogen Mini Kit Protocol: Sample Prep (Spin Protocol), pg33:
9. Add 40 µl Proteinase K and mix the sample by vortexing for 10 s.
10. Incubate the sample at 56°C for 1hr.
11. Add 200 µl of Buffer APL2 to the sample. Close the cap and mix by pulse-vortexing for 30 s. Note: In order to ensure efficient pathogen lysis, it is essential that the sample and Buffer APL2 are mixed thoroughly to yield a homogeneous solution.
12. Incubate at 70°C for 10 min.
    a. This inactivates Proteinase K
12. Briefly spin the tube to remove drops from the inside of the lid.

13. Add 300 µl ethanol to the lysate. Close the cap, and mix thoroughly by pulse-vortexing for 15–30 s.

a. DNA is insoluble in ethanol; this will precipitate it out of solution

14. Carefully apply 600 µl of the mixture from step 6 to the QIAamp UCP Mini spin column (in a 2 ml collection tube) without wetting the rim. Close the cap, and centrifuge at 6000 x g (8000 rpm) for 1 min. Place the QIAamp Mini spin column in a clean 2 ml collection tube (provided), and discard the tube containing the filtrate. Close each spin column in order to avoid aerosol formation during centrifugation.

a. The precipitated DNA stays on the column; it is a silica membrane; proteins and other contaminants go through

15. Repeat step 7 by applying the remaining mixture from step 6 to the QIAamp UCP Mini spin column.

16. Carefully open the QIAamp UCP Mini spin column and add 600 µl Buffer APW1 without wetting the rim. Close the cap and centrifuge at 6000 x g (8000 rpm) for 1 min. Place the QIAamp UCP Mini spin column in a clean 2 ml collection tube (not provided), and discard the collection tube containing the filtrate.*

a. Contains 57% ethanol; allows contaminants to flow through

17. Carefully open the QIAamp UCP Mini spin column and add 750 µl Buffer APW2 without wetting the rim. Close the cap and centrifuge at full speed (20,000 x g; 14,000 rpm) for 3 min.

a. Contains 70% ethanol; more contaminants flow through

18. Recommended: Place the QIAamp UCP Mini spin column in a new 2 ml collection tube (not provided) and discard the old collection tube with the filtrate. Centrifuge at full speed for 1 min. This step helps to eliminate the chance of possible Buffer APW2 carryover.

19. Place the QIAamp UCP Mini column into a new 2 ml collection tube. Open the lid and incubate the assembly at 56°C for 3 min to dry the membrane completely.

a. Drying removes residual ethanol.

20. Place the QIAamp UCP Mini column in a clean 1.5 ml elution tube and discard the collection tube. Carefully apply 20–100 µl of 10:0.1 Tris:EDTA TE buffer to the center of the QIAamp UCP Mini membrane. Close the lid and incubate at room temperature for 1 min.

Important: Ensure that the elution buffer is equilibrated to room temperature. If elution is done in small volumes (<50 µl) the elution buffer has to be dispensed onto the center of the membrane for complete elution of bound DNA. Elution volume is flexible and can be adapted according to the requirements of downstream applications. The recovered eluate volume will up to 5 µl less than the elution volume applied onto the column.

21. Centrifuge at full speed (20,000 x g; 14,000 rpm) for 1 min to elute the DNA.

APPENDIX 2: CHAPTER 4 R CODE

```
#Statistical tests for Walsh et al, 2016

#Dana Walsh

#May 10, 2016

#Contains all the code used for data analysis in the paper above


library("gplots")

library("ggplot2")

library("plyr")

library("compositions")


#Merged and condensed OTU table from Explicet

taxa_counts <-

read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Workin

g_Analysis/OTU_tables/April_8_Explicet_norm_threshold_cntrls_72hrs_for_R.txt",

sep="\t", header=TRUE)

fam.counts <-

read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Workin

g_Analysis/OTU_tables/April_13_explicet_norm_thresh_fam_for_R.txt", sep="\t",

header=TRUE)


#Metadata file
```

```
meta <-
read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Workin
g_Analysis/Metadata/complete_meta_72hrs_feb_29.txt", sep="\t", header=TRUE) #Most
recent and complete metadata file - contains Baux scores
meta$ALI <- as.character(meta$ALI)

taxa.names <- as.character(fam.counts[,1])
patient.id <- colnames(fam.counts)
patient.id <- patient.id[-1]
fam.counts.only <- as.matrix(fam.counts[,-1])
colnames(fam.counts.only) <- NULL #Numbers alone before transposing - keeps data as
matrix type
fam.counts.only.t <- as.data.frame(t(fam.counts.only)) #Transpose the counts as a matrix;
still has controls
colnames(fam.counts.only.t) <- taxa.names #Add taxa back
rownames(fam.counts.only.t) <- patient.id
SampleID <- patient.id
fam.counts.only.t <- cbind(SampleID, fam.counts.only.t)
#rownames(fam.counts.only.t) <- NULL
fam.meta <- merge(meta, fam.counts.only.t, by=SampleID)
fam.meta.no.cntrls <- fam.meta[-1*1:2,]
fam.meta.no.cntrls <- fam.meta.no.cntrls[-9,]
```

```
taxa.names <- taxa_counts[,1]
taxa.counts.only <- taxa_counts[,-1]
ids <- colnames(taxa.counts.only)
taxa.numeric <- matrix(nrow=168, ncol=51)
for(i in 1:51){
  taxa.numeric[,i] <- as.numeric(taxa.counts.only[,i]) #Loops through taxa.counts.only
and makes all columns numeric
}
hist(taxa.numeric) #Very left-skewed; sparse
colnames(taxa.numeric) <- ids
taxa.numeric.names <- rbind(ids, taxa.numeric)
taxa.names <- taxa_counts[,1]
meta$SampleID <- as.character(meta$SampleID)
ali.id <- cbind(meta$SampleID, meta$ALI)
write.table(ali.id, "/Users/walshdm/Desktop/ali.id.txt", sep="\t") #Sorted appropriately in
Excel
ali.ids.sort <- read.table("/Users/walshdm/Desktop/ali.id.txt", sep="\t", header=TRUE)
ali.ids.sort <- ali.ids.sort[,-1]
ali.ids.sort.t <- t(ali.ids.sort)
taxa.numeric.ali <- rbind(ali.ids.sort.t, taxa.numeric)
taxa.numeric.ali <- taxa.numeric.ali[-1,]
taxa.numeric.ali <- taxa.numeric.ali[,-9] #Remove human control
taxa.numeric.ali <- taxa.numeric.ali[,-1*(49:50)] #Remove S. aureus and reagent controls
```

```
spock <- taxa.numeric.ali[1,]
taxa.numeric.ali <- taxa.numeric.ali[-1,]
taxa.numeric.ali.2 <- matrix(nrow=168, ncol=48)
for(i in 1:48){
  taxa.numeric.ali.2[,i] <- as.numeric(taxa.numeric.ali[,i]) #Loops through
taxa.counts.only and makes all columns numeric
}
log.ctr <- clr(taxa.numeric.ali.2)
wil.p <- apply(log.ctr,1,function(x,y){wilcox.test(x~y, paired=FALSE)$p.value},spock)
out.wil.p <- data.frame(Taxa=taxa.names, Wilcox=wil.p, CI=wil.ci, p.adj=p.adjust(wil.p,
method="bonferroni"))
write.table(out.wil.p, "/Users/walshdm/Desktop/out.wil.p.txt", sep="\t")
sig.taxa <- which(out.wil.p$Wilcox<=0.05)

#Nope - nothing of significance here (With low abundance removed)
#Remove low abundance OTUs, convert to log center scale, do wilcoxon again - need to
add ALI after log center scaling
abundance.avg <-cbind(taxa.names, as.data.frame(apply(taxa.numeric.ali.2, 1, mean)))
abundance.avg <- cbind(abundance.avg, taxa.numeric.ali.2)
write.table(abundance.avg, "/Users/walshdm/Desktop/abundance.avg.txt", sep="\t")
oneper <- read.table("/Users/walshdm/Desktop/abundance.avg.oneper.txt", sep="\t",
header=TRUE)
bugs <- oneper[,1]
```

```
oneper.num <- oneper[,-1]
log.ctr <- clr(oneper.num) #log centered scaling for multivariate analysis of
compositional abundance data
log.ctr.t <- t(as.data.frame(log.ctr))

#This loop makes ALI and None into numbers
spock <- taxa.numeric.ali[1,]
kirk <- vector(length=48)
for(i in 1:length(spock)){ #None = 0, ALI = 1
  if(spock[i]=="None"){
    kirk[i] <- 0
  }else{
    kirk[i] <- 1
  }
}

wil.p <- as.data.frame(apply(log.ctr,1,function(x,y){wilcox.test(x~y,
paired=FALSE)$p.value},spock))
wil.p.names <- cbind(bugs, wil.p)
wil.p.names <- cbind(wil.p.names, p.adjust(wil.p, method="bonferroni"))
colnames(wil.p.names) <- c("Taxa", "Wilcoxon P Value")
```

```
write.table(log.ctr, "/Users/walshdm/Desktop/log.ctr.txt", sep="\t")
log.ctr.sort <- read.table("/Users/walshdm/Desktop/log.ctr.txt", sep="\t", header=TRUE)
x <- log.ctr.sort[,1:24]
y <- log.ctr.sort[,25:48]

wil.p <- as.data.frame(apply(log.ctr,1,function(x,y){wilcox.test(x~y,
paired=FALSE)$p.value},status))
wil.ci <- as.data.frame(apply(log.ctr,1,function(x,y){wilcox.test(x~y, paired=FALSE,
conf.int=TRUE)$conf.int},status))


otu.stats <- cbind(taxa.names, wil)
colnames(otu.stats) <- c("OTU ID", "Wilcoxon P Value")
wil.p.adj <- as.data.frame(p.adjust(wil[,1], "bonferroni")) #Nothing significant
otu.stats <- cbind(otu.stats, wil.p.adj)
cis <- confint(wil)


#Alpha Diversity from Explicet
alpha <-
read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Workin
g_Analysis/72hrs/Diversity/R_alpha_diversity_5_9_16.txt", sep="\t", header=TRUE)
```

```
boxplot(alpha$Chao1.Mean~alpha$ALI_Status, main="Chao1 Diversity", ylab="Chao
Index")
wilcox.test(Chao1.Mean~ALI_Status, data=alpha, p.adj="bonferroni")

boxplot(alpha$Chao1.Mean~alpha$Prevotella, main="Chao1 Diversity", ylab="Chao
Index", xlab="Prevotella Detected", names=c("No", "Yes"))
wilcox.test(Chao1.Mean~Prevotella, data=alpha, p.adjust.methods="bonferroni")
#Chao1 is based on number of rare OTUs found in a sample; non-parametric

boxplot(alpha$Chao1.Mean~alpha$ALI_Status + alpha$Prevotella, main="Chao1
Diversity", ylab="Chao Index", xlab="Prevotella Detected", names=c("No", "No", "Yes",
"Yes"), col=c("blue", "red", "blue", "red"))
#blue = ALI, red = no ALI

chao.anova <- aov(Chao1.Mean~Prevotella + ALI_Status, data=alpha)
summary(chao.anova)
TukeyHSD(chao.anova)

chao.lm <- lm(Chao1.Mean~Prevotella + ALI_Status + Prevotella*ALI_Status,
data=alpha)
summary(chao.lm)

#ANOVA for anaerobes/aerobes
```

```
two_way <-
read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Workin
g_Analysis/72hrs/No_Paraprevotella/Files_for_R/Aerobe_ALI_Taxa_Count_Dec_1_AN
OVA.txt", sep="\t", header=TRUE)
o2.anova2 <- aov(Taxa_Per_Seq_Count~Group, data = two_way) #Needed to run post
tests - same results as anova on lm
summary(o2.anova2)
TukeyHSD(o2.anova2, which="Group")
confint(o2.anova2)
aerobe.mod = data.frame(Fitted = fitted(o2.anova2), Residuals = resid(o2.anova2), Group
= two_way$Group)
ggplot(aerobe.mod, aes(Fitted, Residuals, colour = Group)) + geom_point() +
ggtitle("Aerobic Capabilities") #Plots residuals in a single graph



#Plot by Aerobic Capabilities
boxplot(Taxa_Per_Seq_Count~Group, data = two_way, main = "Unique Taxa Per
Bacterial Aerobic Capabilities", ylab = "Taxa/Molecule Tag") #Makes a boxplot to show
differences among unique taxa per aerobic ability
ggplot(two_way, aes(Group, Taxa_Per_Seq_Count, colour = Treatment)) + geom_point()
ggplot(two_way, aes(Treatment, Taxa_Per_Seq_Count, colour = Group)) + geom_point()
boxplot(Taxa_Per_Seq_Count~Group*Treatment, data = two_way, las=2,
par(mar=c(10,5,4,2) + 0.1), main = "Unique Taxa Per Bacterial Aerobic Capabilities",
```

```
ylab = "Taxa/Molecule Tag") #Makes a boxplot to show differences among unique taxa
per aerobic ability

#Calculate SEM for aerobic ability
two_way_ALI <-
read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Workin
g_Analysis/72hrs/No_Paraprevotella/Files_for_R/Aerobe_ALI_Only_Taxa_Count_Dec_
1_ANOVA.txt", sep="\t", header=TRUE)
means.sem_oxygen <- ddply(two_way_ALI, c("Group", "Treatment"), summarise, mean
= mean(Taxa_Per_Seq_Count), sd = sd(Taxa_Per_Seq_Count), sem =
sd(Taxa_Per_Seq_Count)/sqrt(length(Taxa_Per_Seq_Count)))
means.sem_oxygen <- transform(means.sem_oxygen, lower = mean-sem, upper =
mean+sem)

#Make a bar plot with standard deviation error bars for aerobic ability
theme_set(theme_bw(base_size=14))
p <- ggplot(means.sem_oxygen, aes(fill=Treatment, y=mean, x=Group))
p + geom_bar(position="dodge", stat="identity") + geom_errorbar(aes(ymin=mean-sem,
ymax=mean+sem), width = 0.25, position=dodge) +
theme(axis.text.x=element_text(angle=90, hjust=1, size=9),
axis.title.y=element_text(size=10), axis.title.x=element_text(size=10),
legend.title=element_text(size=10)) + xlab("Aerobe/Anaerobe") + ylab("OTUs/Molecule
Tag Count")
```

```
dodge <- position_dodge(width=0.9)

#Bar graph for differences in abundance of bacteria detected as significantly different
among patients with and without ALI
sig_abundance <-
read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Workin
g_Analysis/72hrs/No_Paraprevotella/Files_for_R/Enriched_Taxa_ALI_Abundance_Dec
_4.txt", sep="\t", header=TRUE)
#View(sig_abundance)

#Remove taxa not identified as significant by LEfSe
sig_abundance_2 <- sig_abundance[-1,]
sig_abundance_2 <- sig_abundance_2[-4,]
sig_abundance_2 <- sig_abundance_2[-3,]

#Make a barplot of the abundance data
#png("/Users/walshdm/Documents/Manuscripts/R_figures/lefse_enriched_taxa.png",widt
h=1200,height=800, res=300)
theme_set(theme_gray(base_size=10))
qplot(x = factor(Bacteria), y = Abundance_Increase_In_ALI, fill = Oxygen, data =
sig_abundance_2, geom = "bar", stat = "identity", position = "dodge") + labs(x =
"Significantly Enriched Bacterial Taxa", y = "Abundance Increase in ALI (% of Total
Taxa)") + scale_fill_discrete(name = "Aerobic Ability", labels = c("Facultative
```

```
Anaerobe", "Obligate Anaerobe")) + theme(axis.text.x = element_text(angle = 90, hjust =
1, size=5)) #Makes the barplot from the variables indicated in sig_abundance
ggsave(file="/Users/walshdm/Documents/Meetings/SOT/2016/R_figures/lefse_enriched_
taxa.png", width=1200, height=800, limitsize=FALSE)

dev.off()

#Significant differences among Entero, Staph, Strep in ALI vs none
Entero.strep <- taxa_counts[1:2,]
top.3 <- rbind(Entero.strep, taxa_counts[12,])
top.3 <- top.3[,-10]
top.3 <- top.3[,-1*50:51]
otus <- top.3[,1]
otus.short <- c("Streptococcus", "Enterobacteriaceae", "Staphylococcus")
top.3 <- top.3[,-1]
colnames(top.3) <- spock
top.3.log.ctr <- clr(top.3)
rownames(top.3.log.ctr) <- otus.short
top.3.otus <- cbind(otus.short, top.3.log.ctr)
wrs.out<- apply(top.3.log.ctr,1,function(x,y){wilcox.test(x~y)$p.value},spock)
wilcox.top.3 <- cbind(otus.short, wrs.out)

library(reshape2)
```

```
top.3.melt <- melt(top.3.otus)
top.3.log.melt <- melt(top.3.log.ctr)
top.3.melt <- droplevels(top.3.melt)
top.3.anova <- aov(value~variable*otus, data=top.3.melt)
summary(top.3.anova) #Interaction between OTUs and ALI status is significant; run one-way anovas splitting by taxa
qqnorm(top.3.anova$residuals)
plot(top.3.anova$fitted.values, top.3.anova$residuals, xlab="Fitted", ylab="Residuals")
TukeyHSD(top.3.anova)

top.3.anova.taxa <- aov(value~otus, data=top.3.melt)
summary(top.3.anova.taxa)
TukeyHSD(top.3.anova.taxa)
Strep.top.3 <- subset(top.3.melt, variable ==
"k__Bacteria/p__Firmicutes/c__Bacilli/o__Lactobacillales/f__Streptococcaceae/g__Strep
tococcus/s__")
strep.anova <- aov(value~variable)
Entero.top.3 <- subset(top.3.melt, variable ==
"k__Bacteria/p__Proteobacteria/c__Gammaproteobacteria/o__Enterobacteriales/f__Enter
obacteriaceae/g__/s__")
```

```
Staph.top.3 <- subset(top.3.melt, variable ==
"k__Bacteria/p__Firmicutes/c__Bacilli/o__Bacillales/f__Staphylococcaceae/g__Staphylo
coccus/s__")

none.top.3 <- subset(top.3.melt, variable == "ALI")
none.anova <- aov(value~otus, data=none.top.3)
summary(none.anova)
TukeyHSD(none.anova)
ali.top.3 <- subset(top.3.melt, variable == "None")
```

APPENDIX 3: CHAPTER 5 R CODE

```
#Chapter 5 analysis

#Aug 3 2016

#Dana Walsh


library(compositions)

library("randomForest")

library("e1071")


meta <-

read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Workin

g_Analysis/Metadata/complete_meta_72hrs_feb_29.txt", sep="\t", header=TRUE) #Most

recent and complete metadata file - contains Baux scores


#Merged and condensed OTU table from Explicet

fam.counts <-

read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Workin

g_Analysis/OTU_tables/April_13_explicet_norm_thresh_fam_for_R.txt", sep="\t",

header=TRUE)

taxa.names <- fam.counts[,1] #Put taxonomy in a separate vector

abundances <- fam.counts[,-1]

abundances_t <- t(abundances)

abundances2 <- clr(abundances)
```

```
rownames(abundances2) <- taxa.names
abundances2_t <- t(abundances2)

#For heatmap
scout <- apply(abundances2_t,2,sum)
temp <- which(!(scout==0))
abundances3_t <- abundances2_t[,temp]

#Manhattan distance and Ward clustering
distance <- dist(abundances3_t, method="manhattan")
cluster <- hclust(distance, method="ward.D2") #Used this in R to make a heatmap

#K-means clustering
fit <- kmeans(abundances3_t, 3)

#K-means heatmap
distance <- dist(abundances3_t, method='manhattan')
fit <- kmeans(distance, 3)
heatmap.2(as.matrix(abundances3_t)[order(fit$cluster),], Rowv=NA, Colv=NA,
scale="none", trace="none", col=redgreen, xlab="Taxa", ylab="Patient ID",
margins=c(10,15))
abundances.clust <- cbind(fit$cluster, abundances3_t) #OTU table with K-means cluster
assignments
```

```
cluster.1.taxa <- abundances.clust[which(abundances.clust[,1]==1),]

cluster.2.taxa <- abundances.clust[which(abundances.clust[,1]==2),]

cluster.3.taxa <- abundances.clust[which(abundances.clust[,1]==3),]



#Add BMI

abundances4_t <- abundances3_t

rownames(abundances4_t) <- meta$BMI


#ALI Status of Clusters

cluster.1.ali <- meta.cluster[which(meta.cluster[,2]==1),19]

cluster.2.ali <- meta.cluster[which(meta.cluster[,2]==2),19]

cluster.3.ali <- meta.cluster[which(meta.cluster[,2]==3),19]


#Average BMI per Cluster

cluster.1 <- meta.cluster[which(meta.cluster[,1]==1),37]

cluster.2 <- meta.cluster[which(meta.cluster[,1]==2),37]

cluster.3 <- meta.cluster[which(meta.cluster[,1]==3),37]

cluster.1 <- cluster.1[-1]

cluster.2 <- cluster.2[-1]

cluster.2 <- cluster.2[-5]

cluster.3 <- cluster.3[-9]

cluster.3 <- cluster.3[-10]
```

```
cluster.1 <- as.numeric(levels(cluster.1))[cluster.1]

cluster.2 <- as.numeric(levels(cluster.2))[cluster.2]

cluster.3 <- as.numeric(levels(cluster.3))[cluster.3]

mean(cluster.1)

mean(cluster.2)

mean(cluster.3)


#Average Age per Cluster

cluster.1.age <- meta.cluster[which(meta.cluster[,1]==1),29]

cluster.2.age <- meta.cluster[which(meta.cluster[,1]==2),29]

cluster.3.age <- meta.cluster[which(meta.cluster[,1]==3),29]

cluster.1.age <- cluster.1[-1]

cluster.2.age <- cluster.2[-1]

cluster.2.age <- cluster.2[-5]

cluster.3.age <- as.numeric(levels(cluster.3.age))[cluster.3.age]

mean(cluster.1.age)

mean(cluster.2.age)

mean(cluster.3.age)


#Average SeqCount per Cluster

cluster.1.seq <- meta.cluster[which(meta.cluster[,1]==1),14]

cluster.2.seq <- meta.cluster[which(meta.cluster[,1]==2),14]

cluster.3.seq <- meta.cluster[which(meta.cluster[,1]==3),14]
```

```
cluster.1.seq <- cluster.1.seq[-1]

cluster.2.seq <- cluster.2.seq[-1]

cluster.2.seq <- cluster.2.seq[-5]

cluster.1.seq <- as.numeric(levels(cluster.1.seq))[cluster.1.seq]

cluster.2.seq <- as.numeric(levels(cluster.2.seq))[cluster.2.seq]

cluster.3.seq <- as.numeric(levels(cluster.3.seq))[cluster.3.seq]

mean(cluster.1.seq)

mean(cluster.2.seq)

mean(cluster.3.seq)


#Average IL-8 per Cluster

cluster.1.il8 <- meta.cluster[which(meta.cluster[,1]==1),54]

cluster.2.il8 <- meta.cluster[which(meta.cluster[,1]==2),54]

cluster.3.il8 <- meta.cluster[which(meta.cluster[,1]==3),54]

cluster.1.il8 <- cluster.1.il8[-1]

cluster.2.il8 <- cluster.2.il8[-1]

cluster.2.il8 <- cluster.2.il8[-5]

cluster.1.il8 <- as.numeric(levels(cluster.1.il8))[cluster.1.il8]

cluster.2.il8 <- as.numeric(levels(cluster.2.il8))[cluster.2.il8]

cluster.3.il8 <- as.numeric(levels(cluster.3.il8))[cluster.3.il8]

mean(cluster.1.il8)

mean(cluster.2.il8)

mean(cluster.3.il8)
```

```
#Days on Vent

cluster.1.vent <- meta.cluster[which(meta.cluster[,1]==1),43]

cluster.2.vent <- meta.cluster[which(meta.cluster[,1]==2),43]

cluster.3.vent <- meta.cluster[which(meta.cluster[,1]==3),43]

cluster.1.vent <- cluster.1.vent[-1]

cluster.2.vent <- cluster.2.vent[-1]

cluster.2.vent <- cluster.2.vent[-5]

cluster.1.vent <- as.numeric(levels(cluster.1.vent))[cluster.1.vent]

cluster.2.vent <- as.numeric(levels(cluster.2.vent))[cluster.2.vent]

cluster.3.vent <- as.numeric(levels(cluster.3.vent))[cluster.3.vent]

mean(cluster.1.vent)

mean(cluster.2.vent)

mean(cluster.3.vent)


#IL12p70

cluster.1.il12 <- meta.cluster[which(meta.cluster[,1]==1),50]

cluster.2.il12 <- meta.cluster[which(meta.cluster[,1]==2),50]

cluster.3.il12 <- meta.cluster[which(meta.cluster[,1]==3),50]

cluster.1.il12 <- cluster.1.il12[-1]

cluster.2.il12 <- cluster.2.il12[-1]

cluster.2.il12 <- cluster.2.il12[-5]

cluster.1.il12 <- as.numeric(levels(cluster.1.il12))[cluster.1.il12]

cluster.2.il12 <- as.numeric(levels(cluster.2.il12))[cluster.2.il12]
```

```
cluster.3.il12 <- as.numeric(levels(cluster.3.il12))[cluster.3.il12]
mean(cluster.1.il12)
mean(cluster.2.il12)
mean(cluster.3.il12)

MT_cntrls <-
read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Workin
g_Analysis/72hrs/MT_cntrls.txt", sep="\t", header=TRUE)
meta.cluster <- cbind(MT_cntrls$MT_Count, meta.cluster)

#MT_Counts
cluster.1.mt <- meta.cluster[which(meta.cluster[,2]==1),1]
cluster.2.mt <- meta.cluster[which(meta.cluster[,2]==2),1]
cluster.3.mt <- meta.cluster[which(meta.cluster[,2]==3),1]
mean(cluster.1.mt)
mean(cluster.2.mt)
mean(cluster.3.mt)

#Baux Score
cluster.1.baux <- meta.cluster[which(meta.cluster[,2]==1),14]
cluster.2.baux <- meta.cluster[which(meta.cluster[,2]==2),14]
cluster.3.baux <- meta.cluster[which(meta.cluster[,2]==3),14]
cluster.1.baux <- cluster.1.baux[-1]
```

```
cluster.2.baux <- cluster.2.baux[-1]

cluster.2.baux <- cluster.2.baux[-5]

cluster.1.baux <- as.numeric(levels(cluster.1.baux))[cluster.1.baux]

cluster.2.baux <- as.numeric(levels(cluster.2.baux))[cluster.2.baux]

cluster.3.baux <- as.numeric(levels(cluster.3.baux))[cluster.3.baux]

mean(cluster.1.baux)

mean(cluster.2.baux)

mean(cluster.3.baux)


#Hierarchical clustering

rownames(abundances) <- taxa.names


#3D PCA plot

library(nsprcomp)

#Non-negative sparse PCA (NSPCA)

burn.nspca <- nsprcomp(abundances, nneg=TRUE, scale.=TRUE)


#3D Scatterplots

library("scatterplot3d")

#Colored by ALI status

png("/Users/walshdm/Documents/Dissertation/Chpt_5/pca.png", width=1200,

height=1200, res=300)
```

```
scatterplot3d(burn.nspca$rotation[,1], burn.nspca$rotation[,2], burn.nspca$rotation[,3],
main="ALI Status", color=meta$ALI_color, pch=16, xlab="PC1", ylab="PC2",
zlab="PC3", cex.axis=0.5, cex.lab=0.7)
legend("topright", legend=paste(c('ALI', 'None', 'Human', 'Bacteria', 'Control')), pch=16,
col=c("blue", "red", "black", "yellow", "pink"), cex=0.7, inset=c(0.1, 0.2), bty="n")
dev.off()

#DAPC
library("adegenet")
grp <- find.clusters(abundances_t, max.n.clust=40)
dapc1 <- dapc(abundances_t, grp$grp)
scatter(dapc1)

#Hierarchical clustering
hc <- hclust(dist(abundances_t), method="ward.D2")
par(mfrow=c(1,1))
plot(hc, cex=.6)

#Cut the tree
rect.hclust(hc, k=3)

#Define clusters
mycl <- cutree(hc, k=3)
```

```
#MT_Count
MTs <-
read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Workin
g_Analysis/72hrs/MT_Tag_Counts.txt", sep="\t", header=TRUE)
MT_Counts <- MTs$MT_Count



#Random forest
meta.dapc.rf <- meta.cluster[,-(1:12)]
meta.dapc.rf <- meta.dapc.rf[-(1:2),] #Removes controls
meta.dapc.rf <- meta.dapc.rf[-(9),] #Removes controls
meta.dapc.rf <- meta.dapc.rf[,-2] #Removes SeqCount
meta.dapc.rf <- meta.dapc.rf[,-(4:15)] #Removes non-numeric columns
meta.dapc.rf <- meta.dapc.rf[,-(5:11)] #Removes non-numeric columns
meta.dapc.rf <- meta.dapc.rf[,-(7:8)] #Removes non-numeric columns
meta.dapc.rf <- meta.dapc.rf[,-(10:13)] #Removes non-numeric columns
meta.dapc.rf <- meta.dapc.rf[,-17] #Removes non-numeric columns
meta.dapc.rf <- meta.dapc.rf[,-(19:23)] #Removes non-numeric columns
meta.dapc.rf <- meta.dapc.rf[,-(20:22)] #Removes non-numeric columns
meta.dapc.rf <- cbind(groups.dapc, meta.dapc.rf)
meta.dapc.rf <- meta.dapc.rf[,-24] #Removes non-numeric columns
meta.dapc.rf <- cbind(MT_Counts, meta.dapc.rf)
```

```
choo_train <- meta.dapc.rf[sample(1:nrow(meta.dapc.rf), 38, replace=FALSE),]
tune.cluster <- tune.randomForest(groups.dapc~., data=choo_train, mtry=c(2.8, 5.6,
11.1), ntree=c(250,500,1000), na.rm=TRUE)
x<-summary(tune.cluster)
rf.cluster <- randomForest(groups.dapc~., data=meta.dapc.rf, importance=TRUE,
na.action=na.omit, mtry=as.numeric(x$best.parameters[1]),
ntree=as.numeric(x$best.parameters[2]))
varImpPlot(rf.cluster, cex=.7)
varImpPlot(rf.cluster, type=2, cex=.7)
```

APPENDIX 4: CHAPTER 6 R CODE

#Prediction of Functional Changes Among Bacterial Networks in Patients with

#$PaO_2/FiO_2 \leq 300$

#Code for paper #2

#Dana Walsh

#July 14 2016

```
library(randomForest)
library(e1071)
library(gmodels)
library(matrixStats)
library(compositions)
library(gplots)

#******all_samp is the OTU table used for final network analysis with
SparCC*********
#Import tables with thresholds and averaged duplicates (including controls)
jan_1 <-
read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Workin
g_Analysis/OTU_tables/Threshold_OTU_nums/Jan_1a_averaged_otu_nums.txt",
sep="\t", header=TRUE, stringsAsFactors = FALSE)
jan_2 <-
read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Workin
```

```
g_Analysis/OTU_tables/Threshold_OTU_nums/Jan_2a_averaged_otu_nums.txt",
sep="\t", header=TRUE, stringsAsFactors = FALSE)
dec <-
read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Workin
g_Analysis/OTU_tables/Threshold_OTU_nums/Dec_averaged_otu_nums.txt", sep="\t",
header=TRUE, stringsAsFactors = FALSE)
#cntrls <-
read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Workin
g_Analysis/Picrust/Cntrls_avgs_no_OTUnum_april_8.txt", sep="\t", header=TRUE)
all_jan <- merge(jan_1, jan_2, by="OTUId", all=TRUE)
all_samp <- merge(all_jan, dec, by="OTUId", all=TRUE)
for(i in 1:nrow(all_samp)){ #Converts NAs to zero counts
  all_samp[i, is.na(all_samp[i,])] <- 0
}
write.table(all_samp,
"/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_Analysis
/Community_Analysis/OTU_tables/all_samp_taxa.txt", sep="\t")
all_samp_taxa <- all_samp
# taxa <- as.character(all_samp_taxa$taxonomy)
# all_samp_taxa <- all_samp_taxa[,-25]
# all_samp_taxa <- cbind(taxa, all_samp_taxa)

taxa <- all_samp$taxonomy
```

```
all_samp <- all_samp[,-25]
write.table(all_samp,
"/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_Analysis
/Community_Analysis/OTU_tables/all_samp.txt", sep="\t")

#Add taxonomy back to all_samp
taxonomy_2 <-
read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Workin
g_Analysis/Community_Analysis/OTU_tables/all_otuids_taxa.txt", sep="\t",
header=TRUE, stringsAsFactors = FALSE)
taxa.names.2 <- list() #Adds taxonomy to appropriate OTU ID - some OTUs were
missing from above table
for(i in 1:nrow(all_samp_taxa)){
  taxa.names.2[[i]] <- which(all_samp_taxa[i,1]== taxonomy_2[,1])
  if (all_samp_taxa[i,25]==0){
    all_samp_taxa[i,25] = taxonomy_2[taxa.names.2[[i]],2]
  }
}
write.table(all_samp_taxa,"/Users/walshdm/Documents/Burn_Study/Sequencing_Data_A
nalysis/Working_Analysis/Community_Analysis/OTU_tables/all_samp_taxa.txt",
sep="\t")

#Cluster assignments from Natalie - want to know their abundances
```

```
ALI.comm <-
read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Workin
g_Analysis/Community_Analysis/Community_Assignments/Communities_ALI.txt",
sep="\t", header=TRUE)
#ALI <- as.matrix(ALI.comm)
No.ALI.comm <-
read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Workin
g_Analysis/Community_Analysis/Community_Assignments/Communities_No_ALI.txt",
sep="\t", header=TRUE)

#Add Abundances In
taxonomy <- all_samp_taxa$taxonomy
OTUIDs <- all_samp$OTUId
all_samp_norm <- all_samp_taxa[,-1]
all_samp_norm <- all_samp_norm[,-24]
all_samp_norm <- as.matrix(all_samp_norm)
all_samp_norm <- scale(all_samp_norm, center=F, scale=colSums(all_samp_norm))
#Normalizes the table to relative abundances
all_samp_norm <- cbind(OTUIDs, taxonomy, all_samp_norm)
ALI.norm <- cbind(OTUIDs, taxonomy, all_samp_norm[,4], all_samp_norm[,7],
all_samp_norm[,9], all_samp_norm[,12], all_samp_norm[,13], all_samp_norm[,14],
all_samp_norm[,16], all_samp_norm[,19:22], all_samp_norm[,26], all_samp_norm[,27],
```

```
all_samp_norm[,28], all_samp_norm[,29], all_samp_norm[,32], all_samp_norm[,36:40],
all_samp_norm[,42])
ALI.colnames <- c("OTUIDs", "taxonomy", "X41", "X74", "X79", "X93", "X104",
"X110", "X146", "X172", "X176", "X182", "X202", "X209", "X219", "X229", "X238",
"X307", "X356", "X368", "X372", "X2", "X25", "X149")
colnames(ALI.norm) <- ALI.colnames
No.ALI.norm <- cbind(OTUIDs, taxonomy, all_samp_norm[,3], all_samp_norm[,5],
all_samp_norm[,6], all_samp_norm[,8], all_samp_norm[,10], all_samp_norm[,11],
all_samp_norm[,15], all_samp_norm[,17:18], all_samp_norm[,23:25],
all_samp_norm[,30:31], all_samp_norm[,33:35], all_samp_norm[,41],
all_samp_norm[,43])
No.ALI.colnames <- c("OTUIDs", "taxonomy", "X14", "X63", "X66", "X78", "X81",
"X89", "X124", "X153", "X169", "X207", "X364", "X380", "X246", "X255", "X314",
"X317", "X337", "X128", "X186")
colnames(No.ALI.norm) <- No.ALI.colnames

abundance.matches <- list() #Finds rows in ALI.norm that match ALI.comm
for(i in 1:nrow(ALI.comm)){
  abundance.matches[[i]] <- which(ALI.norm[,1]==ALI.comm[i,1])
}
abundance.match.unlist <- unlist(abundance.matches)
ALI.comm.abundances <- cbind(ALI.comm, ALI.norm[abundance.match.unlist,-1])
ALI.comm.abundances <- ALI.comm.abundances[order(ALI.comm.abundances[,2]),]
```

```
write.table(ALI.comm.abundances,
"/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_Analysis
/Community_Analysis/Community_Assignments/ALI.comm.abundances.txt", sep="\t")

abundance.matches.2 <- list() #Finds rows in No.ALI.norm that match No.ALI.comm
for(i in 1:nrow(No.ALI.comm)){
  abundance.matches.2[[i]] <- which(No.ALI.norm[,1]==No.ALI.comm[i,1])
}
abundance.match.unlist.2 <- unlist(abundance.matches.2)
No.ALI.comm.abundances <- cbind(No.ALI.comm,
No.ALI.norm[abundance.match.unlist.2,-1])
No.ALI.comm.abundances <-
No.ALI.comm.abundances[order(No.ALI.comm.abundances[,2]),]
write.table(No.ALI.comm.abundances,
"/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_Analysis
/Community_Analysis/Community_Assignments/No.ALI.comm.abundances.txt",
sep="\t")


#PICRUSt total predicted functions for all possible taxa/OTU IDs
OTU_functions <-
read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Workin
```

```
g_Analysis/72hrs/PICRUSt/threshold_15/predicted_traits_15_for_R.txt", sep="\t",
header=TRUE)
OTU_nums <- as.character(OTU_functions$OTU_IDs)
OTU_split <- strsplit(OTU_nums, "_") #Splits the numbers from 'OTU'
OTU_split_2 <- matrix(unlist(OTU_split), ncol=2, byrow=TRUE) #Unlists OTU_split
and turns it into a matrix
OTU_functions_2 <- OTU_functions
OTU_functions_2[,1] <- OTU_split_2[,2] #Adds OTU ID numbers without 'OTU'

#Match OTU IDs and add predicted function counts to community assignments
matches <- list() #Finds rows in OTU_functions_2 that match ALI.comm
for(i in 1:nrow(ALI.comm)){
  matches[[i]] <- which(OTU_functions_2[,1]==ALI.comm[i,1])
}
match.unlist <- unlist(matches)
ALI.comm.funct <- cbind(ALI.comm, OTU_functions_2[match.unlist,-1])
#ALI.comm.funct is matrix to use for downstream analysis of functions per community
for patients with ALI
write.table(ALI.comm.funct,
"/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_Analysis
/Community_Analysis/Community_Assignments/ALI.comm.funct.txt", sep="\t")

#Match OTU IDs and add predicted function counts to community assignments
```

```
matches_none <- list() #Finds rows in OTU_functions_2 that match No.ALI.comm
for(i in 1:nrow(No.ALI.comm)){
  matches_none[i] <- which(OTU_functions_2[,1]==No.ALI.comm[i,1])
}
match.none.unlist <- unlist(matches_none)
No.ALI.comm.funct <- cbind(No.ALI.comm, OTU_functions_2[match.none.unlist,-1])
#No.ALI.comm.funct is matrix to use for downstream analysis of functions per
community for patients with no ALI
write.table(No.ALI.comm.funct,
"/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_Analysis
/Community_Analysis/Community_Assignments/No.ALI.comm.funct.txt", sep="\t")

#The only difference between the two files is the community assignment number


#Which functions predict community assignment? For patients with ALI

scout <- apply(ALI.comm.funct,2,sum)
temp <- which(!(scout==0))
ALI.comm.funct2 <- ALI.comm.funct[,temp] #Removes columns which sum to zero
(ALI)
ALI.comm.funct2.summary <- as.matrix(apply(ALI.comm.funct2,2,summary))
```

```
#Add taxa names back to list of communities + functions
taxa.names.2 <- list() #Adds taxonomy to appropriate OTU ID
for(i in 1:nrow(ALI.comm.funct2)){
  taxa.names.2[[i]] <- which(ALI.comm.funct2[i,1]==taxonomy_2[,1])
}

taxa.names.2.unlist <- matrix(unlist(taxa.names.2), ncol=1, byrow=TRUE)
ALI.comm.funct2 <- cbind(taxonomy_2[taxa.names.2.unlist,2], ALI.comm.funct2)

taxa.names.3 <- list() #Adds taxonomy to appropriate OTU ID
for(i in 1:nrow(No.ALI.comm.funct2)){
  taxa.names.3[[i]] <- which(No.ALI.comm.funct2[i,1]==taxonomy_2[,1])
}

taxa.names.3.unlist <- matrix(unlist(taxa.names.3), ncol=1, byrow=TRUE)
No.ALI.comm.funct2 <- cbind(taxonomy_2[taxa.names.3.unlist,2],
No.ALI.comm.funct2)

ALI.comm.funct2.clr <- cbind(ALI.comm.funct2[,(1:3)], clr(ALI.comm.funct2[,-(1:3)]))
#Centered log ratio transformation; adds taxa, OTU number and community assignments
back in
No.ALI.comm.funct2.clr <- cbind(No.ALI.comm.funct2[,(1:3)],
clr(No.ALI.comm.funct2[,-(1:3)]))
```

```
write.table(ALI.comm.funct2.clr,"/Users/walshdm/Documents/Burn_Study/Sequencing_
Data_Analysis/Working_Analysis/Community_Analysis/Heatmaps/ALI.comm.funct2.clr
.txt", sep="\t")
write.table(No.ALI.comm.funct2.clr,"/Users/walshdm/Documents/Burn_Study/Sequenci
ng_Data_Analysis/Working_Analysis/Community_Analysis/Heatmaps/No.ALI.comm.fu
nct2.clr.txt", sep="\t")
#The above files were used with hierarchical clustering to create a large heatmap of all
functions for all OTUs (minus those that sum to zero) - this is the same graph for ALI and
No ALI OTUs
#Make heatmaps per community assignment


#For heatmaps
taxa.ali <- strsplit(as.character(ALI.comm.funct2.clr[,1]), ";") #Each taxonomy level can
be indexed separately
taxa.no.ali <- strsplit(as.character(No.ALI.comm.funct2.clr[,1]), ";")


ALI.comm.funct2.clr <- ALI.comm.funct2.clr[,-(1:2)]
No.ALI.comm.funct2.clr <- No.ALI.comm.funct2.clr[,-(1:2)]


fam <- matrix(data=NA, nrow=372, ncol=1)
for(i in 1:nrow(ALI.comm.funct2.clr)){
```

```
  fam[i,] <- taxa.ali[[i]][5]
} #Add selected taxonomic level for heatmap

#rownames(ALI.comm.funct2.clr) <- fam[,1] Can't - duplicate names; use OTU
ali.community <- ALI.comm.funct2.clr[,1]
ALI.comm.funct2.clr <- ALI.comm.funct2.clr[,-1]

write.table(ALI.comm.funct2.clr,"/Users/walshdm/Documents/Burn_Study/Sequencing_
Data_Analysis/Working_Analysis/Community_Analysis/Heatmaps/ALI.comm.funct2.clr
.txt", sep="\t")
write.table(No.ALI.comm.funct2.clr,"/Users/walshdm/Documents/Burn_Study/Sequenci
ng_Data_Analysis/Working_Analysis/Community_Analysis/Heatmaps/No.ALI.comm.fu
nct2.clr.txt", sep="\t")

scout_2 <- apply(No.ALI.comm.funct,2,sum)
temp_2 <- which(!(scout_2==0))
No.ALI.comm.funct2 <- No.ALI.comm.funct[,temp_2] #Removes columns which sum to
zero (No ALI)
No.ALI.comm.funct2.summary <- as.matrix(apply(No.ALI.comm.funct2,2,summary))

#Heatmap with all functions and ALI + No ALI
ALI.Status <- rep(c("ALI", "No.ALI"), each=372)
```

```
All.comm.funct <- rbind(ALI.comm.funct2, No.ALI.comm.funct2)
All.comm.funct <- cbind(ALI.Status, All.comm.funct)
All.comm.funct.clr <- clr(All.comm.funct[,-(1:3)]) #Centered log ratio transformation - deals with compositional data
write.table(All.comm.funct.clr,
"/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_Analysis/Community_Analysis/Heatmaps/All.comm.funct.clr.txt", sep="\t")

write.table(All.comm.funct,
"/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_Analysis/Community_Analysis/Heatmaps/All.comm.funct.txt", sep="\t")

heatmap2 <-
read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_Analysis/Community_Analysis/Heatmaps/All.comm.funct.txt", sep="\t",
header=TRUE)
heatmap <-
read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_Analysis/Community_Analysis/Heatmaps/All.comm.funct.clr.txt", sep="\t",
header=TRUE)

#Manhattan distance, Ward clustering, heatmap
distance <- dist(All.comm.funct.clr, method="manhattan")
```

```
cluster <- hclust(distance, method="ward.D2")

heatmap.2(All.comm.funct.clr, Rowv=as.dendrogram(cluster), Colv=TRUE,
scale="column", trace="none", col=redgreen, xlab="OTU", ylab="Predicted Function",
margins=c(10,15))



#Which functions predict ALI vs None? Combined functional predictions for patients
with and without ALI
#Which of the most variable functions predict ALI?
All.var <- as.matrix(apply(All.comm.funct,2,var)) #542 > 0.3; use for random forest
features
All.rf <- cbind(ALI.Status, All.comm.funct[,which(All.var>0.3)]) #Use this for random
forest on Kure
All.rf.kure <- All.rf[,-(2:3)]
All.rf.comm <- All.rf[,-(1:2)]
write.table(All.rf.kure,
"/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_Analysis
/Community_Analysis/Community_Assignments/All.rf.kure.txt", sep="\t")

#Random Forest for most variable KOs that predict ALI vs No ALI
set.seed(18)
```

```
choo_train <- All.rf.kure[sample(1:nrow(All.rf.kure), nrow(All.rf.kure)*0.5,
replace=FALSE),]
tune <- tune.randomForest(ALI.Status~., data=choo_train, mtry=c(2.8, 5.6, 11.1),
ntree=c(250,500), na.rm=TRUE)
x<-summary(tune)

#Entire data set
rf.ALI <- randomForest(ALI.Status~., data=All.rf.kure, importance=TRUE,
na.action=na.omit, mtry=as.numeric(x$best.parameters[1]),
ntree=as.numeric(x$best.parameters[2]))
png("/nas02/home/w/a/walshdm/R_Analyses/RF_ALI/rf_ALI.png", width=800,
height=1200, res=300)
varImpPlot(rf.ALI, cex=.7)
dev.off()

#Which functions that are most prevalent predict ALI?
All.sum <- as.matrix(apply(All.comm.funct[,-(1:3)],2,sum)) #726 columns have sums
greater than 410; use these
All.sum.rf <- cbind(ALI.Status, All.comm.funct[,which(All.sum>410)])
All.sum.rf <- All.sum.rf[,-2]
set.seed(2)
choo_train4 <- All.sum.rf[sample(1:nrow(All.sum.rf), nrow(All.sum.rf)*0.5,
replace=FALSE),]
```

```
tune4 <- tune.randomForest(ALI.Status~., data=choo_train4, mtry=c(2.8, 5.6, 11.1),
ntree=c(250,500), na.rm=TRUE)
x4<-summary(tune4)

#Entire data set
rf.ALI.sum <- randomForest(ALI.Status~., data=All.sum.rf, importance=TRUE,
na.action=na.omit, mtry=as.numeric(x4$best.parameters[1]),
ntree=as.numeric(x4$best.parameters[2]))
varImpPlot(rf.ALI.sum, cex=.7)

#Random forest for most variable KOs (same as above) that predict community
assignments (ALI & No ALI together)
All.rf.comm <- cbind(ALI.Status, All.rf.comm)
ALI.comm.char <- paste(All.rf.comm$ALI.Status, All.rf.comm$Community, sep="_")
All.rf.comm <- cbind(ALI.comm.char, All.rf.comm)
All.rf.comm <- All.rf.comm[,-(2:3)]

set.seed(25)
choo_train2 <- All.rf.comm[sample(1:nrow(All.rf.comm), nrow(All.rf.comm)*0.5,
replace=FALSE),]
tune2 <- tune.randomForest(ALI.comm.char~., data=choo_train2, mtry=c(2.8, 5.6, 11.1),
ntree=c(250,500), na.rm=TRUE)
x2<-summary(tune2)
```

```
rf.ALI.comm <- randomForest(ALI.comm.char~., data=All.rf.comm, importance=TRUE,
na.action=na.omit, mtry=as.numeric(x2$best.parameters[1]),
ntree=as.numeric(x2$best.parameters[2]))
varImpPlot(rf.ALI.comm, cex=.7)

All.rf.comm <- cbind(ALI.Status, All.rf.comm)
ALI.comm.char <- paste(All.rf.comm$ALI.Status, All.rf.comm$community, sep="_")
All.rf.comm <- cbind(ALI.comm.char, All.rf.comm)
All.rf.comm <- All.rf.comm[,-(2:3)]

#Random forest for most prevalent KOs with ALI and No ALI together
All.sum.rf <- All.sum.rf[,-1]
All.sum.rf <- cbind(ALI.comm.char, All.sum.rf)
set.seed(50)
choo_train3 <- All.sum.rf[sample(1:nrow(All.sum.rf), nrow(All.sum.rf)*0.5,
replace=FALSE),]
tune3 <- tune.randomForest(ALI.comm.char~., data=choo_train3, mtry=c(2.8, 5.6, 11.1),
ntree=c(250,500), na.rm=TRUE)
x3<-summary(tune3)
rf.All.sum <- randomForest(ALI.comm.char~., data=All.sum.rf, importance=TRUE,
na.action=na.omit, mtry=as.numeric(x3$best.parameters[1]),
ntree=as.numeric(x3$best.parameters[2]))
varImpPlot(rf.All.sum, cex=.7)
```

```
ALI.comm.var <- as.data.frame(t(All.rf.comm)) #Transpose table for Lefse
ALI.comm.most <- as.data.frame(t(All.sum.rf)) #Transpose table for Lefse

write.table(ALI.comm.var,
"/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_Analysis
/Community_Analysis/Lefse/ALI.comm.var.txt", sep="\t")
write.table(ALI.comm.most,
"/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_Analysis
/Community_Analysis/Lefse/ALI.comm.most.txt", sep="\t")


#Summary stats for each community in patients who have ALI
comm_1 <- ALI.comm.funct2[which(ALI.comm.funct2$community==1),]
comm_1_sum <- as.matrix(apply(comm_1,2,sum))
ALI.comm.1 <- comm_1[,which(!(comm_1_sum==0))] #Community 1 functions with no
zero total columns

comm_2 <- ALI.comm.funct2[which(ALI.comm.funct2$community==2),]
comm_2_sum <- as.matrix(apply(comm_2,2,sum))
ALI.comm.2 <- comm_2[,which(!(comm_2_sum==0))] #Community 2 functions with no
zero total columns
```

```
comm_3 <- ALI.comm.funct2[which(ALI.comm.funct2$community==3),]
comm_3_sum <- as.matrix(apply(comm_3,2,sum))
ALI.comm.3 <- comm_3[,which(!(comm_3_sum==0))] #Community 3 functions with no zero total columns

comm_4 <- ALI.comm.funct2[which(ALI.comm.funct2$community==4),]
comm_4_sum <- as.matrix(apply(comm_4,2,sum))
ALI.comm.4 <- comm_4[,which(!(comm_4_sum==0))] #Community 4 functions with no zero total columns

#Summary stats for each community within patients without ALI
No_comm_1 <- No.ALI.comm.funct2[which(No.ALI.comm.funct2$community==1),]
No_comm_1_sum <- as.matrix(apply(No_comm_1,2,sum))
No.ALI.comm.1 <- No_comm_1[,which(!(No_comm_1_sum==0))] #Community 1 functions with no zero total columns

No_comm_2 <- No.ALI.comm.funct2[which(No.ALI.comm.funct2$community==2),]
No_comm_2_sum <- as.matrix(apply(No_comm_2,2,sum))
No.ALI.comm.2 <- No_comm_2[,which(!(No_comm_2_sum==0))] #Community 2 functions with no zero total columns

No_comm_3 <- No.ALI.comm.funct2[which(No.ALI.comm.funct2$community==3),]
No_comm_3_sum <- as.matrix(apply(No_comm_3,2,sum))
```

```
No.ALI.comm.3 <- No_comm_3[,which(!(No_comm_3_sum==0))] #Community 3
functions with no zero total columns

No_comm_4 <- No.ALI.comm.funct2[which(No.ALI.comm.funct2$community==4),]
No_comm_4_sum <- as.matrix(apply(No_comm_4,2,sum))
No.ALI.comm.4 <- No_comm_4[,which(!(No_comm_4_sum==0))] #Community 4
functions with no zero total columns
No.ALI.comm.4.summary <- as.matrix(apply(No.ALI.comm.4,2,summary))

#Contingency table
All.comm <- ALI.comm
colnames(All.comm) <- c("OTU", "ALI_Comm")
No_ALI_comm <- No.ALI.comm[,2]
All.comm <- cbind(All.comm, No_ALI_comm)
cont.table <- as.data.frame(CrossTable(All.comm$ALI_Comm,
All.comm$No_ALI_comm, prop.t=TRUE, prop.r=TRUE, prop.c=TRUE, chisq =
TRUE))

#Overlap between ALI and No ALI OTU community assignments
otus.ali <- as.character(ALI.comm[,1])
otus.no.ali <- as.character(No.ALI.comm[,1])
ALI.comm[,1] <- otus.ali
No.ALI.comm[,1] <- otus.no.ali
```

```
ALI.comm.1 <- ALI.comm[which(ALI.comm[,2]==1),]
ALI.comm.2 <- ALI.comm[which(ALI.comm[,2]==2),]
ALI.comm.3 <- ALI.comm[which(ALI.comm[,2]==3),]
ALI.comm.4 <- ALI.comm[which(ALI.comm[,2]==4),]

No.ALI.comm.1 <- No.ALI.comm[which(No.ALI.comm[,2]==1),]
No.ALI.comm.2 <- No.ALI.comm[which(No.ALI.comm[,2]==2),]
No.ALI.comm.3 <- No.ALI.comm[which(No.ALI.comm[,2]==3),]
No.ALI.comm.4 <- No.ALI.comm[which(No.ALI.comm[,2]==4),]

#Matches input to all No.ALI community assignments and returns No.ALI.comm
matches
No.ALI.overlap <- function(x){
same <- list()
for(i in 1:nrow(x)){
  same[i] <- which(x[i,1]==No.ALI.comm[,1])
  same.1 <- unlist(as.matrix(same))
  matches <- No.ALI.comm[same.1,]
  }
 return(matches)
}
```

```
ALI.1.matches <- No.ALI.overlap(ALI.comm.1) #These contain the community
assignments for matching OTUs within No.ALI.comm
ALI.2.matches <- No.ALI.overlap(ALI.comm.2)
ALI.3.matches <- No.ALI.overlap(ALI.comm.3)
ALI.4.matches <- No.ALI.overlap(ALI.comm.4)

ALI.overlap <- function(x){
 same <- list()
 for(i in 1:nrow(x)){
  same[i] <- which(x[i,1]==ALI.comm[,1])
  same.1 <- unlist(as.matrix(same))
  matches <- ALI.comm[same.1,]
 }
 return(matches)
}
No.ALI.1.matches <- ALI.overlap(No.ALI.comm.1) #These contain the community
assignments for matching OTUs within ALI.comm
No.ALI.2.matches <- ALI.overlap(No.ALI.comm.2)
No.ALI.3.matches <- ALI.overlap(No.ALI.comm.3)
No.ALI.4.matches <- ALI.overlap(No.ALI.comm.4)

#These match the contingency table - can identify which OTUs overlap
```

```
#________This has to be done on Kure - file is too large_____________________
#Split data into train set
choo_train <- ALI.comm.funct[sample(1:nrow(ALI.comm.funct),
nrow(ALI.comm.funct)*0.5, replace=FALSE),]
tune_ALI <- tune.randomForest(community~., data=choo_train, mtry=c(2.8, 5.6, 11.1),
ntree=c(250,500,1000), na.rm=TRUE)
x<-summary(tune_ALI)

#Entire data set
rf.ALI <- randomForest(community~., data=ALI.comm.funct[,-2], importance=TRUE,
na.action=na.omit, mtry=as.numeric(x$best.parameters[1]),
ntree=as.numeric(x$best.parameters[2]))
importance(rf.ALI)
varImpPlot(rf.ALI, cex=.7)
#_____________________________________________________________________________

#This isn't tested - use to add taxonomy to All.comm (from taxonomy_2)
taxa.names.2 <- list() #Adds taxonomy to appropriate OTU ID - some OTUs were
missing from above table
for(i in 1:nrow(all_samp_taxa)){
  taxa.names.2[[i]] <- which(all_samp_taxa[i,1]== taxonomy_2[,1])
  if (all_samp_taxa[i,25]==0){
    all_samp_taxa[i,25] = taxonomy_2[taxa.names.2[[i]],2]
```

```
  }
}

#_________This has to be done on Kure - file is too large_____________________
#Code submitted to Kure: RF.paper.2.kure.R



#PICRUSt data per patient sample
#Analysis of predicted functions per patient - importing and preprocessing
Patient_functions_15 <-
read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Workin
g_Analysis/72hrs/PICRUSt/threshold_15/predicted_metagenome_15_for_R.txt",
sep="\t", header=TRUE)
Patient_functions_35 <-
read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Workin
g_Analysis/72hrs/PICRUSt/threshold_35/predicted_metagenome_35_for_R.txt",
sep="\t", header=TRUE)
All_patient_functions <- merge(Patient_functions_15, Patient_functions_35, by="KO")
write.table(All_patient_functions,
file="/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_An
alysis/72hrs/PICRUSt/all_patient_functions.txt", sep="\t")
KOs <- as.character(All_patient_functions$KO) #Assign KO IDs to a vector
Counts <- All_patient_functions[,-1] #Remove IDs from the count table
```

```
rownames(All_patient_functions) <- All_patient_functions[,1]

#Random forest indicates these are most important patient communities driving
clustering:
#ALI: 25, 219, 79
#No ALI: 63, 246, 255, 124
#Center log ratio transformation
All.patient.funct.clr <- clr(All_patient_functions)

#Remove rows which total zero
bella <- apply(All.patient.funct.clr,1,sum)
temp <- which(!(bella==0))
All.patient.funct2 <- All.patient.funct.clr[temp,] #Removes columns which sum to zero
(ALI)
All.patient.funct2 <- as.data.frame(All.patient.funct2)
functs <- rownames(All.patient.funct2)

#Patient functions ranked as most important by random forest analysis in determining
SparCC clustering
ALI.rf.patients <- cbind(All.patient.funct2$X25, All.patient.funct2$X79)
rownames(ALI.rf.patients) <- functs
colnames(ALI.rf.patients) <- c("X25", "X79")
```

```
No.ALI.rf.patients <- cbind(All.patient.funct2$X63, All.patient.funct2$X246,
All.patient.funct2$X255, All.patient.funct2$X124)
rownames(No.ALI.rf.patients) <- functs
colnames(No.ALI.rf.patients) <- c("X63", "X246", "X255", "X124")
ALI.rf.patients.summary <- as.matrix(apply(ALI.rf.patients, 1, summary))
No.ALI.rf.patients.summary <- as.matrix(apply(No.ALI.rf.patients, 1, summary))
ALI.rf.mean <- apply(ALI.rf.patients,1,mean)
No.ALI.rf.mean <- apply(No.ALI.rf.patients,1,mean)
difference <- ALI.rf.mean-No.ALI.rf.mean

#Summary stats prior to scaling the data
s_prime <- summary(All_patient_functions)
max_per_ko <- as.matrix(apply(All_patient_functions[,-1], 1, max))
max_per_ko_id <- cbind(KOs, max_per_ko)
colnames(max_per_ko_id) <- c("KOs", "Max")
mean_per_ko <- as.matrix(apply(All_patient_functions[,-1], 1, mean))

max_per_patient_KO <- as.matrix(apply(All_patient_functions, 2, which.max))#This
gives the row index number - how to pull out actual value?
for(i in 1:52){
  KO_list[i] <- rownames(All_patient_functions[max_per_patient_KO[i],]) #Pulls out
row index for max KO, matches to KO ID and adds to list
}
```

```
max_per_patient_KO_ID <- cbind(max_per_patient_KO, KO_list) #Puts KO IDs
together with max KOs per patient
write.table(max_per_patient_KO_ID,
file="/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_An
alysis/72hrs/PICRUSt/max_per_patient_KO_ID.txt", sep="\t")

#Alpha diversity stats
alpha <-
read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Workin
g_Analysis/Community_Analysis/Alpha_diversity/alpha_div_june_13.txt", sep="\t",
header=TRUE, stringsAsFactors = FALSE)

#For ALI
boxplot(alpha$Chao1.Mean~alpha$ALI, main="Chao1 Diversity", ylab="Chao Index")
wilcox.test(Chao1.Mean~ALI, data=alpha, p.adj="bonferroni") #Ties - some values are
the same; can't compute p
kruskal.test(Chao1.Mean~ALI, data=alpha, p.adj="bonferroni")
t.test(Chao1.Mean~ALI, data=alpha, p.adj="bonferroni") #No errors here

#For Prevotella melaninogenica
boxplot(alpha$Chao1.Mean~alpha$Prevotella, main="Chao1 Diversity", ylab="Chao
Index")
```

```
wilcox.test(Chao1.Mean~Prevotella, data=alpha, p.adj="bonferroni") #Ties - some values are the same; can't compute p
t.test(Chao1.Mean~Prevotella, data=alpha, p.adj="bonferroni")

#Prev and ALI boxplot
boxplot(alpha$Chao1.Mean~alpha$ALI + alpha$Prevotella, main="Chao1 Diversity", ylab="Chao Index", xlab="Prevotella Detected", names=c("No", "No", "Yes", "Yes"), col=c("blue", "red", "blue", "red"))

#Gemellaceae
boxplot(alpha$Chao1.Mean~alpha$Gemellaceae, main="Chao1 Diversity", ylab="Chao Index", names=c("None", "Gemellaceae Present"))
wilcox.test(Chao1.Mean~Gemellaceae, data=alpha, p.adj="bonferroni") #Ties - some values are the same; can't compute p
t.test(Chao1.Mean~Gemellaceae, data=alpha, p.adj="bonferroni")

#Gemellaceae and ALI boxplot
boxplot(alpha$Chao1.Mean~alpha$ALI + alpha$Gemellaceae, main="Chao1 Diversity", ylab="Chao Index", xlab="Gemellaceae Detected", names=c("No", "No", "Yes", "Yes"), col=c("blue", "red", "blue", "red"))

#Enterobacteriaceae
```

```
boxplot(alpha$Chao1.Mean~alpha$Enterobacteriaceae, main="Chao1 Diversity",
ylab="Chao Index", names=c("None", "Enterobacteriaceae Present"))
wilcox.test(Chao1.Mean~Enterobacteriaceae, data=alpha, p.adj="bonferroni") #Ties -
some values are the same; can't compute p
t.test(Chao1.Mean~Enterobacteriaceae, data=alpha, p.adj="bonferroni")

#Enterobacteriaceae and ALI boxplot
boxplot(alpha$Chao1.Mean~alpha$ALI + alpha$Enterobacteriaceae, main="Chao1
Diversity", ylab="Chao Index", xlab="Enterobacteriaceae Detected", names=c("No",
"No", "Yes", "Yes"), col=c("blue", "red", "blue", "red"))

#Staphylococcus
boxplot(alpha$Chao1.Mean~alpha$Staphylococcus, main="Chao1 Diversity",
ylab="Chao Index", names=c("None", "Staphylococcus Present"))
wilcox.test(Chao1.Mean~Staphylococcus, data=alpha, p.adj="bonferroni") #Ties - some
values are the same; can't compute p
t.test(Chao1.Mean~Staphylococcus, data=alpha, p.adj="bonferroni")

#Staphylococcus and ALI boxplot
boxplot(alpha$Chao1.Mean~alpha$ALI + alpha$Staphylococcus, main="Chao1
Diversity", ylab="Chao Index", xlab="Staphylococcus Detected", names=c("No", "No",
"Yes", "Yes"), col=c("blue", "red", "blue", "red"))
```

```
#Dana Walsh
#August 10 2016
#Analysis for paper #2 (Predicted functions of bacterial communities)
#Heatmaps for predicted OTU functions

#Import the files
ALI.comm <-
read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Workin
g_Analysis/Community_Analysis/Heatmaps/ALI.comm.funct2.clr.txt", sep="\t",
header=TRUE)

No.ALI.comm <-
read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Workin
g_Analysis/Community_Analysis/Heatmaps/No.ALI.comm.funct2.clr.txt", sep="\t",
header=TRUE)

#Remove the taxonomy, OTU ID numbers, and community assignments
ALI.comm.labels <- ALI.comm[,1:3]
ALI.comm.funct <- ALI.comm[,-(1:3)]
No.ALI.comm.labels <- No.ALI.comm[,1:3]
No.ALI.comm.funct <- No.ALI.comm[,-(1:3)]

#Convert to a matrix
```

```
ALI.comm.funct <- as.matrix(ALI.comm.funct)
No.ALI.comm.funct <- as.matrix(No.ALI.comm.funct)

#Assign OTU IDs as row names
ALI.OTUs <- ALI.comm.labels[,2]
No.ALI.OTUs <- No.ALI.comm.labels[,2]
rownames(ALI.comm.funct) <- ALI.OTUs
rownames(No.ALI.comm.funct) <- No.ALI.OTUs

#The heatmap below is the same for both ALI and No ALI
distance <- dist(ALI.comm.funct, method="manhattan")
cluster <- hclust(distance, method="ward.D2")
png(filename="/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Wo
rking_Analysis/Community_Analysis/Heatmaps/ali.comm.heatmap.png", width=1200,
height=800)
heatmap.2(ALI.comm.funct, Rowv=as.dendrogram(cluster), Colv=TRUE,
scale="column", trace="none", col=redgreen, xlab="Functions", ylab="OTU IDs",
margins=c(10,15))
dev.off()

#Make per community heatmaps
#Replace rownames with community assignments
ALI.comms <- ALI.comm.labels[,3]
```

```
No.ALI.comms <- No.ALI.comm.labels[,3]

rownames(ALI.comm.funct) <- ALI.comms

rownames(No.ALI.comm.funct) <- No.ALI.comms


#Add OTU IDs back in

ALI.comm.funct <- cbind(ALI.comm.labels[,2], ALI.comm.funct)

No.ALI.comm.funct <- cbind(No.ALI.comm.labels[,2], No.ALI.comm.funct)


#Select each community

ALI.comm.1 <- ALI.comm.funct[which(rownames(ALI.comm.funct)==1),]

ALI.comm.2 <- ALI.comm.funct[which(rownames(ALI.comm.funct)==2),]

ALI.comm.3 <- ALI.comm.funct[which(rownames(ALI.comm.funct)==3),]

ALI.comm.4 <- ALI.comm.funct[which(rownames(ALI.comm.funct)==4),]


No.ALI.comm.1 <- No.ALI.comm.funct[which(rownames(No.ALI.comm.funct)==1),]

No.ALI.comm.2 <- No.ALI.comm.funct[which(rownames(No.ALI.comm.funct)==2),]

No.ALI.comm.3 <- No.ALI.comm.funct[which(rownames(No.ALI.comm.funct)==3),]

No.ALI.comm.4 <- No.ALI.comm.funct[which(rownames(No.ALI.comm.funct)==4),]


#Remove the columns with zero sum

sum.ali.1 <- apply(ALI.comm.1,2,sum)

sum.ali.2 <- apply(ALI.comm.2,2,sum)

sum.ali.3 <- apply(ALI.comm.3,2,sum)
```

```
sum.ali.4 <- apply(ALI.comm.4,2,sum)

no.sum.ali.1 <- apply(No.ALI.comm.1,2,sum)
no.sum.ali.2 <- apply(No.ALI.comm.2,2,sum)
no.sum.ali.3 <- apply(No.ALI.comm.3,2,sum)
no.sum.ali.4 <- apply(No.ALI.comm.4,2,sum)

ali.keep.1 <- which(!(sum.ali.1==0))
ali.keep.2 <- which(!(sum.ali.2==0))
ali.keep.3 <- which(!(sum.ali.3==0))
ali.keep.4 <- which(!(sum.ali.4==0))

no.ali.keep.1 <- which(!(no.sum.ali.1==0))
no.ali.keep.2 <- which(!(no.sum.ali.2==0))
no.ali.keep.3 <- which(!(no.sum.ali.3==0))
no.ali.keep.4 <- which(!(no.sum.ali.4==0))

library("plotrix")

gap.boxplot(sum.ali.1, sum.ali.2, sum.ali.3, sum.ali.4, no.sum.ali.1, no.sum.ali.2,
no.sum.ali.3, no.sum.ali.4, gap=list(top=c(1e+1, 5e+5), bottom=c(0,1e+1)), las=2,
names=c("ALI 1", "ALI 2", "ALI 3", "ALI 4", "No ALI 1", "No ALI 2", "No ALI 3",
"No ALI 4"))
```

```
boxplot(sum.ali.1, sum.ali.2, sum.ali.3, sum.ali.4, no.sum.ali.1, no.sum.ali.2,
no.sum.ali.3, no.sum.ali.4, las=2, names=c("ALI 1", "ALI 2", "ALI 3", "ALI 4", "No ALI
1", "No ALI 2", "No ALI 3", "No ALI 4"))

ALI.comm.1 <- ALI.comm.1[,ali.keep.1]
ALI.comm.2 <- ALI.comm.2[,ali.keep.2]
ALI.comm.3 <- ALI.comm.3[,ali.keep.3]
ALI.comm.4 <- ALI.comm.4[,ali.keep.4]

No.ALI.comm.1 <- No.ALI.comm.1[,no.ali.keep.1]
No.ALI.comm.2 <- No.ALI.comm.2[,no.ali.keep.2]
No.ALI.comm.3 <- No.ALI.comm.3[,no.ali.keep.3]
No.ALI.comm.4 <- No.ALI.comm.4[,no.ali.keep.4]

#Make OTU IDs rownames
ALI.comm.1.otus <- ALI.comm.1[,1]
rownames(ALI.comm.1) <- ALI.comm.1.otus
ALI.comm.1 <- ALI.comm.1[,-1]

ALI.comm.2.otus <- ALI.comm.2[,1]
rownames(ALI.comm.2) <- ALI.comm.2.otus
ALI.comm.2 <- ALI.comm.2[,-1]
```

```
ALI.comm.3.otus <- ALI.comm.3[,1]
rownames(ALI.comm.3) <- ALI.comm.3.otus
ALI.comm.3 <- ALI.comm.3[,-1]

ALI.comm.4.otus <- ALI.comm.4[,1]
rownames(ALI.comm.4) <- ALI.comm.4.otus
ALI.comm.4 <- ALI.comm.4[,-1]

No.ALI.comm.1.otus <- No.ALI.comm.1[,1]
rownames(No.ALI.comm.1) <- No.ALI.comm.1.otus
No.ALI.comm.1 <- No.ALI.comm.1[,-1]

No.ALI.comm.2.otus <- No.ALI.comm.2[,1]
rownames(No.ALI.comm.2) <- No.ALI.comm.2.otus
No.ALI.comm.2 <- No.ALI.comm.2[,-1]

No.ALI.comm.3.otus <- No.ALI.comm.3[,1]
rownames(No.ALI.comm.3) <- No.ALI.comm.3.otus
No.ALI.comm.3 <- No.ALI.comm.3[,-1]

No.ALI.comm.4.otus <- No.ALI.comm.4[,1]
rownames(No.ALI.comm.4) <- No.ALI.comm.4.otus
No.ALI.comm.4 <- No.ALI.comm.4[,-1]
```

```
#Heatmaps
dist.ali.1 <- dist(ALI.comm.1, method="manhattan")
clust.ali.1 <- hclust(dist.ali.1, method="ward.D2")
png(filename="/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_Analysis/Community_Analysis/Heatmaps/ali.comm.1.heatmap.png", width=1200, height=800)
heatmap.2(ALI.comm.1, Rowv=as.dendrogram(clust.ali.1), Colv=TRUE,
scale="column", trace="none", col=redgreen, xlab="Predicted Functions", ylab="OTU ID", margins=c(10,15))
dev.off()

dist.ali.2 <- dist(ALI.comm.2, method="manhattan")
clust.ali.2 <- hclust(dist.ali.2, method="ward.D2")
png(filename="/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_Analysis/Community_Analysis/Heatmaps/ali.comm.2.heatmap.png", width=1200, height=800)
heatmap.2(ALI.comm.2, Rowv=as.dendrogram(clust.ali.2), Colv=TRUE,
scale="column", trace="none", col=redgreen, xlab="Predicted Functions", ylab="OTU ID", margins=c(10,15))
dev.off()

dist.ali.3 <- dist(ALI.comm.3, method="manhattan")
clust.ali.3 <- hclust(dist.ali.3, method="ward.D2")
```

```
png(filename="/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_Analysis/Community_Analysis/Heatmaps/ali.comm.3.heatmap.png", width=1200, height=800)
heatmap.2(ALI.comm.3, Rowv=as.dendrogram(clust.ali.3), Colv=TRUE, scale="column", trace="none", col=redgreen, xlab="Predicted Functions", ylab="OTU ID", margins=c(10,15))
dev.off()

dist.ali.4 <- dist(ALI.comm.4, method="manhattan")
clust.ali.4 <- hclust(dist.ali.4, method="ward.D2")
png(filename="/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_Analysis/Community_Analysis/Heatmaps/ali.comm.4.heatmap.png", width=1200, height=800)
heatmap.2(ALI.comm.4, Rowv=as.dendrogram(clust.ali.4), Colv=TRUE, scale="column", trace="none", col=redgreen, xlab="Predicted Functions", ylab="OTU ID", margins=c(10,15))
dev.off()

dist.no.ali.1 <- dist(No.ALI.comm.1, method="manhattan")
clust.no.ali.1 <- hclust(dist.no.ali.1, method="ward.D2")
png(filename="/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_Analysis/Community_Analysis/Heatmaps/no.ali.comm.1.heatmap.png", width=1200, height=800)
```

```
heatmap.2(No.ALI.comm.1, Rowv=as.dendrogram(clust.no.ali.1), Colv=TRUE,
scale="column", trace="none", col=redgreen, xlab="Predicted Functions", ylab="OTU
ID", margins=c(10,15))
dev.off()

dist.no.ali.2 <- dist(No.ALI.comm.2, method="manhattan")
clust.no.ali.2 <- hclust(dist.no.ali.2, method="ward.D2")
png(filename="/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Wo
rking_Analysis/Community_Analysis/Heatmaps/no.ali.comm.2.heatmap.png",
width=1200, height=800)
heatmap.2(No.ALI.comm.2, Rowv=as.dendrogram(clust.no.ali.2), Colv=TRUE,
scale="column", trace="none", col=redgreen, xlab="Predicted Functions", ylab="OTU
ID", margins=c(10,15))
dev.off()

dist.no.ali.3 <- dist(No.ALI.comm.3, method="manhattan")
clust.no.ali.3 <- hclust(dist.no.ali.3, method="ward.D2")
png(filename="/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Wo
rking_Analysis/Community_Analysis/Heatmaps/no.ali.comm.3.heatmap.png",
width=1200, height=800)
heatmap.2(No.ALI.comm.3, Rowv=as.dendrogram(clust.no.ali.3), Colv=TRUE,
scale="column", trace="none", col=redgreen, xlab="Predicted Functions", ylab="OTU
ID", margins=c(10,15))
```

```
dev.off()

dist.no.ali.4 <- dist(No.ALI.comm.4, method="manhattan")
clust.no.ali.4 <- hclust(dist.no.ali.4, method="ward.D2")
png(filename="/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_Analysis/Community_Analysis/Heatmaps/no.ali.comm.4.heatmap.png",
width=1200, height=800)
heatmap.2(No.ALI.comm.4, Rowv=as.dendrogram(clust.no.ali.4), Colv=TRUE,
scale="column", trace="none", col=redgreen, xlab="Predicted Functions", ylab="OTU
ID", margins=c(10,15))
dev.off()

#Heatmaps for Abundance Data with Community Assignments
library("compositions")
ALI.comm.abundances <-
read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_Analysis/Community_Analysis/Community_Assignments/ALI.comm.abundances.txt",
sep="\t", header=TRUE)
ALI.comm.orig <- ALI.comm.abundances
#Draw heatmap without re-ordering by the dendogram
color.map <- function(community){
        if(community=="1") "red"
        else if(community=="2") "green"
```

```
        else if(community=="3") "blue"
        else if(community=="4") "purple"
}
sidebarcolors <- unlist(lapply(ALI.comm.abundances$community, color.map))
ALI.otus <- ALI.comm.abundances[,1]
ALI.comms <- ALI.comm.abundances[,2]
ALI.taxa <- ALI.comm.abundances[,3]
ALI.comm.abundances <- ALI.comm.abundances[,-(1:3)]
rownames(ALI.comm.abundances) <- ALI.otus
ALI.comm.abundances <- clr(ALI.comm.abundances)

#Heatmap ordered by community assignment
png(filename="/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Wo
rking_Analysis/Community_Analysis/Heatmaps/ali.comm.abundances.2.heatmap.png",
width=1200, height=800)
heatmap.2(as.matrix(ALI.comm.abundances), Rowv=NA, Colv=NA, scale="none",
trace="none", col=redgreen, xlab="Patient ID", ylab="OTUs", margins=c(10,15),
RowSideColors=sidebarcolors)
dev.off()

#Hierarchical clustering-based heatmap
dist.ali.abund <- dist(ALI.comm.abundances, method="manhattan")
clust.ali.abund <- hclust(dist.ali.abund, method="ward.D2")
```

```
png(filename="/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_Analysis/Community_Analysis/Heatmaps/ali.comm.abundances.heatmap.png", width=1200, height=800)
heatmap.2(ALI.comm.abundances, Rowv=as.dendrogram(clust.ali.abund), Colv=TRUE, scale="column", trace="none", col=redgreen, xlab="Predicted Functions", ylab="OTU ID", margins=c(10,15))
dev.off()



No.ALI.comm.abundances <-
read.table("/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_Analysis/Community_Analysis/Community_Assignments/No.ALI.comm.abundances.txt", sep="\t", header=TRUE)
No.ALI.comm.orig <- No.ALI.comm.abundances
sidebarcolors.2 <- unlist(lapply(No.ALI.comm.abundances$community, color.map))
#Draw heatmap without re-ordering by the dendogram
No.ALI.otus <- No.ALI.comm.abundances[,1]
No.ALI.comms <- No.ALI.comm.abundances[,2]
No.ALI.taxa <- No.ALI.comm.abundances[,3]
No.ALI.comm.abundances <- No.ALI.comm.abundances[,-(1:3)]
rownames(No.ALI.comm.abundances) <- No.ALI.otus
No.ALI.comm.abundances <- clr(No.ALI.comm.abundances)
```

```
#Heatmap ordered by community assignment
png(filename="/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_Analysis/Community_Analysis/Heatmaps/no.ali.comm.abundances.2.heatmap.png", width=1200, height=800)
heatmap.2(as.matrix(No.ALI.comm.abundances), Rowv=NA, Colv=NA, scale="none", trace="none", col=redgreen, xlab="Patient ID", ylab="OTUs", margins=c(10,15), RowSideColors=sidebarcolors.2)
dev.off()

#Hierarchical clustering-based heatmap
dist.no.ali.abund <- dist(No.ALI.comm.abundances, method="manhattan")
clust.no.ali.abund <- hclust(dist.no.ali.abund, method="ward.D2")
png(filename="/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Working_Analysis/Community_Analysis/Heatmaps/no.ali.comm.abundances.heatmap.png", width=1200, height=800)
heatmap.2(No.ALI.comm.abundances, Rowv=as.dendrogram(clust.no.ali.abund), Colv=TRUE, scale="column", trace="none", col=redgreen, xlab="Predicted Functions", ylab="OTU ID", margins=c(10,15))
dev.off()

#Random forest for taxa that predict community assignments among ALI and None
rownames(ALI.comm.orig) <- ALI.otus
ALI.comm.only <- ALI.comm.orig[,-1]
```

```
ALI.comm.only <- ALI.comm.only[,-2]
community <- as.character(ALI.comm.only$community)
ALI.comm.only <- ALI.comm.only[,-1]
ALI.comm.only <- cbind(community, ALI.comm.only)

rownames(No.ALI.comm.orig) <- No.ALI.otus
No.ALI.comm.only <- No.ALI.comm.orig[,-1]
No.ALI.comm.only <- No.ALI.comm.only[,-2]
community.2 <- as.character(No.ALI.comm.only$community)
No.ALI.comm.only <- No.ALI.comm.only[,-1]
No.ALI.comm.only <- cbind(community.2, No.ALI.comm.only)

library("randomForest")
library("e1071")

set.seed(18)
choo_train <- ALI.comm.only[sample(1:nrow(ALI.comm.only),
nrow(ALI.comm.only)*0.5, replace=FALSE),]
tune <- tune.randomForest(community~., data=choo_train, mtry=c(2.8, 5.6, 11.1),
ntree=c(250,500), na.rm=TRUE)
x<-summary(tune)

#Entire data set
```

```
rf.ALI.comm <- randomForest(community~., data=ALI.comm.only, importance=TRUE,
na.action=na.omit, mtry=as.numeric(x$best.parameters[1]),
ntree=as.numeric(x$best.parameters[2]))
png(filename="/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Wo
rking_Analysis/Community_Analysis/Random_Forests/ALI.comm.abundances.heatmap.
png", width=800, height=1200, res=300)
varImpPlot(rf.ALI.comm, cex=.7)
dev.off()

set.seed(23)
choo_train.2 <- No.ALI.comm.only[sample(1:nrow(No.ALI.comm.only),
nrow(No.ALI.comm.only)*0.5, replace=FALSE),]
tune <- tune.randomForest(community.2~., data=choo_train.2, mtry=c(2.8, 5.6, 11.1),
ntree=c(250,500), na.rm=TRUE)
x.2<-summary(tune)

#Entire data set
rf.No.ALI.comm <- randomForest(community.2~., data=No.ALI.comm.only,
importance=TRUE, na.action=na.omit, mtry=as.numeric(x.2$best.parameters[1]),
ntree=as.numeric(x.2$best.parameters[2]))
png(filename="/Users/walshdm/Documents/Burn_Study/Sequencing_Data_Analysis/Wo
rking_Analysis/Community_Analysis/Random_Forests/No.ALI.comm.abundances.heatm
ap.png", width=800, height=1200, res=300)
```

```
varImpPlot(rf.No.ALI.comm, cex=.7)

dev.off()
```

REFERENCES

[1] X.C. Morgan, C. Huttenhower, Chapter 12: Human Microbiome Analysis, PLoS Comput. Biol. 8 (2012). doi:10.1371/journal.pcbi.1002808.

[2] J.F. Petrosino, S. Highlander, R.A. Luna, R. a Gibbs, J. Versalovic, Metagenomic pyrosequencing and microbial identification., Clin. Chem. 55 (2009) 856–66. doi:10.1373/clinchem.2008.107565.

[3] G.M. Weinstock, Genomic approaches to studying the human microbiota., Nature. 489 (2012) 250–6. doi:10.1038/nature11553.

[4] J.R. Marchesi, J. Ravel, The vocabulary of microbiome research: a proposal, Microbiome. 3 (2015) 31. doi:10.1186/s40168-015-0094-5.

[5] R. Sender, S. Fuchs, R. Milo, Revised estimates for the number of human and bacteria cells in the body, bioRxiv. Jan (2016) 1–21. doi:http://dx.doi.org/10.1101/036103.

[6] S.R. Bordenstein, K.R. Theis, Host Biology in Light of the Microbiome: Ten Principles of Holobionts and Hologenomes, PLOS Biol. 13 (2015) e1002226. doi:10.1371/journal.pbio.1002226.

[7] P.J. Turnbaugh, R.E. Ley, M. Hamady, C.M. Fraser-Liggett, R. Knight, J.I. Gordon, The human microbiome project., Nature. 449 (2007) 804–10. doi:10.1038/nature06244.

[8] The Human Microbiome Project Consortium, Structure, function and diversity of the healthy human microbiome., Nature. 486 (2012) 207–14. doi:10.1038/nature11234.

[9] L. V Hooper, D.R. Littman, A.J. Macpherson, Interactions between the microbiota and the immune system., Science. 336 (2012) 1268–73. doi:10.1126/science.1223490.

[10] Y. Lai, A. Di Nardo, T. Nakatsuji, A. Leichtle, Y. Yang, A.L. Cogen, Z.-R. Wu, L. V Hooper, R.R. Schmidt, S. von Aulock, K. a Radek, C.-M. Huang, A.F. Ryan, R.L. Gallo, Commensal bacteria regulate Toll-like receptor 3-dependent inflammation after skin injury., Nat. Med. 15 (2009) 1377–82. doi:10.1038/nm.2062.

[11] W.S.F. Chung, A.W. Walker, P. Louis, J. Parkhill, J. Vermeiren, D. Bosscher, S.H. Duncan, H.J. Flint, Modulation of the human gut microbiota by dietary fibres occurs at the species level, BMC Biol. 14 (2016) 3. doi:10.1186/s12915-015-0224-

3.

[12] F. Bäckhed, C.M.M. Fraser, Y. Ringel, M.E.E. Sanders, R.B.B. Sartor, P.M.M. Sherman, J. Versalovic, V. Young, B.B.B. Finlay, Defining a Healthy Human Gut Microbiome: Current Concepts, Future Directions, and Clinical Applications, Cell Host Microbe. 12 (2012) 611–622. doi:10.1016/j.chom.2012.10.012.

[13] M.A. Fischbach, J.A. Segre, Signaling in Host-Associated Microbial Communities, Cell. 164 (2016) 1288–1300. doi:10.1016/j.cell.2016.02.037.

[14] R.R. Dietert, The Microbiome in early life: Self-completion and microbiota protection as health priorities, Birth Defects Res. Part B - Dev. Reprod. Toxicol. 101 (2014) 333–340. doi:10.1002/bdrb.21116.

[15] M.R.S. Walther-Antonio, P. Jeraldo, M.E. Berg Miller, C.J. Yeoman, K.E. Nelson, B.A. Wilson, B.A. White, N. Chia, D.J. Creedon, Pregnancy's stronghold on the vaginal microbiome, PLoS One. 9 (2014) 1–10. doi:10.1371/journal.pone.0098514.

[16] N. a Abreu, N. a Nagalingam, Y. Song, F.C. Roediger, S.D. Pletcher, A.N. Goldberg, S. V Lynch, Sinus microbiome diversity depletion and Corynebacterium tuberculostearicum enrichment mediates rhinosinusitis., Sci. Transl. Med. 4 (2012) 151ra124. doi:10.1126/scitranslmed.3003783.

[17] G. Hajishengallis, S. Liang, M.A. Payne, A. Hashim, R. Jotwani, M.A. Eskan, M.L. McIntosh, A. Alsam, K.L. Kirkwood, J.D. Lambris, R.P. Darveau, M.A. Curtis, Low-abundance biofilm species orchestrates inflammatory periodontal disease through the commensal microbiota and complement., Cell Host Microbe. 10 (2011) 497–506. doi:10.1016/j.chom.2011.10.006.

[18] F.S. Collins, E.D. Green, A.E. Guttmacher, M. S., A vision for the future of genomics research, Nature. 422 (2003) 15–17.

[19] M.S. Rappé, S.J. Giovannoni, The uncultured microbial majority, Annu. Rev. Microbiol. 57 (2003) 369–394. doi:10.1146/annurev.micro.57.030502.090759.

[20] The Human Microbiome Project Consortium, A framework for human microbiome research., Nature. 486 (2012) 215–21. doi:10.1038/nature11209.

[21] A. Shade, J. Handelsman, Beyond the Venn diagram: the hunt for a core microbiome, Environ. Microbiol. 14 (2012) 4–12. doi:10.1111/j.1462-2920.2011.02585.x.

[22] M. Hamady, R. Knight, Microbial community profiling for human microbiome projects: Tools, techniques, and challenges., Genome Res. 19 (2009) 1141–52. doi:10.1101/gr.085464.108.

[23] S.M. Huse, Y. Ye, Y. Zhou, A. a Fodor, A Core Human Microbiome as Viewed through 16S rRNA Sequence Clusters., PLoS One. 7 (2012) e34242. doi:10.1371/journal.pone.0034242.

[24] H. Zheng, L. Xu, Z. Wang, L. Li, J. Zhang, Q. Zhang, T. Chen, J. Lin, F. Chen, Subgingival microbiome in patients with healthy and ailing dental implants, Sci. Rep. 5 (2015) 10948. doi:10.1038/srep10948.

[25] D. Bogaert, B. Keijser, S. Huse, J. Rossen, R. Veenhoven, E. van Gils, J. Bruin, R. Montijn, M. Bonten, E. Sanders, Variability and diversity of nasopharyngeal microbiota in children: a metagenomic analysis., PLoS One. 6 (2011) e17035. doi:10.1371/journal.pone.0017035.

[26] E.R. Morton, J. Lynch, A. Froment, S. Lafosse, E. Heyer, M. Przeworski, R. Blekhman, L. Segurel, Variation in rural African gut microbiomes is strongly shaped by parasitism and diet, PLOS Genet. 11 (2015) e1005658. doi:10.1101/016949.

[27] T. Yatsunenko, F.E. Rey, M.J. Manary, I. Trehan, M.G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R.N. Baldassano, A.P. Anokhin, A.C. Heath, B. Warner, J. Reeder, J. Kuczynski, J.G. Caporaso, C.A. Lozupone, C. Lauber, J.C. Clemente, D. Knights, R. Knight, J.I. Gordon, Human gut microbiome viewed across age and geography, Nature. 486 (2012) 222–227. doi:10.1038/nature11053.

[28] B.D. Muegge, J. Kuczynski, D. Knights, J.C. Clemente, A. González, L. Fontana, B. Henrissat, R. Knight, J.I. Gordon, Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans., Science. 332 (2011) 970–4. doi:10.1126/science.1198719.

[29] C.F. Maurice, H.J. Haiser, P.J. Turnbaugh, Xenobiotics shape the physiology and gene expression of the active human gut microbiome., Cell. 152 (2013) 39–50. doi:10.1016/j.cell.2012.10.052.

[30] J. Wu, B.A. Peters, C. Dominianni, Y. Zhang, Z. Pei, L. Yang, Y. Ma, M.P. Purdue, E.J. Jacobs, S.M. Gapstur, H. Li, A. V Alekseyenko, R.B. Hayes, J. Ahn, Cigarette smoking and the oral microbiome in a large study of American adults, Isme J. (2016) 1–12. doi:10.1038/ismej.2016.37.

[31] N. Goldstein-Daruech, E.K. Cope, K.-Q. Zhao, K. Vukovic, J.M. Kofonow, L. Doghramji, B. González, A.G. Chiu, D.W. Kennedy, J.N. Palmer, J.G. Leid, J.L. Kreindler, N. a Cohen, Tobacco smoke mediated induction of sinonasal microbial biofilms., PLoS One. 6 (2011) e15700. doi:10.1371/journal.pone.0015700.

[32] J. Hu, V. Raikhel, K. Gopalakrishnan, H. Fernandez-Hernandez, L. Lambertini, F. Manservisi, L. Falcioni, L. Bua, F. Belpoggi, S. L.Teitelbaum, J. Chen, Effect of postnatal low-dose exposure to environmental chemicals on the gut microbiome in

a rodent model, Microbiome. 4 (2016) 26. doi:10.1186/s40168-016-0173-2.

[33] J.K. Goodrich, J.L. Waters, A.C. Poole, J.L. Sutter, O. Koren, R. Blekhman, M. Beaumont, W. Van Treuren, R. Knight, J.T. Bell, T.D. Spector, A.G. Clark, R.E. Ley, Human Genetics Shape the Gut Microbiome, Cell. 159 (2014) 789–799. doi:10.1016/j.cell.2014.09.053.

[34] S. Liu, A. Pires, R.M. Rezende, L.E. Comstock, R. Gandhi, H.L. Weiner, S. Liu, A. Pires, R.M. Rezende, R. Cialic, Z. Wei, L. Bry, L.E. Comstock, The Host Shapes the Gut Microbiota via Fecal MicroRNA, Cell Host Microbe. 19 (2016) 32–43. doi:10.1016/j.chom.2015.12.005.

[35] B.J. Marsland, K. Yadava, L.P. Nicod, The airway microbiome and disease., Chest. 144 (2013) 632–7. doi:10.1378/chest.12-2854.

[36] A.S. Neish, Mucosal immunity and the microbiome, Ann. Am. Thorac. Soc. 11 (2014). doi:10.1513/AnnalsATS.201306-161MG.

[37] K. Honda, D.R. Littman, The microbiome in infectious disease and inflammation., Annu. Rev. Immunol. 30 (2012) 759–95. doi:10.1146/annurev-immunol-020711-074937.

[38] A.M. Schubert, M. a M. Rogers, C. Ring, J. Mogle, J.P. Petrosino, V.B. Young, D.M. Aronoff, P.D. Schloss, Microbiome Data Distinguish Patients with Clostridium difficile Infection and Non- C . difficile -Associated Diarrhea from Healthy, MBio. 5 (2014) 1–9. doi:10.1128/mBio.01021-14.Editor.

[39] M.C. Abt, D. Artis, The dynamic influence of commensal bacteria on the immune response to pathogens., Curr. Opin. Microbiol. 16 (2013) 4–9. doi:10.1016/j.mib.2012.12.002.

[40] A.L. Prince, J. Ma, P.S. Kannan, M. Alvarez, T. Gisslen, R.A. Harris, E.L. Sweeney, C.L. Knox, D.S. Lambers, A.H. Jobe, C.A. Chougnet, S.G. Kallapur, K.M. Aagaard, The placental membrane microbiome is altered among subjects with spontaneous preterm birth with and without chorioamnionitis, Am. J. Obstet. Gynecol. 214 (2016) 627e1–627e16. doi:10.1016/j.ajog.2016.01.193.

[41] C. Fox, K. Eichelberger, Maternal microbiome and pregnancy outcomes, Fertil. Steril. 104 (2015) 1358–1363. doi:10.1016/j.fertnstert.2015.09.037.

[42] B.J. Marsland, O. Salami, Microbiome influences on allergy in mice and humans, Curr. Opin. Immunol. 36 (2015) 94–100. doi:10.1016/j.coi.2015.07.005.

[43] J. Lampi, D. Canoy, D. Jarvis, A.L. Hartikainen, L. Keski-Nisula, M.R. J??rvelin, J. Pekkanen, Farming environment and prevalence of atopy at age 31: Prospective birth cohort study in Finland, Clin. Exp. Allergy. 41 (2011) 987–993.

doi:10.1111/j.1365-2222.2011.03777.x.

[44] D.P. Strachan, Family size, infection and atopy: the first decade of the'hygiene hypothesis', Thorax. 55 (2000) S2.

[45] M.C. Abt, L.C. Osborne, L. a Monticelli, T. a Doering, T. Alenghat, G.F. Sonnenberg, M. a Paley, M. Antenus, K.L. Williams, J. Erikson, E.J. Wherry, D. Artis, Commensal bacteria calibrate the activation threshold of innate antiviral immunity., Immunity. 37 (2012) 158–70. doi:10.1016/j.immuni.2012.04.011.

[46] M. Costalonga, P.P. Cleary, L. a Fischer, Z. Zhao, Intranasal bacteria induce Th1 but not Treg or Th2., Mucosal Immunol. 2 (2009) 85–95. doi:10.1038/mi.2008.67.

[47] C. Manichanh, N. Borruel, F. Casellas, F. Guarner, The gut microbiota in IBD., Nat. Rev. Gastroenterol. Hepatol. 9 (2012) 599–608. doi:10.1038/nrgastro.2012.152.

[48] L. Mira-Pascual, R. Cabrera-Rubio, S. Ocon, P. Costales, A. Parra, A. Suarez, F. Moris, L. Rodrigo, A. Mira, M.C. Collado, Microbial mucosal colonic shifts associated with the development of colorectal cancer reveal the presence of different bacterial and archaeal biomarkers, J. Gastroenterol. 50 (2014) 167–179. doi:10.1007/s00535-014-0963-x.

[49] N. Larsen, F.K. Vogensen, F.W.J. van den Berg, D.S. Nielsen, A.S. Andreasen, B.K. Pedersen, W.A. Al-Soud, S.J. Sørensen, L.H. Hansen, M. Jakobsen, Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults., PLoS One. 5 (2010) e9085. doi:10.1371/journal.pone.0009085.

[50] B. Zhu, X. Wang, L. Li, Human gut microbiome: the second genome of human body., Protein Cell. 1 (2010) 718–25. doi:10.1007/s13238-010-0093-z.

[51] M. Zozaya, M.J. Ferris, J.D. Siren, R. Lillis, L. Myers, M.J. Nsuami, A.M. Eren, J. Brown, C.M. Taylor, D.H. Martin, Bacterial communities in penile skin, male urethra, and vaginas of heterosexual couples with and without bacterial vaginosis, Microbiome. 4 (2016) 16. doi:10.1186/s40168-016-0161-6.

[52] E.A. Baldwin, M. Walther-Antonio, A.M. MacLean, D.M. Gohl, K.B. Beckman, J. Chen, B. White, D.J. Creedon, N. Chia, Persistent microbial dysbiosis in preterm premature rupture of membranes from onset until delivery., PeerJ. 3 (2015) e1398. doi:10.7717/peerj.1398.

[53] Z.M. Earley, S. Akhtar, S.J. Green, A. Naqib, O. Khan, A.R. Cannon, A.M. Hammer, N.L. Morris, X. Li, J.M. Eberhardt, R.L. Gamelli, R.H. Kennedy, M.A. Choudhry, Burn injury alters the intestinal microbiome and increases gut permeability and bacterial translocation, PLoS One. 10 (2015) 1–16. doi:10.1371/journal.pone.0129996.

[54] R. Dheer, J. Patterson, M. Dudash, E.N. Stachler, K.J. Bibby, D.B. Stolz, S. Shiva, Z. Wang, S.L. Hazen, A. Barchowsky, J.F. Stolz, Arsenic induces structural and compositional colonic microbiome change and promotes host nitrogen and amino acid metabolism, Toxicol. Appl. Pharmacol. 289 (2015) 397–408. doi:10.1016/j.taap.2015.10.020.

[55] D.M. Comer, J.S. Elborn, M. Ennis, Inflammatory and cytotoxic effects of acrolein, nicotine, acetylaldehdye and cigarette smoke extract on human nasal epithelial cells, BMC Pulm. Med. 14 (2014).

[56] D.S. Olivera, S.E. Boggs, C. Beenhouwer, J. Aden, C. Knall, Cellular mechanisms of mainstream cigarette smoke-induced lung epithelial tight junction permeability changes in vitro., Inhal. Toxicol. 19 (2007) 13–22. doi:10.1080/08958370600985768.

[57] D.M. DeMarini, Genotoxicity of tobacco smoke and tobacco smoke condensate: a review., Mutat. Res. 567 (2004) 447–74. doi:10.1016/j.mrrev.2004.02.001.

[58] A. Morris, J.M. Beck, P.D. Schloss, T.B. Campbell, K. Crothers, J.L. Curtis, S.C. Flores, A.P. Fontenot, E. Ghedin, L. Huang, K. Jablonski, E. Kleerup, S. V. Lynch, E. Sodergren, H. Twigg, V.B. Young, C.M. Bassis, A. Venkataraman, T.M. Schmidt, G.M. Weinstock, Comparison of the Respiratory Microbiome in Healthy Nonsmokers and Smokers, Am. J. Respir. Crit. Care Med. 187 (2013) 1067–1075. doi:10.1164/rccm.201210-1913OC.

[59] E.S. Charlson, J. Chen, R. Custers-Allen, K. Bittinger, H. Li, R. Sinha, J. Hwang, F.D. Bushman, R.G. Collman, Disordered microbial communities in the upper respiratory tract of cigarette smokers., PLoS One. 5 (2010) e15216. doi:10.1371/journal.pone.0015216.

[60] A.D. Kostic, M.R. Howitt, W.S. Garrett, Exploring host – microbiota interactions in animal models and humans, Genes Dev. 27 (2013) 701–718. doi:10.1101/gad.212522.112.of.

[61] V.A. Miracle, The Belmont Report: The Triple Crown of Research Ethics, Dimens. Crit. Care Nurs. 35 (2016) 223–228. doi:10.1097/DCC.0000000000000186.

[62] K. Venema, P. van den Abbeele, Experimental models of the gut microbiome., Best Pract. Res. Clin. Gastroenterol. 27 (2013) 115–26. doi:10.1016/j.bpg.2013.03.002.

[63] M.K. Skinner, Environmental stress and epigenetic transgenerational inheritance., BMC Med. 12 (2014) 153. doi:10.1186/s12916-014-0153-y.

[64] S.F. Gilbert, A holobiont birth narrative: The epigenetic transmission of the human

microbiome, Front. Genet. 5 (2014) 1–7. doi:10.3389/fgene.2014.00282.

[65] M. a J. Hullar, B.C. Fu, Diet, the gut microbiome, and epigenetics., Cancer J. 20 (2014) 170–5. doi:10.1097/PPO.0000000000000053.

[66] J. V Fritz, M.S. Desai, P. Shah, J.G. Schneider, P. Wilmes, From meta-omics to causality: experimental models for human microbiome research., Microbiome. 1 (2013) 14. doi:10.1186/2049-2618-1-14.

[67] V.K. Ridaura, J.J. Faith, F.E. Rey, J. Cheng, A.E. Duncan, A.L. Kau, N.W. Griffin, V. Lombard, B. Henrissat, J.R. Bain, M.J. Muehlbauer, O. Ilkayeva, C.F. Semenkovich, K. Funai, D.K. Hayashi, B.J. Lyle, M.C. Martini, L.K. Ursell, J.C. Clemente, W. Van Treuren, W. a Walters, R. Knight, C.B. Newgard, A.C. Heath, J.I. Gordon, Gut microbiota from twins discordant for obesity modulate metabolism in mice., Science. 341 (2013) 1241214. doi:10.1126/science.1241214.

[68] A.E. Hoban, R.M. Stilling, F.J. Ryan, F. Shanahan, T.G. Dinan, M.J. Claesson, G. Clarke, J.F. Cryan, Regulation of prefrontal cortex myelination by the microbiota, Transl. Psychiatry. 6 (2016) e774. doi:10.1038/tp.2016.42.

[69] L. Desbonnet, G. Clarke, A. Traplin, O. O'Sullivan, F. Crispie, R.D. Moloney, P.D. Cotter, T.G. Dinan, J.F. Cryan, Gut microbiota depletion from early adolescence in mice: Implications for brain and behaviour, Brain. Behav. Immun. 48 (2015) 165–173. doi:10.1016/j.bbi.2015.04.004.

[70] F. Sommer, F. Bäckhed, The gut microbiota--masters of host development and physiology., Nat. Rev. Microbiol. 11 (2013) 227–38. doi:10.1038/nrmicro2974.

[71] R. Lundberg, M.F. Toft, B. August, A.K. Hansen, C.H.F. Hansen, Antibiotic-treated versus germ-free rodents for microbiota transplantation studies, Gut Microbes. 7 (2016) 68–74. doi:10.1080/19490976.2015.1127463.

[72] M. Ellekilde, E. Selfjord, C.S. Larsen, M. Jakesevic, I. Rune, B. Tranberg, F.K. Vogensen, D.S. Nielsen, M.I. Bahl, T.R. Licht, A.K. Hansen, C.H.F. Hansen, Transfer of gut microbiota from lean and obese mice to antibiotic-treated mice, Sci. Rep. 4 (2014) 5922. doi:10.1038/srep05922.

[73] D.H. Reikvam, A. Erofeev, A. Sandvik, V. Grcic, F.L. Jahnsen, P. Gaustad, K.D. McCoy, A.J. Macpherson, L.A. Meza-Zepeda, F.E. Johansen, Depletion of murine intestinal microbiota: Effects on gut mucosa and epithelial gene expression, PLoS One. 6 (2011) 1–13. doi:10.1371/journal.pone.0017996.

[74] I.I. Ivanov, K. Atarashi, N. Manel, E.L. Brodie, T. Shima, U. Karaoz, D. Wei, K.C. Goldfarb, C.A. Santee, S. V. Lynch, T. Tanoue, A. Imaoka, K. Itoh, K. Takeda, Y. Umesaki, K. Honda, D.R. Littman, Induction of Intestinal Th17 Cells by Segmented Filamentous Bacteria, Cell. 139 (2009) 485–498.

doi:10.1016/j.cell.2009.09.033.

[75] A.C. Ericsson, J.W. Davis, W. Spollen, N. Bivens, S. Givan, C.E. Hagan, M. McIntosh, C.L. Franklin, Effects of vendor and genetic background on the composition of the fecal microbiota of inbred mice, PLoS One. 10 (2015) 1–19. doi:10.1371/journal.pone.0116704.

[76] J. McCafferty, M. Mühlbauer, R.Z. Gharaibeh, J.C. Arthur, E. Perez-Chanona, W. Sha, C. Jobin, A. a Fodor, Stochastic changes over time and not founder effects drive cage effects in microbial community assembly in a mouse model., ISME J. 7 (2013) 2116–25. doi:10.1038/ismej.2013.106.

[77] R.B. Pyles, K.L. Vincent, M.M. Baum, B. Elsom, A.L. Miller, C. Maxwell, T.D. Eaves-Pyles, G. Li, V.L. Popov, R.J. Nusbaum, M.R. Ferguson, Cultivated vaginal microbiomes alter HIV-1 infection and antiretroviral efficacy in colonized epithelial multilayer cultures, PLoS One. 9 (2014) 1–12. doi:10.1371/journal.pone.0093419.

[78] P. Shah, J. V Fritz, E. Glaab, M.S. Desai, K. Greenhalgh, A. Frachet, M. Niegowska, M. Estes, C. Jager, C. Seguin-Devaux, F. Zenhausern, P. Wilmes, A microfluidics-based in vitro model of the gastrointestinal human-microbe interface, Nat Commun. 7 (2016). doi:10.1038/ncomms11535.

[79] W.A. Rose, C.L. McGowin, R.A. Spagnuolo, T.D. Eaves-Pyles, V.L. Popov, R.B. Pyles, Commensal bacteria modulate innate immune responses of vaginal epithelial cell multilayer cultures, PLoS One. 7 (2012) 1–11. doi:10.1371/journal.pone.0032728.

[80] T. Tsuruta, S. Saito, Y. Osaki, A. Hamada, A. Aoki-Yoshida, K. Sonoyama, Organoids as an ex vivo model for studying the serotonin system in the murine small intestine and colon epithelium, Biochem. Biophys. Res. Commun. 474 (2016) 161–167. doi:10.1016/j.bbrc.2016.03.165.

[81] S. Lukovac, C. Belzer, L. Pellis, B.J. Keijser, W.M. de Vos, R.C. Montijn, G. Roeselers, Differential modulation by Akkermansia muciniphila and faecalibacterium prausnitzii of host peripheral lipid metabolism and histone acetylation in mouse gut organoids, MBio. 5 (2014) 1–10. doi:10.1128/mBio.01438-14.

[82] T. Thomas, J. Gilbert, F. Meyer, Metagenomics - a guide from sampling to data analysis., Microb. Inform. Exp. 2 (2012) 3. doi:10.1186/2042-5783-2-3.

[83] J.C. Wooley, A. Godzik, I. Friedberg, A primer on metagenomics., PLoS Comput. Biol. 6 (2010) e1000667. doi:10.1371/journal.pcbi.1000667.

[84] Z. Liu, C. Lozupone, M. Hamady, F.D. Bushman, R. Knight, Short

pyrosequencing reads suffice for accurate microbial community analysis, Nucleic Acids Res. 35 (2007). doi:10.1093/nar/gkm541.

[85] P.J. Turnbaugh, M. Hamady, T. Yatsunenko, B.L. Cantarel, A. Duncan, R.E. Ley, M.L. Sogin, W.J. Jones, B.A. Roe, J.P. Affourtit, M. Egholm, B. Henrissat, A.C. Heath, R. Knight, J.I. Gordon, A core gut microbiome in obese and lean twins, Nature. 457 (2009) 480–485.

[86] S.E. Levy, R.M. Myers, Advancements in Next-Generation Sequencing, Annu. Rev. Genomics Hum. Genet. 17 (2016) annurev–genom–083115–022413. doi:10.1146/annurev-genom-083115-022413.

[87] H.P.J. Buermans, J.T. den Dunnen, Next generation sequencing technology: Advances and applications, Biochim. Biophys. Acta - Mol. Basis Dis. 1842 (2014) 1932–1941. doi:10.1016/j.bbadis.2014.06.015.

[88] Illumina, An Introduction to Next-Generation Sequencing Technology Table of Contents, 2015.

[89] M.-A. Madoui, S. Engelen, C. Cruaud, C. Belser, L. Bertrand, A. Alberti, A. Lemainque, P. Wincker, J.-M. Aury, Genome assembly using Nanopore-guided long and error-free DNA reads, BMC Genomics. 16 (2015) 327. doi:10.1186/s12864-015-1519-z.

[90] T. Laver, J. Harrison, P.A. O'Neill, K. Moore, A. Farbos, K. Paszkiewicz, D.J. Studholme, Assessing the performance of the Oxford Nanopore Technologies MinION, Biomol. Detect. Quantif. 3 (2015) 1–8. doi:10.1016/j.bdq.2015.02.001.

[91] J. Kuczynski, C.L. Lauber, W. a Walters, L.W. Parfrey, J.C. Clemente, D. Gevers, R. Knight, Experimental and analytical tools for studying the human microbiome., Nat. Rev. Genet. 13 (2012) 47–58. doi:10.1038/nrg3129.

[92] G.M. Weinstock, Genomic approaches to studying the human microbiota., Nature. 489 (2012) 250–6. doi:10.1038/nature11553.

[93] E.A. Grice, J.A. Segre, The Human Microbiome: Our Second Genome *, Annu. Rev. Genomics Hum. Genet. 13 (2012) 151–170. doi:10.1146/annurev-genom-090711-163814.

[94] J.G. Caporaso, C.L. Lauber, W.A. Walters, D. Berg-lyons, C.A. Lozupone, P.J. Turnbaugh, N. Fierer, R. Knight, Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample, (2010). doi:10.1073/pnas.1000080107/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1000080107.

[95] M.J. Claesson, Q. Wang, O. O'Sullivan, R. Greene-Diniz, J.R. Cole, R.P. Ross, P.W. O'Toole, Comparison of two next-generation sequencing technologies for

resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions., Nucleic Acids Res. 38 (2010) e200. doi:10.1093/nar/gkq873.

[96] Y. He, J.G. Caporaso, X.-T. Jiang, H.-F. Sheng, S.M. Huse, J.R. Rideout, R.C. Edgar, E. Kopylova, W. a Walters, R. Knight, H.-W. Zhou, Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity, Microbiome. 3 (2015) 1–10. doi:10.1186/s40168-015-0081-x.

[97] P.D. Schloss, S.L. Westcott, T. Ryabin, J.R. Hall, M. Hartmann, E.B. Hollister, R.A. Lesniewski, B.B. Oakley, D.H. Parks, C.J. Robinson, J.W. Sahl, B. Stres, G.G. Thallinger, D.J. Van Horn, C.F. Weber, Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities, Appl. Environ. Microbiol. 75 (2009) 7537–7541. doi:10.1128/AEM.01541-09.

[98] J.G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F.D. Bushman, E.K. Costello, N. Fierer, A.G. Peña, J.K. Goodrich, J.I. Gordon, G. a Huttley, S.T. Kelley, D. Knights, J.E. Koenig, R.E. Ley, C. a Lozupone, D. McDonald, B.D. Muegge, M. Pirrung, J. Reeder, J.R. Sevinsky, P.J. Turnbaugh, W. a Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, R. Knight, QIIME allows analysis of high-throughput community sequencing data, Nat. Methods. 7 (2010) 335–336.

[99] S.M. Huse, D.B. Mark Welch, A. Voorhis, A. Shipunova, H.G. Morrison, A.M. Eren, M.L. Sogin, VAMPS: a website for visualization and analysis of microbial population structures., BMC Bioinformatics. 15 (2014) 41. doi:10.1186/1471-2105-15-41.

[100] G. Van Rossum, Python, Python Softw. Found. (1997).

[101] R.C. Edgar, Search and clustering orders of magnitude faster than BLAST, Bioinformatics. 26 (2010) 2460–2461. doi:10.1093/bioinformatics/btq461.

[102] A. Oulas, C. Pavloudi, P. Polymenakou, G.A. Pavlopoulos, N. Papanikolaou, G. Kotoulas, C. Arvanitidis, I. Iliopoulos, Metagenomics: Tools and insights for analyzing next-generation sequencing data derived from biodiversity studies, Bioinform. Biol. Insights. 9 (2015) 75–88. doi:10.4137/BBIi.Ss12462.

[103] A.C. Howe, J. Jansson, S. a Malfatti, S.G. Tringe, J.M. Tiedje, C.T. Brown, Assembling large, complex environmental metagenomes, Audio, Trans. IRE Prof. Gr. (2012) 1–29. http://pubget.com/site/paper/pgtmp_12122832?institution=\npapers2://publication/uuid/4531F22E-ADCF-431E-BEE4-50FBB38C3F28.

[104] M.R. Crusoe, H.F. Alameldin, S. Awad, E. Boucher, A. Caldwell, R. Cartwright, A. Charbonneau, B. Constantinides, G. Edvenson, S. Fay, J. Fenton, T. Fenzl, J. Fish, L. Garcia-Gutierrez, P. Garland, J. Gluck, I. González, S. Guermond, J. Guo,

A. Gupta, J.R. Herr, A. Howe, A. Hyer, A. Härpfer, L. Irber, R. Kidd, D. Lin, J. Lippi, T. Mansour, P. McA'Nulty, E. McDonald, J. Mizzi, K.D. Murray, J.R. Nahum, K. Nanlohy, A.J. Nederbragt, H. Ortiz-Zuazaga, J. Ory, J. Pell, C. Pepe-Ranney, Z.N. Russ, E. Schwarz, C. Scott, J. Seaman, S. Sievert, J. Simpson, C.T. Skennerton, J. Spencer, R. Srinivasan, D. Standage, J.A. Stapleton, S.R. Steinman, J. Stein, B. Taylor, W. Trimble, H.L. Wiencko, M. Wright, B. Wyss, Q. Zhang, E. Zyme, C.T. Brown, The khmer software package: enabling efficient nucleotide sequence analysis., F1000Research. 4 (2015) 900. doi:10.12688/f1000research.6924.1.

[105] P.S. La Rosa, J.P. Brooks, E. Deych, E.L. Boone, D.J. Edwards, Q. Wang, E. Sodergren, G. Weinstock, W.D. Shannon, Hypothesis testing and power calculations for taxonomic-based human microbiome data., PLoS One. 7 (2012) e52078. doi:10.1371/journal.pone.0052078.

[106] H. Li, Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis, Annu. Rev. Stat. Its Appl. 2 (2015) 73–94. doi:10.1146/annurev-statistics-010814-020351.

[107] M.C.B. Tsilimigras, A.A. Fodor, Compositional Data Analysis of the Microbiome: Fundamentals, Tools, and Challenges, Ann. Epidemiol. 26 (2016) 330–335. doi:10.1016/j.annepidem.2016.03.002.

[108] P.J. McMurdie, S. Holmes, Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible, PLoS Comput. Biol. 10 (2014). doi:10.1371/journal.pcbi.1003531.

[109] T.C.J. Hill, K.A. Walsh, J.A. Harris, B.F. Moffett, Using ecological diversity measures with bacterial communities, FEMS Microbiol. Ecol. 43 (2003) 1–11.

[110] E.K. Morris, T. Caruso, F. Buscot, M. Fischer, C. Hancock, T.S. Maier, T. Meiners, C. Müller, E. Obermaier, D. Prati, S.A. Socher, I. Sonnemann, N. Wäschke, T. Wubet, S. Wurst, M.C. Rillig, Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories, Ecol. Evol. 4 (2014) 3514–3524. doi:10.1002/ece3.1155.

[111] B. Haegeman, J. Hamelin, J. Moriarty, P. Neal, J. Dushoff, J.S. Weitz, Robust estimation of microbial diversity in theory and in practice., ISME J. (2013) 1–10. doi:10.1038/ismej.2013.10.

[112] J. Oksanen, F.G. Blanchet, R. Kindt, P. Legendre, P.R. Minchin, R.B. O'Hara, G.L. Simpson, P. Solymos, M. Henry, H. Stevens, H. Wagner, vegan: Community Ecology Package R Version 2.3-4, R. (2016).

[113] P.J. McMurdie, S. Holmes, Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data, PLoS One. 8 (2013).

doi:10.1371/journal.pone.0061217.

[114] C.E. Robertson, J.K. Harris, B.D. Wagner, D. Granger, K. Browne, B. Tatem, L.M. Feazel, K. Park, N.R. Pace, D.N. Frank, Explicet: graphical user interface software for metadata-driven management, analysis and visualization of microbiome data., Bioinformatics. 29 (2013) 3100–3101. doi:10.1093/bioinformatics/btt526.

[115] M. Hamady, C. Lozupone, R. Knight, Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data., ISME J. 4 (2010) 17–27. doi:10.1038/ismej.2009.97.

[116] C. Lozupone, R. Knight, UniFrac : a New Phylogenetic Method for Comparing Microbial Communities UniFrac : a New Phylogenetic Method for Comparing Microbial Communities, 71 (2005). doi:10.1128/AEM.71.12.8228.

[117] A. Gonzalez-Martinez, A. Rodriguez-Sanchez, T. Lotti, M.J. Garcia-Ruiz, F. Osorio, J. Gonzalez-Lopez, M.C. van Loosdrecht, Comparison of bacterial communities of conventional and A-stage activated sludge systems, Sci Rep. 6 (2016) 18786. doi:10.1038/srep18786.

[118] M.G. Howard, W.J.F. McDonald, P.I. Forster, W.J. Kress, D. Erickson, D.P. Faith, A. Shapcott, Patterns of Phylogenetic Diversity of Subtropical Rainforest of the Great Sandy Region, Australia Indicate Long Term Climatic Refugia., PLoS One. 11 (2016) e0153565. doi:10.1371/journal.pone.0153565.

[119] Y. Kim, Y.S. Choi, K.J. Baek, S.-H. Yoon, H.K. Park, Y. Choi, Mucosal and salivary microbiota associated with recurrent aphthous stomatitis, BMC Microbiol. 16 (2016) 57. doi:10.1186/s12866-016-0673-z.

[120] I. Holmes, K. Harris, C. Quince, Dirichlet multinomial mixtures: generative models for microbial metagenomics., PLoS One. 7 (2012) e30126. doi:10.1371/journal.pone.0030126.

[121] E. Afgan, D. Baker, M. van den Beek, D. Blankenberg, D. Bouvier, M. Čech, J. Chilton, D. Clements, N. Coraor, C. Eberhard, B. Grüning, A. Guerler, J. Hillman-Jackson, G. Von Kuster, E. Rasche, N. Soranzo, N. Turaga, J. Taylor, A. Nekrutenko, J. Goecks, The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update, Nucleic Acids Res. 44 (2016) gkw343. doi:10.1093/nar/gkw343.

[122] N. Segata, J. Izard, L. Waldron, D. Gevers, L. Miropolsky, W.S. Garrett, C. Huttenhower, Metagenomic biomarker discovery and explanation., Genome Biol. 12 (2011) R60. doi:10.1186/gb-2011-12-6-r60.

[123] J. Wu, X.W. Wen, C. Faulk, K. Boehnke, H. Zhang, D.C. Dolinoy, C. Xi, Perinatal Lead Exposure Alters Gut Microbiota Composition and Results in Sex-specific Bodyweight Increases in Adult Mice, Toxicol. Sci. 151 (2016) kfw046. doi:10.1093/toxsci/kfw046.

[124] K. Faust, J.F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes, C. Huttenhower, Microbial co-occurrence relationships in the human microbiome., PLoS Comput. Biol. 8 (2012) e1002606. doi:10.1371/journal.pcbi.1002606.

[125] C.C. Evans, K.J. LePard, J.W. Kwak, M.C. Stancukas, S. Laskowski, J. Dougherty, L. Moulton, A. Glawe, Y. Wang, V. Leone, D.A. Antonopoulos, D. Smith, E.B. Chang, M.J. Ciancio, Exercise prevents weight gain and alters the gut microbiota in a mouse model of high fat diet-induced obesity, PLoS One. 9 (2014). doi:10.1371/journal.pone.0092193.

[126] I.H. McHardy, M. Goudarzi, M. Tong, P.M. Ruegger, E. Schwager, J.R. Weger, T.G. Graeber, J.L. Sonnenburg, S. Horvath, C. Huttenhower, D.P. McGovern, A.J. Fornace, J. Borneman, J. Braun, Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships., Microbiome. 1 (2013) 17. doi:10.1186/2049-2618-1-17.

[127] R. Artusi, P. Verderio, E. Marubini, Bravais-Pearson and Spearman correlation coefficients: Meaning, test of hypothesis and confidence interval, Int. J. Biol. Markers. 17 (2002) 148–151.

[128] J. Friedman, E.J. Alm, Inferring Correlation Networks from Genomic Survey Data, PLoS Comput. Biol. 8 (2012) 1–11. doi:10.1371/journal.pcbi.1002687.

[129] Z.D. Kurtz, C.L. Muller, E.R. Miraldi, D.R. Littman, M.J. Blaser, R.A. Bonneau, Sparse and Compositionally Robust Inference of Microbial Ecological Networks, PLoS Comput. Biol. 11 (2015) 1–25. doi:10.1371/journal.pcbi.1004226.

[130] T.T. Rasmussen, L.P. Kirkeby, K. Poulsen, J. Reinholdt, M. Kilian, Resident aerobic microbiota of the adult human nasal cavity, APMIS. 108 (2000) 663–675.

[131] A. Camarinha-Silva, R. Jáuregui, D.H. Pieper, M.L. Wos-Oxley, The temporal dynamics of bacterial communities across human anterior nares., Environ. Microbiol. Rep. 4 (2012) 126–32. doi:10.1111/j.1758-2229.2011.00313.x.

[132] M. Yan, S.J. Pamp, J. Fukuyama, P.H. Hwang, D.-Y. Cho, S. Holmes, D.A. Relman, Nasal Microenvironments and Interspecific Interactions Influence Nasal Microbiota Complexity and S. aureus Carriage, Cell Host Microbe. 14 (2013) 631–640. doi:10.1016/j.chom.2013.11.005.

[133] R.P. Dickson, J.R. Erb-Downward, F.J. Martinez, G.B. Huffnagle, The Microbiome and the Respiratory Tract., Annu. Rev. Physiol. (2015) 1–24.

doi:10.1146/annurev-physiol-021115-105238.

[134] E.S. Charlson, K. Bittinger, A.R. Haas, A.S. Fitzgerald, I. Frank, A. Yadav, F.D. Bushman, R.G. Collman, Topographical continuity of bacterial populations in the healthy human respiratory tract., Am. J. Respir. Crit. Care Med. 184 (2011) 957–63. doi:10.1164/rccm.201104-0655OC.

[135] B.J. Kelly, I. Imai, K. Bittinger, A. Laughlin, B.D. Fuchs, F.D. Bushman, R.G. Collman, Composition and dynamics of the respiratory tract microbiome in intubated patients, Microbiome. 4 (2016) 7. doi:10.1186/s40168-016-0151-8.

[136] A. Tam, S. Wadsworth, D. Dorscheid, S.F.P. Man, D.D. Sin, The airway epithelium: more than just a structural barrier., Ther. Adv. Respir. Dis. 5 (2011) 255–73. doi:10.1177/1753465810396539.

[137] M.I. Gómez, A. Prince, Airway epithelial cell signaling in response to bacterial pathogens., Pediatr. Pulmonol. 43 (2008) 11–9. doi:10.1002/ppul.20735.

[138] K. Hara, D. Zhang, Bacterial abundance and viability in long-range transported dust, Atmos. Environ. 47 (2012) 20–25. doi:10.1016/j.atmosenv.2011.11.050.

[139] R.P. Dickson, J.R. Erb-Downward, G.B. Huffnagle, Homeostasis and its Disruption in the Lung Microbiome, Am. J. Physiol. - Lung Cell. Mol. Physiol. (2015) ajplung.00279.2015. doi:10.1152/ajplung.00279.2015.

[140] B.H. Warren, D. Simberloff, R.E. Ricklefs, R. Aguilée, F.L. Condamine, D. Gravel, H. Morlon, N. Mouquet, J. Rosindell, J. Casquet, E. Conti, J. Cornuault, J.M. Fernández-Palacios, T. Hengl, S.J. Norder, K.F. Rijsdijk, I. Sanmartín, D. Strasberg, K.A. Triantis, L.M. Valente, R.J. Whittaker, R.G. Gillespie, B.C. Emerson, C. Thébaud, Islands as model systems in ecology and evolution: Prospects fifty years after MacArthur-Wilson, Ecol. Lett. 18 (2015) 200–217. doi:10.1111/ele.12398.

[141] R.P. Dickson, The microbiome and critical illness, Lancet Respir. Med. 4 (2016) 59–72. doi:10.1016/S2213-2600(15)00427-0.

[142] M. Pérez-Losada, E. Castro-Nallar, M.L. Bendall, R.J. Freishtat, K.A. Crandall, Dual Transcriptomic Profiling of Host and Microbiota during Health and Disease in Pediatric Asthma., PLoS One. 10 (2015) e0131819. doi:10.1371/journal.pone.0131819.

[143] A.A. Pragman, H.B. Kim, C.S. Reilly, C. Wendt, R.E. Isaacson, The lung microbiome in moderate and severe chronic obstructive pulmonary disease., PLoS One. 7 (2012) e47305. doi:10.1371/journal.pone.0047305.

[144] J.R. Erb-Downward, D.L. Thompson, M.K. Han, C.M. Freeman, L. McCloskey, L.

a Schmidt, V.B. Young, G.B. Toews, J.L. Curtis, B. Sundaram, F.J. Martinez, G.B. Huffnagle, Analysis of the lung microbiome in the "healthy" smoker and in COPD., PLoS One. 6 (2011) e16384. doi:10.1371/journal.pone.0016384.

[145] Y.J. Huang, C.E. Nelson, E.L. Brodie, T.Z. Desantis, M.S. Baek, J. Liu, T. Woyke, M. Allgaier, J. Bristow, J.P. Wiener-Kronish, E.R. Sutherland, T.S. King, N. Icitovic, R.J. Martin, W.J. Calhoun, M. Castro, L.C. Denlinger, E. Dimango, M. Kraft, S.P. Peters, S.I. Wasserman, M.E. Wechsler, H. a Boushey, S. V Lynch, Airway microbiota and bronchial hyperresponsiveness in patients with suboptimally controlled asthma., J. Allergy Clin. Immunol. 127 (2011) 372–381.e1–3. doi:10.1016/j.jaci.2010.10.048.

[146] M.K. Han, Y. Zhou, S. Murray, N. Tayob, I. Noth, V.N. Lama, B.B. Moore, E.S. White, K.R. Flaherty, G.B. Huffnagle, F.J. Martinez, Lung microbiome and disease progression in idiopathic pulmonary fibrosis: An analysis of the COMET study, Lancet Respir. Med. 2 (2014) 448–456. doi:10.1016/S2213-2600(14)70069-4.

[147] E.T. Zemanick, B.D. Wagner, C.E. Robertson, M.J. Stevens, S.J. Szefler, F.J. Accurso, S.D. Sagel, J.K. Harris, Assessment of Airway Microbiota and Inflammation in Cystic Fibrosis Using Multiple Sampling Methods, Ann. Am. Thorac. Soc. 12 (2015) 221–229. doi:10.1513/AnnalsATS.201407-310OC.

[148] L.A. Carmody, J. Zhao, L.M. Kalikin, W. LeBar, R.H. Simon, A. Venkataraman, T.M. Schmidt, Z. Abdo, P.D. Schloss, J.J. LiPuma, The daily dynamics of cystic fibrosis airway microbiota during clinical stability and at exacerbation, Microbiome. 3 (2015) 12. doi:10.1186/s40168-015-0074-9.

[149] R.J. Boyton, C.J. Reynolds, K.J. Quigley, D.M. Altmann, Immune mechanisms and the impact of the disrupted lung microbiome in chronic bacterial lung infection and bronchiectasis., Clin. Exp. Immunol. 171 (2013) 117–23. doi:10.1111/cei.12003.

[150] D. Collie, L. Glendinning, J. Govan, S. Wright, E. Thornton, P. Tennant, C. Doherty, G. McLachlan, Lung Microbiota Changes Associated with Chronic Pseudomonas aeruginosa Lung Infection and the Impact of Intravenous Colistimethate Sodium, PLoS One. 10 (2015) e0142097. doi:10.1371/journal.pone.0142097.

[151] W.T. Sloan, M. Lunn, S. Woodcock, I.M. Head, S. Nee, T.P. Curtis, Quantifying the roles of immigration and chance in shaping prokaryote community structure, Environ. Microbiol. 8 (2006) 732–740. doi:10.1111/j.1462-2920.2005.00956.x.

[152] F. Fouhy, J. Deane, M.C. Rea, Ó. O'Sullivan, R.P. Ross, G. O'Callaghan, B.J. Plant, C. Stanton, The effects of freezing on faecal microbiota as determined using miseq sequencing and culture-based investigations, PLoS One. 10 (2015) 1–12.

doi:10.1371/journal.pone.0119355.

[153] G.A. Cangelosi, J.S. Meschke, Dead or alive: Molecular assessment of microbial viability, Appl. Environ. Microbiol. 80 (2014) 5884–5891. doi:10.1128/AEM.01763-14.

[154] J.G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F.D. Bushman, E.K. Costello, N. Fierer, A.G. Peña, J.K. Goodrich, J.I. Gordon, G. a Huttley, S.T. Kelley, D. Knights, J.E. Koenig, R.E. Ley, C. a Lozupone, D. McDonald, B.D. Muegge, M. Pirrung, J. Reeder, J.R. Sevinsky, P.J. Turnbaugh, W. a Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, R. Knight, QIIME allows analysis of high-throughput community sequencing data., Nat. Methods. 7 (2010) 335–6. doi:10.1038/nmeth.f.303.

[155] T. Nordahl Petersen, S. Rasmussen, H. Hasman, C. Carøe, J. Bælum, A. Charlotte Schultz, L. Bergmark, C. a. Svendsen, O. Lund, T. Sicheritz-Pontén, F.M. Aarestrup, Meta-genomic analysis of toilet waste from long distance flights; a step towards global surveillance of infectious diseases and antimicrobial resistance, Sci. Rep. 5 (2015) 11444. doi:10.1038/srep11444.

[156] P.-A. Neumann, S. Koch, R.S. Hilgarth, E. Perez-Chanona, P. Denning, C. Jobin, A. Nusrat, Gut commensal bacteria and regional wnt gene expression in the proximal versus distal colon., Am. J. Pathol. 184 (2014) 592–9. doi:10.1016/j.ajpath.2013.11.029.

[157] C.T. Brown, I. Sharon, B.C. Thomas, C.J. Castelle, M.J. Morowitz, J.F. Banfield, Genome resolved analysis of a premature infant gut microbial community reveals a Varibaculum cambriense genome and a shift towards fermentation-based metabolism during the third week of life., Microbiome. 1 (2013) 30. doi:10.1186/2049-2618-1-30.

[158] S. Bunyavanich, N. Shen, A. Grishin, R. Wood, W. Burks, P. Dawson, S.M. Jones, D. Leung, H. Sampson, S. Sicherer, J.C. Clemente, Early-life gut microbiome composition and milk allergy resolution, J. Allergy Clin. Immunol. (2016) 1–9. doi:10.1016/j.jaci.2016.03.041.

[159] P. McInnes, M. Cutting, Manual of Procedures for Human Microbiome Project, 2010.

[160] C. Yeates, M. Gillings, A. Davison, N. Altavilla, D. Veal, Methods for microbial DNA extraction from soil for PCR amplification, Biol. Proced. Online. 1 (1998) 40–47. doi:10.1251/bpo6.

[161] J.R. de Lipthay, C. Enzinger, K. Johnsen, J. Aamand, S.J. Sørensen, Impact of DNA extraction method on bacterial community composition measured by denaturing gradient gel electrophoresis, Soil Biol. Biochem. 36 (2004) 1607–1614.

doi:10.1016/j.soilbio.2004.03.011.

[162] R. de Boer, R. Peters, S. Gierveld, T. Schuurman, M. Kooistra-Smid, P. Savelkoul, Improved detection of microbial DNA after bead-beating before DNA isolation, J. Microbiol. Methods. 80 (2010) 209–211. doi:10.1016/j.mimet.2009.11.009.

[163] S. Yuan, D.B. Cohen, J. Ravel, Z. Abdo, L.J. Forney, Evaluation of methods for the extraction and purification of DNA from the human microbiome., PLoS One. 7 (2012) e33865. doi:10.1371/journal.pone.0033865.

[164] G. Biesbroek, E. a M. Sanders, G. Roeselers, X. Wang, M.P.M. Caspers, K. Trzciński, D. Bogaert, B.J.F. Keijser, Deep sequencing analyses of low density microbial communities: working at the boundary of accurate microbiota detection., PLoS One. 7 (2012) e32942. doi:10.1371/journal.pone.0032942.

[165] D. Willner, J. Daly, D. Whiley, K. Grimwood, C.E. Wainwright, P. Hugenholtz, Comparison of DNA extraction methods for microbial community profiling with an application to pediatric bronchoalveolar lavage samples., PLoS One. 7 (2012) e34605. doi:10.1371/journal.pone.0034605.

[166] J.P. Brooks, D.J. Edwards, M.D. Harwich, M.C. Rivera, J.M. Fettweis, M.G. Serrano, R.A. Reris, N.U. Sheth, B. Huang, P. Girerd, J.F. Strauss, K.K. Jefferson, G.A. Buck, The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies., BMC Microbiol. 15 (2015) 66. doi:10.1186/s12866-015-0351-6.

[167] H.P. Horz, M.E. Vianna, B.P.F.A. Gomes, G. Conrads, Evaluation of Universal Probes and Primer Sets for Assessing Total Bacterial Load in Clinical Samples : General Implications and Practical Use in Endodontic Antimicrobial Therapy, J. Clin. Microbiol. 43 (2005) 5332–5337. doi:10.1128/JCM.43.10.5332.

[168] M.A. Nadkarni, F.E. Martin, N.A. Jacques, N. Hunter, Determination of bacterial load by real-time PCR using a broad-range (universal) probe and primers set, (2002) 257–266.

[169] R. Rastogi, M. Wu, I. Dasgupta, G.E. Fox, Visualization of ribosomal RNA operon copy number distribution., BMC Microbiol. 9 (2009) 208. doi:10.1186/1471-2180-9-208.

[170] M.G.I. Langille, J. Zaneveld, J.G. Caporaso, D. McDonald, D. Knights, J. a Reyes, J.C. Clemente, D.E. Burkepile, R.L. Vega Thurber, R. Knight, R.G. Beiko, C. Huttenhower, Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences., Nat. Biotechnol. 31 (2013) 814–21. doi:10.1038/nbt.2676.

[171] F.E. Angly, P.G. Dennis, A. Skarshewski, I. Vanwonterghem, P. Hugenholtz,

G.W. Tyson, CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction., Microbiome. 2 (2014) 11. doi:10.1186/2049-2618-2-11.

[172] R.J. Case, Y. Boucher, I. Dahllöf, C. Holmström, W.F. Doolittle, S. Kjelleberg, Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies., Appl. Environ. Microbiol. 73 (2007) 278–88. doi:10.1128/AEM.01177-06.

[173] M. Vos, C. Quince, A.S. Pijl, M. de Hollander, G. a Kowalchuk, A comparison of rpoB and 16S rRNA as markers in pyrosequencing studies of bacterial diversity., PLoS One. 7 (2012) e30600. doi:10.1371/journal.pone.0030600.

[174] C.K. Murray, Burns, in: Mand. Douglas, Bennett's Princ. Pract. Infect. Dis., Eighth, 2015.

[175] American Burn Association, National Burn Repository 2015 Report version 11.0, 2015.

[176] F.F. Ferri, Burns, in: Ferri's Clin. Advis. 2017, 2017: pp. 219–221.

[177] M.G. Jeschke, D.N. Herndon, Burns, in: Sabist. Textb. Surg., Twentieth, 2017: pp. 505–531.

[178] L.C. Cancio, Airway management and smoke inhalation injury in the burn patient., Clin. Plast. Surg. 36 (2009) 555–67. doi:10.1016/j.cps.2009.05.013.

[179] C.S. Davis, S.E. Janus, M.J. Mosier, S.R. Carter, J.T. Gibbs, L. Ramirez, R.L. Gamelli, E.J. Kovacs, Inhalation Injury Severity and Systemic Immune Perturbations in Burned Adults, Ann. Surg. 257 (2013) 1137–1146.

[180] D. a Edelman, N. Khan, K. Kempf, M.T. White, Pneumonia after inhalation injury., J. Burn Care Res. 28 (2007) 241–6. doi:10.1097/BCR.0B013E318031D049.

[181] D.J. Dries, F.W. Endorf, Inhalation injury: epidemiology, pathology, treatment strategies., Scand. J. Trauma. Resusc. Emerg. Med. 21 (2013) 31. doi:10.1186/1757-7241-21-31.

[182] Z. Hassan, J.K. Wong, J. Bush, a Bayat, K.W. Dunn, Assessing the severity of inhalation injuries in adults., Burns. 36 (2010) 212–6. doi:10.1016/j.burns.2009.06.205.

[183] N. Brusselaers, D. Logie, D. Vogelaers, S. Monstrey, S. Blot, Burns, inhalation injury and ventilator-associated pneumonia: value of routine surveillance cultures., Burns. 38 (2012) 364–70. doi:10.1016/j.burns.2011.09.005.

[184] T. Prien, D.L. Traber, Toxic smoke compounds and inhalation injury--a review., Burns. Incl. Therm. Inj. 14 (1988) 451–60. http://www.ncbi.nlm.nih.gov/pubmed/2855039.

[185] P. Enkhbaatar, D.L. Traber, Pathophysiology of acute lung injury in combined burn and smoke inhalation injury, Clin. Sci. 107 (2004) 137–143.

[186] M. Stefanidou, S. Athanaselis, C. Spiliopoulou, Health impacts of fire smoke inhalation., Inhal. Toxicol. 20 (2008) 761–6. doi:10.1080/08958370801975311.

[187] M.H. Toon, M.O. Maybauer, J.E. Greenwood, D.M. Maybauer, J.F. Fraser, Management of acute smoke inhalation injury., Crit. Care Resusc. 12 (2010) 53–61. http://www.ncbi.nlm.nih.gov/pubmed/20196715.

[188] D.L. Traber, H.A. Linares, D.N. Herndon, The pathophysiology of inhalation injury - a review, Burns. 14 (1988) 357–364.

[189] R.H. Demling, Smoke inhalation lung injury: an update., Eplasty. 8 (2008) e27. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2396464&tool=pmcentrez&rendertype=abstract.

[190] R. a Balk, Systemic inflammatory response syndrome (SIRS): Where did it come from and is it still relevant today?, Virulence. 5 (2014) 20–26. doi:10.4161/viru.27135.

[191] N.S. Ward, B. Casserly, A. Ayala, The compensatory anti-inflammatory response syndrome (CARS) in critically ill patients., Clin. Chest Med. 29 (2008) 617–25, viii. doi:10.1016/j.ccm.2008.06.010.

[192] W. Xiao, M.N. Mindrinos, J. Seok, J. Cuschieri, A.G. Cuenca, H. Gao, D.L. Hayden, L. Hennessy, E.E. Moore, J.P. Minei, P.E. Bankey, J.L. Johnson, J. Sperry, A.B. Nathens, T.R. Billiar, M. a West, B.H. Brownstein, P.H. Mason, H. V Baker, C.C. Finnerty, M.G. Jeschke, M.C. López, M.B. Klein, R.L. Gamelli, N.S. Gibran, B. Arnoldo, W. Xu, Y. Zhang, S.E. Calvano, G.P. McDonald-Smith, D. a Schoenfeld, J.D. Storey, J.P. Cobb, H.S. Warren, L.L. Moldawer, D.N. Herndon, S.F. Lowry, R. V Maier, R.W. Davis, R.G. Tompkins, A genomic storm in critically injured humans., J. Exp. Med. 208 (2011) 2581–90. doi:10.1084/jem.20111354.

[193] C.S. Davis, J.M. Albright, S.R. Carter, L. Ramirez, H. Kim, R.L. Gamelli, E.J. Kovacs, Early pulmonary immune hyporesponsiveness is associated with mortality after burn and smoke inhalation injury., J. Burn Care Res. 33 (2011) 26–35. doi:10.1097/BCR.0b013e318234d903.

[194] C.S. Davis, S.E. Janus, M.J. Mosier, S.R. Carter, J.T. Gibbs, L. Ramirez, R.L. Gamelli, E.J. Kovacs, Inhalation injury severity and systemic immune

perturbations in burned adults., Ann. Surg. 257 (2013) 1137–46. doi:10.1097/SLA.0b013e318275f424.

[195] A. Baldea, R.L. Gamelli, Burns and Inhalation Injury, in: Textb. Crit. Care, 6th ed., 2011: pp. 491–497.

[196] J.A. Bastarache, L.B. Ware, G.R. Bernard, Acute Lung Injury and Respiratory Distress Syndrome, in: Textb. Crit. Care, 2011: pp. 388–397.

[197] S.D. Pouwels, I.H. Heijink, N.H. Ten Hacken, P. Vandenabeele, D. V Krysko, M.C. Nawijn, a J. van Oosterhout, DAMPs activating innate and adaptive immune responses in COPD., Mucosal Immunol. (2013) 1–12. doi:10.1038/mi.2013.77.

[198] Z. Gregus, Mechanisms of Toxicity, in: C.D. Klaassen (Ed.), Casarett Doull's Toxicol. Basic Sci. Poisons, Seventh, McGraw Hill, New York, 2008: pp. 45–106.

[199] S.W. Jones, H. Zhou, S.M. Ortiz-Pujols, R. Maile, M. Herbst, B.L. Joyner, H. Zhang, M. Kesic, I. Jaspers, K. a Short, A. a Meyer, D.B. Peden, B. a Cairns, T.L. Noah, Bronchoscopy-derived correlates of lung injury following inhalational injuries: a prospective observational study., PLoS One. 8 (2013) e64250. doi:10.1371/journal.pone.0064250.

[200] W. Wu, D. Peden, D. Diaz-Sanchez, Role of GSTM1 in resistance to lung inflammation., Free Radic. Biol. Med. 53 (2012) 721–9. doi:10.1016/j.freeradbiomed.2012.05.037.

[201] G. Smith, M. Shlipak, E. Havranek, J. Foody, F. Masoudi, S. Rathore, H. Krumholz, Serum urea nitrogen, creatinine, and estimators of renal function: mortality in older patients with cardiovascular disease., Arch. Intern. Med. 166 (2006) 1134–1142.

[202] R.G. Schnellmann, Toxic Responses of the Kidney, in: C.D. Klaassen (Ed.), Casarett Doull's Toxicol. Basic Sci. Poisons, Seventh, McGraw Hill, New York, 2008: pp. 583–608.

[203] E.S. Charlson, K. Bittinger, J. Chen, J.M. Diamond, H. Li, R.G. Collman, F.D. Bushman, Assessing bacterial populations in the lung by replicate analysis of samples from the upper and lower respiratory tracts., PLoS One. 7 (2012) e42786. doi:10.1371/journal.pone.0042786.

[204] R.P. Dickson, G.B. Huffnagle, The Lung Microbiome: New Principles for Respiratory Bacteriology in Health and Disease, PLOS Pathog. 11 (2015) e1004923. doi:10.1371/journal.ppat.1004923.

[205] R.P. Dickson, J.R. Erb-Downward, G.B. Huffnagle, Towards an ecology of the lung: New conceptual models of pulmonary microbiology and pneumonia

pathogenesis, Lancet Respir. Med. 2 (2014) 238–246. doi:10.1016/S2213-2600(14)70028-1.

[206] A.J. Ghio, C.B. Smith, M.C. Madden, Diesel exhaust particles and airway inflammation., Curr. Opin. Pulm. Med. 18 (2012) 144–50. doi:10.1097/MCP.0b013e32834f0e2a.

[207] L. Jayaran, N.R. Labiris, A. Efthimiadis, H. Vlachos-Mayer, F.E. Hargreave, A.P. Freitag, The efficiency of sputum cell counts in cystic fibrosis, Can. Respir. J. 14 (2007) 99–103.

[208] M.M. Tunney, T.R. Field, T.F. Moriarty, S. Patrick, G. Doering, M.S. Muhlebach, M.C. Wolfgang, R. Boucher, D.F. Gilpin, A. McDowell, J.S. Elborn, Detection of Anaerobic Bacteria in High Numbers in Sputum from Patients with Cystic Fibrosis, Am. J. Respir. Crit. Care Med. 177 (2008) 995–1001. doi:10.1164/rccm.200708-1151OC.

[209] G.R. Feehery, E. Yigit, S.O. Oyola, B.W. Langhorst, V.T. Schmidt, F.J. Stewart, E.T. Dimalanta, L. a Amaral-Zettler, T. Davis, M. a Quail, S. Pradhan, A method for selectively enriching microbial DNA from contaminating vertebrate host DNA., PLoS One. 8 (2013) e76096. doi:10.1371/journal.pone.0076096.

[210] M.R.J. Salton, K.-S. Kim, Structure, in: S. Baron (Ed.), Med. Microbiol., Fourth, University of Texas Medical Branch at Galveston, Galveston, TX, 1996.

[211] S.M. Travis, B. a Conway, J. Zabner, J.J. Smith, N.N. Anderson, P.K. Singh, E.P. Greenberg, M.J. Welsh, Activity of abundant antimicrobials of the human airway., Am. J. Respir. Cell Mol. Biol. 20 (1999) 872–9. doi:10.1165/ajrcmb.20.5.3572.

[212] T.J. Foster, Immune evasion by staphylococci., Nat. Rev. Microbiol. 3 (2005) 948–58. doi:10.1038/nrmicro1289.

[213] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, Isolating, Cloning, and Sequencing DNA, in: Mol. Biol. Cell, Fourth, Garland Science, New York, 2002.

[214] H. Maeda, C. Fujimoto, Y. Haruki, T. Maeda, S. Kokeguchi, M. Petelin, H. Arai, I. Tanimoto, F. Nishimura, S. Takashiba, Quantitative real-time PCR using TaqMan and SYBR Green for Actinobacillus actinomycetemcomitans , Porphyromonas gingivalis , Prevotella intermedia , tetQ gene and total bacteria, FEMS Immunol. Med. Microbiol. 39 (2003) 81–86. doi:10.1016/S0928-8244(03)00224-4.

[215] D.S. Lundberg, S. Yourstone, P. Mieczkowski, C.D. Jones, J.L. Dangl, Practical innovations for high-throughput amplicon sequencing., Nat. Methods. 10 (2013) 999–1002. doi:10.1038/nmeth.2634.

[216] N.A. Bokulich, S. Subramanian, J.J. Faith, D. Gevers, J.I. Gordon, R. Knight, D.A. Mills, J.G. Caporaso, Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing., Nat. Methods. 10 (2013) 57–9. doi:10.1038/nmeth.2276.

[217] S.M. Yourstone, D.S. Lundberg, J.L. Dangl, C.D. Jones, MT-Toolbox: improved amplicon sequencing using molecule tags, BMC Bioinformatics. 15 (2014) 284. doi:10.1186/1471-2105-15-284.

[218] R.C. Edgar, UPARSE: highly accurate OTU sequences from microbial amplicon reads., Nat. Methods. 10 (2013) 996–8. doi:10.1038/nmeth.2604.

[219] T.Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E.L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, G.L. Andersen, Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB, Appl. Environ. Microbiol. 72 (2006) 5069–5072. doi:10.1128/AEM.03006-05.

[220] R.H. El-Helbawy, F.M. Ghareeb, Inhalation injury as a prognostic factor for mortality in burn patients., Ann. Burns Fire Disasters. 24 (2011) 82–8. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3230152&tool=pmcentrez&rendertype=abstract.

[221] The ARDS Task Force, Acute Respiratory Distress Syndrome, Jama. 307 (2012) 2526–2533. doi:10.1001/jama.2012.5669.

[222] D. Dreyfuss, J.-D. Ricard, Acute Lung Injury and Bacterial Infection., Clin. Chest Med. 26 (2005) 105–112.

[223] A. Zumla, J.A. Al-Tawfiq, V.I. Enne, M. Kidd, C. Drosten, J. Breuer, M.A. Muller, D. Hui, M. Maeurer, M. Bates, P. Mwaba, R. Al-Hakeem, G. Gray, P. Gautret, A.A. Al-Rabeeah, Z.A. Memish, V. Gant, Rapid point of care diagnostic tests for viral and bacterial respiratory tract infections--needs, advances, and future prospects., Lancet. Infect. Dis. 14 (2014) 1123–35. doi:10.1016/S1473-3099(14)70827-8.

[224] S.D. McCullough, X. Xu, S.Y.R. Dent, S. Bekiranov, R.G. Roeder, P. a Grant, Reelin is a target of polyglutamine expanded ataxin-7 in human spinocerebellar ataxia type 7 (SCA7) astrocytes., Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 21319–24. doi:10.1073/pnas.1218331110.

[225] A.L. Cozens, M.J. Yezzi, K. Kunzelmann, T. Ohrui, L. Chin, K. Eng, W.E. Finkbeiner, J.H. Widdicombe, D.C. Gruenert, CFTR expression and chloride secretion in polarized immortal human bronchial epithelial cells., Am. J. Respir. Cell Mol. Biol. 10 (1994) 38–47. doi:10.1165/ajrcmb.10.1.7507342.

[226] D.S. Lundberg, S.L. Lebeis, S.H. Paredes, S. Yourstone, J. Gehring, S. Malfatti, J.

Tremblay, A. Engelbrektson, V. Kunin, T.G. Del Rio, R.C. Edgar, T. Eickhorst, R.E. Ley, P. Hugenholtz, S.G. Tringe, J.L. Dangl, Defining the core Arabidopsis thaliana root microbiome, Nature. 488 (2012) 86–90. doi:10.1038/nature11237.

[227] R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, (2013). http://www.r-project.org/.

[228] B. Haegeman, J. Hamelin, J. Moriarty, P. Neal, J. Dushoff, J.S. Weitz, Robust estimation of microbial diversity in theory and in practice., ISME J. (2013) 1092–1101. doi:10.1038/ismej.2013.10.

[229] C.-C. Lai, M.-I. Sung, H.-H. Liu, C.-M. Chen, S.-R. Chiang, W.-L. Liu, C.-M. Chao, C.-H. Ho, S.-F. Weng, S.-C. Hsing, K.-C. Cheng, The Ratio of Partial Pressure Arterial Oxygen and Fraction of Inspired Oxygen 1 Day After Acute Respiratory Distress Syndrome Onset Can Predict the Outcomes of Involving Patients., Medicine (Baltimore). 95 (2016) e3333. doi:10.1097/MD.0000000000003333.

[230] C.-C. Lin, a a Liem, C.-K. Wu, Y.-F. Wu, J.-Y. Yang, C.-H. Feng, Severity score for predicting pneumonia in inhalation injury patients., Burns. 38 (2012) 203–7. doi:10.1016/j.burns.2011.08.010.

[231] R. Cohen-Poradosu, D.L. Kasper, Anaerobic Infections, in: J.E. Bennett, R. Dolin, M.J. Blaser (Eds.), Mand. Douglas, Bennett's Princ. Pract. Infect. Dis., Eight, 2015: p. 2736 0 2743.

[232] J.C. Arthur, R.Z. Gharaibeh, M. Mühlbauer, E. Perez-Chanona, J.M. Uronis, J. McCafferty, A.A. Fodor, C. Jobin, Microbial genomic analysis reveals the essential role of inflammation in bacteria-induced colorectal cancer., Nat. Commun. 5 (2014) 4724. doi:10.1038/ncomms5724.

[233] W. a a de Steenhuijsen Piters, E.G.W. Huijskens, A.L. Wyllie, G. Biesbroek, M.R. van den Bergh, R.H. Veenhoven, X. Wang, K. Trzciński, M.J. Bonten, J.W. a Rossen, E. a M. Sanders, D. Bogaert, Dysbiosis of upper respiratory tract microbiota in elderly pneumonia patients., ISME J. (2015) 1–12. doi:10.1038/ismej.2015.99.

[234] K.P. Lemon, V. Klepac-ceraj, H.K. Schiffer, E.L. Brodie, S. V Lynch, R. Kolter, Comparative Analyses of the Bacterial Microbiota of the Human Nostril and Oropharynx, MBio. 1 (2010). doi:10.1128/mBio.00129-10.Editor.

[235] D.N. Frank, L.M. Feazel, M.T. Bessesen, C.S. Price, E.N. Janoff, N.R. Pace, The human nasal microbiota and Staphylococcus aureus carriage., PLoS One. 5 (2010) e10598. doi:10.1371/journal.pone.0010598.

[236] R.P. Dickson, J.R. Erb-Downward, G.B. Huffnagle, The role of the bacterial

microbiome in lung disease, Expert Rev Respir Med. 7 (2013) 245–257. doi:10.1037/a0030561.Striving.

[237] S.C. Buckingham, Bacteroides, Fusobacterium, and Prevotella, in: J.D. Cherry, G.J. Harrison, S.L. Kaplan, W.J. Steinbach, P.J. Hotez (Eds.), Feigin Cherry's Textb. Pediatr. Infect. Dis., Seventh, 2014: pp. 1825–1834.

[238] J.M. Larsen, D.B. Steen-Jensen, J.M. Laursen, J.N. Søndergaard, H.S. Musavian, T.M. Butt, S. Brix, Divergent pro-inflammatory profile of human dendritic cells in response to commensal and pathogenic bacteria associated with the airway microbiota., PLoS One. 7 (2012) e31976. doi:10.1371/journal.pone.0031976.

[239] C. Agvald-Ohman, J. Wernerman, C.E. Nord, C. Edlund, Anaerobic bacteria commonly colonize the lower airways of intubated ICU patients, Clin Microbiol Infect. 9 (2003) 397–405. doi:551 [pii].

[240] T.R. Field, C.D. Sibley, M.D. Parkins, H.R. Rabin, M.G. Surette, The genus Prevotella in cystic fibrosis airways, Anaerobe. 16 (2010) 337–344. doi:10.1016/j.anaerobe.2010.04.002.

[241] K.Z. Shirani, B. a Pruitt, a D. Mason, The influence of inhalation injury and pneumonia on burn mortality., Ann. Surg. 205 (1987) 82–7. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1492872&tool=pmcentrez&rendertype=abstract.

[242] L. Schultz, S. a N. Walker, M. Elligsen, S.E. Walker, A. Simor, S. Mubareka, N. Daneman, Identification of predictors of early infection in acute burn patients., Burns. 39 (2013) 1355–66. doi:10.1016/j.burns.2013.04.009.

[243] M.O. a Sommer, G. Dantas, Antibiotics and the resistant microbiome., Curr. Opin. Microbiol. 14 (2011) 556–63. doi:10.1016/j.mib.2011.07.005.

[244] C. Ubeda, E.G. Pamer, Antibiotics, microbiota, and immune defense., Trends Immunol. 33 (2012) 459–66. doi:10.1016/j.it.2012.05.003.

[245] D. a Hill, C. Hoffmann, M.C. Abt, Y. Du, D. Kobuley, T.J. Kirn, F.D. Bushman, D. Artis, Metagenomic analyses reveal antibiotic-induced temporal and spatial changes in intestinal microbiota with associated alterations in immune cell homeostasis., Mucosal Immunol. 3 (2010) 148–58. doi:10.1038/mi.2009.132.

[246] S.E. Brill, M. Law, E. El-Emir, J.P. Allinson, P. James, V. Maddox, G.C. Donaldson, T.D. McHugh, W.O. Cookson, M.F. Moffatt, I. Nazareth, J.R. Hurst, P.M.A. Calverley, M.J. Sweeting, J.A. Wedzicha, Effects of different antibiotic classes on airway bacteria in stable COPD using culture and molecular techniques: a randomised controlled trial., Thorax. 70 (2015) 930–8. doi:10.1136/thoraxjnl-2015-207194.

[247] E. Pasolli, T. Truong, F. Malik, L. Waldron, N. Segata, Machine learning meta-analysis of large metagenomic datasets : tools and biological insights, PLoS Comput. Biol. accepted (2016) 1–26. doi:10.1371/journal.pcbi.1004977.

[248] A. Barberan, S.T. Bates, E.O. Casamayor, N. Fierer, Using network analysis to explore co-occurrence patterns in soil microbial communities, ISME J. 6 (2012) 343–351. doi:10.1038/ismej.2011.119.

[249] A. Statnikov, M. Henaff, V. Narendra, K. Konganti, Z. Li, L. Yang, Z. Pei, M.J. Blaser, C.F. Aliferis, A. V Alekseyenko, A comprehensive evaluation of multicategory classification methods for microbiomic data., Microbiome. 1 (2013) 11. doi:10.1186/2049-2618-1-11.

[250] P.E. Larsen, D. Field, J.A. Gilbert, Predicting bacterial community assemblages using an artificial neural network approach., Nat. Methods. 9 (2012) 621–5. doi:10.1038/nmeth.1975.

[251] R. Kop, M. Hoogendoorn, A. ten Teije, F.L. Büchner, P. Slottje, L.M.G. Moons, M.E. Numans, Predictive modeling of colorectal cancer using a dedicated pre-processing pipeline on routine electronic medical records, Comput. Biol. Med. 76 (2016) 30–38. doi:10.1016/j.compbiomed.2016.06.019.

[252] F. Iorio, T.A. Knijnenburg, D.J. Vis, J. Saez-Rodriguez, U. Mcdermott, M.J.G. Correspondence, G.R. Bignell, M.P. Menden, M. Schubert, N. Aben, E. Gonç Alves, S. Barthorpe, H. Lightfoot, T. Cokelaer, P. Greninger, E. Van Dyk, H. Chang, H. De Silva, H. Heyn, X. Deng, R.K. Egan, Q. Liu, T. Mironenko, X. Mitropoulos, L. Richardson, J. Wang, M.J. Garnett, A Landscape of Pharmacogenomic Interactions in Cancer, Cell Tinghu Zhang. 16613616 (2016) 1–15. doi:10.1016/j.cell.2016.06.017.

[253] A. Monsalve-Torra, D. Ruiz-Fernandez, O. Marin-Alonso, A. Soriano-Payá, J. Camacho-Mackenzie, M. Carreñ-Jaimes, Using Machine Learning Methods for Predicting Inhospital Mortality in Patients Undergoing Open Repair of Abdominal Aortic Aneurysm, J. Biomed. Inform. (2016). doi:10.1016/j.jbi.2016.07.007.

[254] M. Kubat, An Introduction to Machine Learning, 2015. doi:10.1007/978-3-319-20010-1.

[255] D.L. Donoho, A. Averbuch, T. Hastie, I. Johnstone, A. Owen, D. Scott, B. Stine, R. Tibshirani Stanford University Walter Willinger, G.U. Piatetsky-Shapiro XChange Edward Bosch Army Topographic Eng Ctr John Elder, A. Inselberg, J.-L. Starck Centre Europeen, A. Arne Stoschek, P. Weinberger, M. Clerc, A. Georgina Flesia, K. Jennings, O. Levi, High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality, 2000.

[256] M. Verleysen, D. François, The Curse of Dimensionality in Data Mining,

Analysis. 3512 (2005) 758 – 770. doi:10.1007/11494669_93.

[257] G. Casella, S. Fienberg, I. Olkin, An Introduction to Statistical Learning, 2006. doi:10.1016/j.peva.2007.06.006.

[258] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods. 6 (2014) 2812–2831. doi:10.1016/0169-7439(87)80084-9.

[259] T. Jombart, S. Devillard…, Discriminant analysis of principal components: a new method for the analysis of genetically structured populations, BMC Genet. (2010). doi:doi:10.1186/1471-2156-11-94.

[260] T. Jombart, A tutorial for Discriminant Analysis of Principal Components (DAPC) using adegenet 1 . 3-4, Rvignette. (2012) 1–37.

[261] M. Kuhn, K. Johnson, Applied Predictive Modeling [Hardcover], 2013. doi:10.1007/978-1-4614-6849-3.

[262] Y. Ranganathan, R.M. Borges, To transform or not to transform: that is the dilemma in the statistical analysis of plant volatiles., Plant Signal. Behav. 6 (2011) 113–116. doi:10.4161/psb.6.1.14191.

[263] C.E. Robertson, J.K. Harris, B.D. Wagner, D. Granger, K. Browne, B. Tatem, L.M. Feazel, K. Park, N.R. Pace, D.N. Frank, Explicet: Graphical user interface software for metadata-driven management, analysis and visualization of microbiome data, Bioinformatics. 29 (2013) 3100–3101. doi:10.1093/bioinformatics/btt526.

[264] D. Beck, J.A. Foster, Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics, PLoS One. 9 (2014). doi:10.1371/journal.pone.0087830.

[265] D. Chiumello, A. Colombo, I. Algieri, C. Mietto, E. Carlesso, F. Crimella, M. Cressoni, M. Quintel, L. Gattinoni, T. Asai, Effect of body mass index in acute respiratory distress syndrome, Br. J. Anaesth. 116 (2016) 113–121. doi:10.1093/bja/aev378.

[266] S. Tanizaki, K. Suzuki, No influence of burn size on ventilator-associated pneumonia in burn patients with inhalation injury., Burns. 38 (2012) 1109–13. doi:10.1016/j.burns.2012.08.008.

[267] E. Hoffmann, O. Dittrich-Breiholz, H. Holtmann, M. Kracht, Multiple control of interleukin-8 gene expression., J. Leukoc. Biol. 72 (2002) 847–55. http://www.ncbi.nlm.nih.gov/pubmed/12429706.

[268] J. Nadigel, S. Audusseau, C.J. Baglole, D.H. Eidelman, Q. Hamid, IL-8 production

in response to cigarette smoke is decreased in epithelial cells from COPD patients., Pulm. Pharmacol. Ther. 26 (2013) 596–602. doi:10.1016/j.pupt.2013.03.002.

[269] K. Gee, C. Guzzo, N.F. Che Mat, W. Ma, A. Kumar, The IL-12 family of cytokines in infection, inflammation and autoimmune disorders., Inflamm. Allergy Drug Targets. 8 (2009) 40–52. http://www.ncbi.nlm.nih.gov/pubmed/19275692.

[270] S. Goriely, M.F. Neurath, M. Goldman, How microorganisms tip the balance between interleukin-12 family members., Nat. Rev. Immunol. 8 (2008) 81–6. doi:10.1038/nri2225.

[271] E. Boon, C.J. Meehan, C. Whidden, D.H.-J. Wong, M.G.I. Langille, R.G. Beiko, Interactions in the microbiome: communities of organisms and communities of genes., FEMS Microbiol. Rev. 38 (2014) 90–118. doi:10.1111/1574-6976.12035.

[272] T.L.F. Leung, R. Poulin, Parasitism , Commensalism , and Mutualism : Exploring the Many Shades of Symbioses, 58 (2008) 107–115.

[273] S. Bikel, A. Valdez-Lara, F. Cornejo-Granados, K. Rico, S. Canizales-Quinteros, X. Soberón, L. Del Pozo-Yauner, A. Ochoa-Leyva, Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: Towards a systems-level understanding of human microbiome, Comput. Struct. Biotechnol. J. 13 (2015) 390–401. doi:10.1016/j.csbj.2015.06.001.

[274] A. Liaw, M. Wiener, M. Andy Liaw, Breiman and Cutler's Random Forests for Classification and Regression, (2015). https://www.stat.berkeley.edu/~breiman/RandomForests/.

[275] S. Abubucker, N. Segata, J. Goll, A.M. Schubert, J. Izard, B.L. Cantarel, B. Rodriguez-Mueller, J. Zucker, M. Thiagarajan, B. Henrissat, O. White, S.T. Kelley, B. Methé, P.D. Schloss, D. Gevers, M. Mitreva, C. Huttenhower, Metabolic reconstruction for metagenomic data and its application to the human microbiome., PLoS Comput. Biol. 8 (2012) e1002358. doi:10.1371/journal.pcbi.1002358.

[276] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, M. Tanabe, KEGG for integration and interpretation of large-scale molecular data sets, Nucleic Acids Res. 40 (2012) 109–114. doi:10.1093/nar/gkr988.

[277] A. Bateman, M.J. Martin, C. O'Donovan, M. Magrane, R. Apweiler, E. Alpi, R. Antunes, J. Arganiska, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, G. Chavali, E. Cibrian-Uhalte, A. Da Silva, M. De Giorgi, T. Dogan, F. Fazzini, P. Gane, L.G. Castro, P. Garmiri, E. Hatton-Ellis, R. Hieta, R. Huntley, D. Legge, W. Liu, J. Luo, A. Macdougall, P. Mutowo, A. Nightingale, S. Orchard, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, S. Rosanoff, R. Saidi, T. Sawford, A. Shypitsyna, E. Turner, V. Volynkin, T. Wardell, X. Watkins, H. Zellner, A.

Cowley, L. Figueira, W. Li, H. McWilliam, R. Lopez, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. Argoud-Puy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M.C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. De Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, F. Jungo, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nouspikel, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A.L. Veuthey, C.H. Wu, C.N. Arighi, L. Arminski, C. Chen, Y. Chen, J.S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D.A. Natale, B.E. Suzek, C.R. Vinayaka, Q. Wang, Y. Wang, L.S. Yeh, M.S. Yerramalla, J. Zhang, UniProt: A hub for protein information, Nucleic Acids Res. 43 (2015) D204–D212. doi:10.1093/nar/gku989.

[278] S.A. Tunio, N.J. Oldfield, S. Bano, D. Ala'Aldeen, K.G. Wooldridge, D.P. Turner, The moonlighting protein glyceraldehyde 3-phosphate dehydrogenase (GapA1) in Neisseria meningitidis is required for optimal association to human cells, 110th ASM Gen. Meet. (2010).

[279] A.J. Ghio, L. a Dailey, J.M. Soukup, J. Stonehuerner, J.H. Richards, R.B. Devlin, Growth of human bronchial epithelial cells at an air-liquid interface alters the response to particle exposure., Part. Fibre Toxicol. 10 (2013) 25. doi:10.1186/1743-8977-10-25.

[280] A.J. Ghio, J.M. Soukup, M. Case, L. a Dailey, J. Richards, J. Berntsen, R.B. Devlin, S. Stone, A. Rappold, Exposure to wood smoke particles produces inflammation in healthy volunteers., Occup. Environ. Med. 69 (2012) 170–5. doi:10.1136/oem.2011.065276.

[281] M.W. Pfaffl, A new mathematical model for relative quantification in real-time RT-PCR., Nucleic Acids Res. 29 (2001) e45. doi:10.1093/nar/29.9.e45.

[282] I. Kruman, Q. Guo, M.P. Mattson, Calcium and reactive oxygen species mediate staurosporine-induced mitochondrial dysfunction and apoptosis in PC12 cells, J. Neurosci. Res. 51 (1998) 293–308. doi:10.1002/(SICI)1097-4547(19980201)51:3<293::AID-JNR3>3.0.CO;2-B.

[283] M.L. Fulcher, S.H. Randell, Epithelial Cell Culture Protocols, in: S.H. Randell, M.L. Fulcher (Eds.), Methods Mol. Biol., Humana Press, Totowa, NJ, 2013: pp. 109–121. doi:10.1007/978-1-62703-125-7.

[284] S.D. McCullough, K.E. Duncan, S.M. Swanton, L. a Dailey, D. Diaz-Sanchez, R.B. Devlin, Ozone induces a proinflammatory response in primary human bronchial epithelial cells through mitogen-activated protein kinase activation without nuclear factor-κB activation., Am. J. Respir. Cell Mol. Biol. 51 (2014)

426–35. doi:10.1165/rcmb.2013-0515OC.

[285] S.H. Randell, Mammalian Cell Cultures: The Example of Airway Epithelial Cell Cultures for Cystic Fibrosis Research, in: P. Langton (Ed.), Essent. Guid. to Read. Biomed. Pap. Recognising Interpret. Best Pract., First, John Wiley & Sons, Ltd., 2013: pp. 49–58.

[286] M.L. Fulcher, S.E. Gabriel, J.C. Olsen, J.R. Tatreau, M. Gentzsch, E. Livanos, M.T. Saavedra, P. Salmon, S.H. Randell, Novel human bronchial epithelial cell lines for cystic fibrosis research, Am. J. Physiol. Lung Cell. Mol. Physiol. 296 (2009) L82–91. doi:10.1152/ajplung.90314.2008.

[287] A.C. Schamberger, N. Mise, J. Jia, E. Genoyer, A.Ö. Yildirim, S. Meiners, O. Eickelberg, Cigarette smoke-induced disruption of bronchial epithelial tight junctions is prevented by transforming growth factor-β, Am. J. Respir. Cell Mol. Biol. 50 (2014) 1040–1052. doi:10.1165/rcmb.2013-0090OC.

[288] A.-S. Jang, The Apical Junctional Complex in Respiratory Diseases., Chonnam Med. J. 50 (2014) 1–5. doi:10.4068/cmj.2014.50.1.1.

[289] T.W. Kensler, N. Wakabayashi, S. Biswal, Cell survival responses to environmental stresses via the Keap1-Nrf2-ARE pathway., Annu. Rev. Pharmacol. Toxicol. 47 (2007) 89–116. doi:10.1146/annurev.pharmtox.46.120604.141046.

[290] X. Zhang, X. Chen, H. Song, H.Z. Chen, B.H. Rovin, Activation of the Nrf2/antioxidant response pathway increases IL-8 expression, Eur. J. Immunol. 35 (2005) 3258–3267. doi:10.1002/eji.200526116.

[291] A. Loboda, A. Stachurska, U. Florczyk, D. Rudnicka, A. Jazwa, J. Wegrzyn, M. Kozakowska, K. Stalinska, L. Poellinger, A.-L. Levonen, S. Yla-Herttuala, A. Jozkowicz, J. Dulak, HIF-1 induction attenuates Nrf2-dependent IL-8 expression in human endothelial cells, Antioxid. Redox Signal. 11 (2009) 1501–1517. doi:10.1089/ARS.2008.2211.

[292] B.-N. Yoon, N.-G. Choi, H.-S. Lee, K.-S. Cho, H.-J. Roh, Induction of interleukin-8 from nasal epithelial cells during bacterial infection: the role of IL-8 for neutrophil recruitment in chronic rhinosinusitis., Mediators Inflamm. 2010 (2010) 813610. doi:10.1155/2010/813610.

[293] Y.-J. Kim, S.-H. Paek, S. Jin, B.S. Park, U.-H. Ha, A novel Pseudomonas aeruginosa-derived effector cooperates with flagella to mediate the upregulation of interleukin 8 in human epithelial cells., Microb. Pathog. 66 (2014) 24–8. doi:10.1016/j.micpath.2013.12.001.

[294] P.G. Czaikoski, D.C. Nascimento, F. Sônego, A. de Freitas, W.M. Turato, M. a de Carvalho, R.S. Santos, G.P. de Oliveira, C. dos Santos Samary, C. Tefe-Silva, J.C.

Alves-Filho, S.H. Ferreira, M.A. Rossi, P.R.M. Rocco, F. Spiller, F.Q. Cunha, Heme oxygenase inhibition enhances neutrophil migration into the bronchoalveolar spaces and improves the outcome of murine pneumonia-induced sepsis., Shock. 39 (2013) 389–96. doi:10.1097/SHK.0b013e31828bbcf9.

[295] D. Moranta, V. Regueiro, C. March, E. Llobet, J. Margareto, E. Larrarte, E. Larrate, J. Garmendia, J. a Bengoechea, Klebsiella pneumoniae capsule polysaccharide impedes the expression of beta-defensins by airway epithelial cells., Infect. Immun. 78 (2010) 1135–46. doi:10.1128/IAI.00940-09.

[296] S. Chillappagari, S. Venkatesan, V. Garapati, P. Mahavadi, A. Munder, A. Seubert, G. Sarode, A. Guenther, B.T. Schmeck, B. Tümmler, M.O. Henke, Impaired TLR4 and HIF expression in cystic fibrosis bronchial epithelial cells downregulates hemeoxygenase-1 and alters iron homeostasis, 49 (2014) 791–799. doi:10.1152/ajplung.00167.2014.

[297] C.-C. Wu, C.-K. Wang, Y.-C. Chen, T.-H. Lin, T.-R. Jinn, C.-T. Lin, IscR Regulation of Capsular Polysaccharide Biosynthesis and Iron-Acquisition Systems in Klebsiella pneumoniae CG43, PLoS One. 9 (2014) e107812. doi:10.1371/journal.pone.0107812.

[298] V. Regueiro, D. Moranta, M. a Campos, J. Margareto, J. Garmendia, J. a Bengoechea, Klebsiella pneumoniae increases the levels of Toll-like receptors 2 and 4 in human airway epithelial cells., Infect. Immun. 77 (2009) 714–24. doi:10.1128/IAI.00852-08.

[299] D.W. Good, T. George, B. a Watts, Toll-like receptor 2 is required for LPS-induced Toll-like receptor 4 signaling and inhibition of ion transport in renal thick ascending limb., J. Biol. Chem. 287 (2012) 20208–20. doi:10.1074/jbc.M111.336255.

[300] S. Greenblum, P.J. Turnbaugh, E. Borenstein, Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease, Proc. Natl. Acad. Sci. 109 (2012) 594–599. doi:10.1073/pnas.1116053109/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1116053109.

[301] M.H. Alhagamhmad, A.S. Day, D.A. Lemberg, S.T. Leach, An overview of the bacterial contribution to Crohn disease pathogenesis., J. Med. Microbiol. (2016). doi:10.1099/jmm.0.000331.

[302] M.B. Azad, A.L. Kozyrskyj, Perinatal programming of asthma: the role of gut microbiota., Clin. Dev. Immunol. 2012 (2012) 932072. doi:10.1155/2012/932072.

[303] H. Daniel, A.M. Gholami, D. Berry, C. Desmarchelier, H. Hahne, G. Loh, S. Mondot, P. Lepage, M. Rothballer, A. Walker, C. Böhm, M. Wenning, M.

Wagner, M. Blaut, P. Schmitt-Kopplin, B. Kuster, D. Haller, T. Clavel, High-fat diet alters gut microbiota physiology in mice., ISME J. (2013) 295–308. doi:10.1038/ismej.2013.155.

[304] a. M. Schubert, H. Sinani, P.D. Schloss, Antibiotic-induced alterations of the murine gut microbiota and subsequent effects on colonization resistance against Clostridium difficile, MBio. 6 (2015) e00974–15–. doi:10.1128/mBio.00974-15.

[305] Z. Kassam, C.H. Lee, Y. Yuan, R.H. Hunt, Fecal microbiota transplantation for Clostridium difficile infection: systematic review and meta-analysis., Am. J. Gastroenterol. 108 (2013) 500–8. doi:10.1038/ajg.2013.59.

[306] S. Angelberger, W. Reinisch, A. Makristathis, C. Lichtenberger, C. Dejaco, P. Papay, G. Novacek, M. Trauner, A. Loy, D. Berry, Temporal Bacterial Community Dynamics Vary Among Ulcerative Colitis Patients After Fecal Microbiota Transplantation, Am. J. Gastroenterol. 108 (2013) 1620–1630. doi:10.1038/ajg.2013.257.

[307] K.K. Barfod, M. Roggenbuck, L.H. Hansen, S. Schjørring, S.T. Larsen, S.J. Sørensen, K.A. Krogfelt, The murine lung microbiome in relation to the intestinal and vaginal bacterial communities., BMC Microbiol. 13 (2013) 303. doi:10.1186/1471-2180-13-303.