Quantitative Inferences from the Lung Microbiome

by

Laura Tipton

BA, University of Virginia, 2007

MS, George Washington University, 2011

Submitted to the Graduate Faculty of

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Laura Tipton

It was defended on

December 14, 2016

and approved by

Penayiotis Benos, PhD, Department of Computational & Systems Biology

Alison Morris, MD, MS, Department of Immunology

Kyle Bibby, PhD, PE, Swanson School of Engineering, Civil and Environmental Engineering

Kathryn Roeder, PhD, CMU Department of Statistics

Dissertation Advisor: Elodie Ghedin, PhD, NYU Department of Biology

QUANTITATIVE INFERENCE FROM THE LUNG MICROBIOME

Laura Tipton, PhD

University of Pittsburgh, 2016

Within the last decade, we have progressed from the belief that the healthy human lung is a sterile environment to attempts to study inter-kingdom interactions between microbial residents of the lungs. It has been repeatedly confirmed that the lungs contain both bacteria, predominantly from the *Streptococcus*, *Veillonella*, and *Prevotella* genera, and fungi, predominantly from the *Cladosporium*, *Eurotium*, *Penicillium*, and *Aspergillus* genera. The community composition as a whole undergoes shifts in every lung disease and condition that has been studied, including asthma, chronic obstructive pulmonary disorder, and cystic fibrosis. The studies that have observed these shifts have largely been descriptive, comparing the taxonomies present in healthy lungs to taxonomies in diseased lungs. Here we investigated the lung microbiome and relationships within the microbial community and between microbes and the host in a more quantitative and inferential manner. First, we introduced the lasso-penalized generalized linear mixed model (LassoGLMM) for microbiomes. LassoGLMM was applied to a short time-course study of the human oral bacterial microbiome with standard blood chemical measurements and to repeated measurements of the human lung bacterial microbiome and fungal mycobiome with local and systemic markers of inflammation. We sought to show that increased inflammation and other continuous clinical variables in human hosts are associated with distinct microbes present in the lung or oral microbiomes. Then, we examined cross-domain interactions between bacteria and fungi. Ecological interaction networks were inferred for the human lung and skin micro- and myco-biomes. Networks limited to a single domain of life were compared with those that

include both bacteria and fungi to identify important components of the microbial community that would be overlooked in a single domain study. Finally, we explored the metabolism of the bacteria within the human lung using three different "-omics" datasets: taxonomic assignments from 16S rRNA gene sequences, gene families from metatranscriptomic sequences, and mass-to-charge ratio (*m/z*) features from metabolomics. Correlations were examined between pairs of datasets and all three datasets were integrated to identify bacteria contributing metabolic processes that may have otherwise gone unnoticed, resulting in the first complete characterization of the metabolism of the human lung bacterial microbiome.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to start by thanking my advisor and mentor, Elodie Ghedin, for being a continuous source of wisdom and *pain au chocolat*, without which I would not have made it this far.

I would also like to thank each of my committee members, Takis Benos, Kyle Bibby, Alison Morris, and Kathryn Roeder, for their support and encouragement since I first asked each one to be on my committee. I am honored that each of you have taken the time to help me improve my work.

Although not members of my committee, I owe a great debt of gratitude to Rich Bonneau and Karen T. Cuenco, for guiding sections of my research in ever more interesting directions.

To all of the members of the Ghedin lab, past, present, and honorary, thank you for your support and comradery. I promise the chocolate velvet cake recipe protocol is in the works.

To my parents, Timothy and Sara Moore, and my sister Heather, who always encouraged me to be “the smart one”.

I am forever thankful for my husband Chris, without whom I would have been less well well-fed, well-tempered, and well-dressed.

Finally, I’d like to thank my own microbiome for keeping me healthy despite the stress I have caused it over the course of grad school.

1.0 INTRODUCTION

As far back as the 1880's, healthy human intestines were known to be home to a multitude of bacteria (1); however, it is only in the past few years that scientists have recognized that the healthy lung also harbors bacteria (2, 3). Part of this discrepancy may be due to the fact that, until recently, only culturable bacteria could be studied. The rapid rise of next-generation sequencing (NGS) has enabled the recognition of unculturable bacteria, fungi, and other microbes in the lung and in other habitats.

Although the existence of a distinct lung microbiome has been confirmed (4), it is still analyzed primarily using descriptive statistics (5, 6). This work focuses on investigating the lung microbiome and its relationship with the host in a quantitative and inferential manner. Other host-associated microbiomes, including in the human oral cavity, on the human skin, and in the macaque lung, were included to further validate the methods developed and used throughout. The results significantly enhance both our knowledge of the lung microbiome and the methodology available to analyze other host-associated microbiomes.

1.1 BACTERIA IN THE LUNGS

Studies of the human-associated microbiome came on the heels of the human genome project and all its technological advancements. It allowed the characterization of this 'second genome' suspected of contributing to health and normal physiology. More recently, explorations of the respiratory tract have demonstrated the presence of bacteria and other micro-organisms in healthy lungs, including members of the Firmicutes, Bacteroidetes, and Proteobacteria phyla (2, 3, 7). Each of these phyla is commonly found in other human-associated bacterial microbiomes (8, 9). Prominent genera from these phyla include *Streptococcus* and *Veillonella* from the Firmicutes phylum, and *Prevotella* from the Bacteroidetes phylum (**Figure 1.1**) (2, 4, 10–12). None of these genera are unique to the lung microbiome; what is likely to be specific to the lung environment are particular species or strains of microbes. Alternatively, some of these genera and species may have translocated to the lungs from the gut or other microbiome, a process that is known to occur under compromised immune conditions such as HIV-infection (13, 14). However, most studies rely on target gene sequencing of the 16S rRNA, which is an approach not considered to be reliable for taxonomic assignments below the genus level.

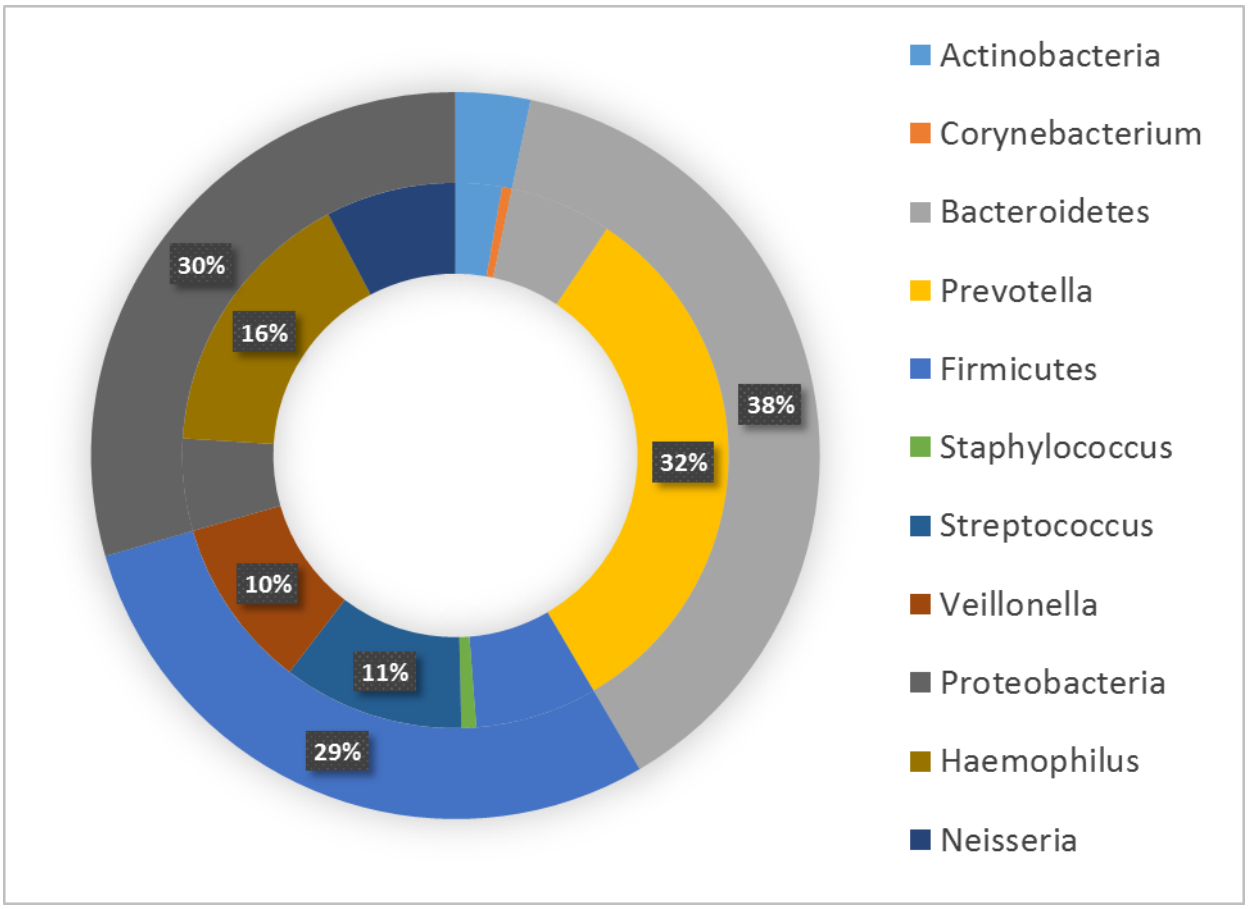

Figure 1.1 Relative abundance of bacterial genera and phyla in the lung brushings of healthy and asthmatic individuals. The inner ring displays the genera of bacteria while the outer ring displays the phyla. Any genera or phyla that represent over 10% of the reads are labeled with the percentage of reads. Figure adapted from (2).

Because the genera prevalent in the lung microbiota are also highly abundant in the human oral microbiota, care has been taken to repeatedly prove that the bacteria within the lungs form their own community. Proof has come from both models and observations. From the modeling side, investigators have applied an ecological community assembly model for the neutral model of biodiversity. This model assumes that all inhabitable locations are the same and that all species have an equal chance of survival upon arrival in a given location (15). The

abundance of each species in a given community is dependent only on the total size of the local community and the immigration rate from the source community. The expected abundance, *N*, of the i^{th} species is $E(N_i) = N_T P_i$, where N_T is the local community size and P_i is the relative abundance of species *i* in the source community (16). When this model was applied to the lung microbiome, using the oral microbiome as the source community, individual bacterial species were more abundant than would be expected based solely on immigration from the mouth (4, 17). Both the details of how the neutral model was applied and the population studied impacted the results. A recent study found that the genera *Ralstonia* and *Bosea* were more prevalent in the lungs than expected from the oral wash source community of healthy non-smokers (4). In another study, the genera *Catonella* and *Selenomonas* were found to be more prevalent than expected from an oral wash source community in healthy patients (17). Both studies concluded that there are bacterial genera present in the lung microbiome that are not simply neutral immigrants from the mouth. Other studies have used ordination methods—an approach that plots the multi-dimensional community in a 2- or 3-dimensional space—to observe distinct community compositions of the mouth and lung microbiotas, displaying a separation between the oral and lung communities in the ordination plots (18, 19). These ordination plots have been used to show that the bacterial community found in the lung samples could not have originated solely as carry-over or contamination from the oral cavity.

1.1.1 Bacteria During Disease

Several human diseases have been associated with shifts in the composition of the bacteria in the lungs. Most of the conditions studied have been respiratory diseases, including cystic fibrosis (CF) (20, 21), asthma (2, 7), and chronic obstructive pulmonary disease (COPD) (3, 22–24), or led to lung transplantation (11, 25). These conditions have mostly unknown or unclear etiology, but it was hypothesized that the microbiome may play an important role. While some diseases studied have a clearer link to the bacteria present in the lungs, including active *Mycobacterium tuberculosis* (the causative agent of tuberculosis) infection (26), others, including HIV, have an indirect link to the lung microbiota (14, 27). In the case of HIV, an association is suspected between the microbiota and subtle lung immune deficits seen even in well-controlled HIV infection.

Each disease studied has its own unique shifts in the composition of bacteria present. Cystic fibrosis patients have decreased community diversity in their sputum (28, 29) while asthmatic patients have increased community diversity in their lower respiratory tract (7, 30). Specifically, asthma has been associated with increased abundance of members of the *Proteobacteria* phylum (2, 7, 30). COPD may be unique among the respiratory diseases as shifts in the microbiota are only seen when the disease is severe (3, 22, 31, 32). In severe COPD there is decreased abundance of members of the *Bacteroidetes* phylum, and accompanying increases in potentially pathogenic members of the *Proteobacteria*, including members of the *Pseudomonas* and *Haemophilus* genera (3, 33, 34). Similarly, bronchoalveolar lavages (BALs) from lung transplant patients have been shown to be enriched with *Pseudomonas* and other members of the *Proteobacteria* phylum (11, 25, 35, 36). However, each disease appears to be associated with

lung bacterial communities that are different, driven by the variety of genera in the *Proteobacteria* phylum and diversity of species and strains within the *Pseudomonas* genera (**Figure 1.2**) (12).

One commonality across all diseases studied is that the direction of causality remains unknown. Because most studies are cross-sectional, or represent a single point in time, investigators are unable to determine if the shifts in the disease are the result or the cause of a shifting microbiota. This directionality will become more clear as microbiome studies incorporate other technologies to study the metabolism and mechanisms of the community, and through prospective, longitudinal studies that follow patients from early disease onset through clinical exacerbations.

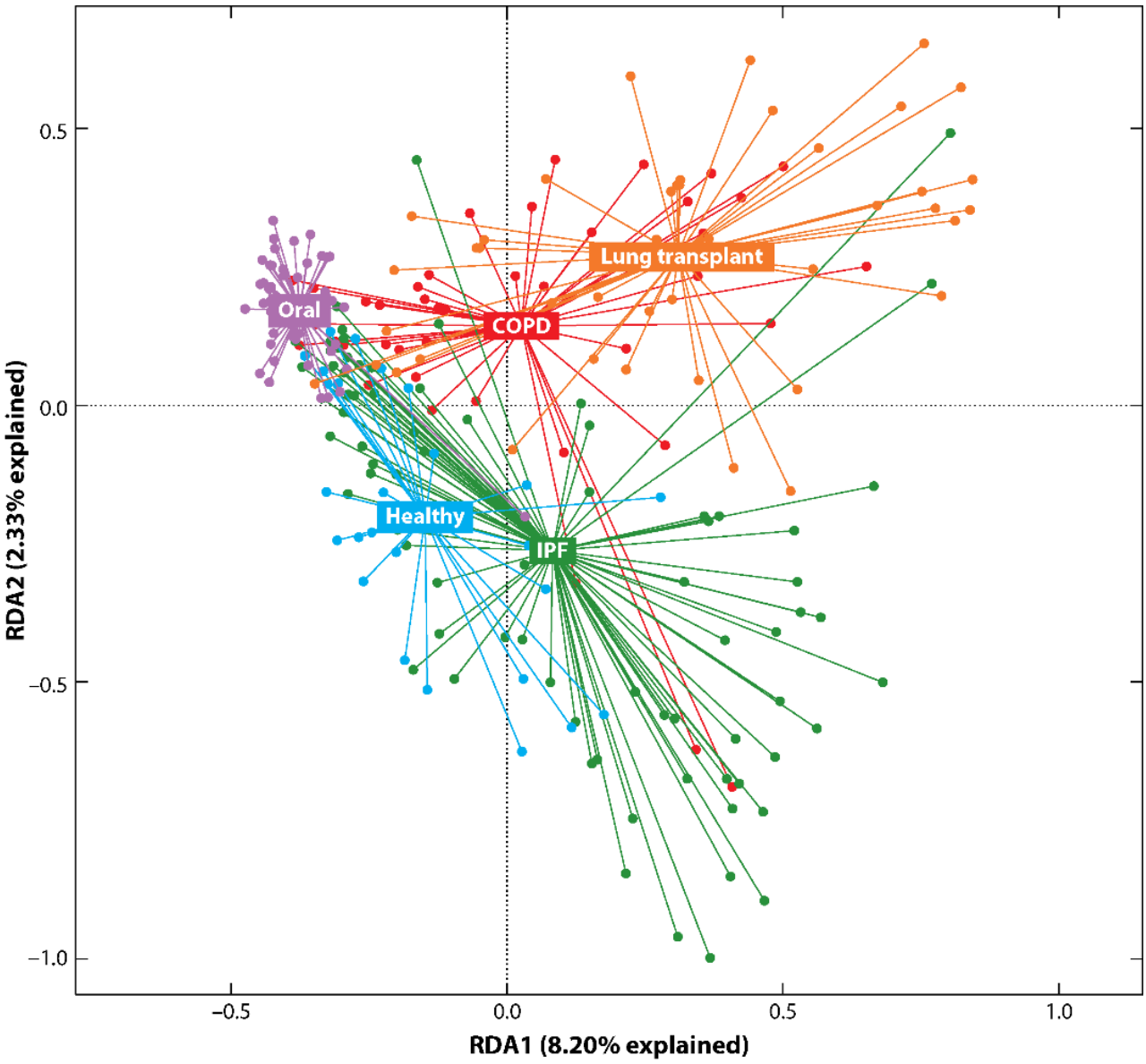

Figure 1.2 Ordination plot of bacterial communities from the lungs of patients with different respiratory diseases as labeled. When plotted together, each disease separates from healthy lungs in its own way. COPD = chronic obstructive pulmonary disease; IPF = idiopathic pulmonary fibrosis. Figure from (12), Copyright 2016, Annual Reviews.

1.2 FUNGI IN THE LUNGS[1]

In less than a decade, we have progressed from believing that healthy lungs are a sterile environment to studying inter-kingdom interactions between microbial residents of the lung. In part due to the debate about the sterility of the lungs, next generation sequencing (NGS)-based studies of the lung microbiome have lagged behind those of the gut microbiome, with the first studies of the lung microbiome being published in 2010 and 2011 (2, 3, 7). These early NGS studies, and many studies since, focused exclusively on the bacteria present in the lungs under health and disease. However, the microbial community that inhabits the lungs also contains viruses, fungi, and other eukaryotes.

1.2.1 Why is the lung mycobiome important?

In addition to causing clinical infections, the lung mycobiome may have profound inflammatory effects that can cause or worsen lung disease. Similar to bacterial pathogens, fungi contain pathogen-associated molecular patterns (PAMPS) such as glucans, chitin, and mannans present in the fungal cell wall (37, 38). These PAMPs are recognized by pathogen recognition receptors (PRRs) that then activate immune cells leading to inflammation (**Figure 1.3**). Activation of macrophages, T cells, and B cells leads to cytokine release and immune activation. Both the adaptive and innate immune responses are triggered by fungi, and the respiratory epithelium plays a key role in the response to fungi. Fungi have been linked to such chronic lung diseases as

[1] This work was published in *Virulence* as "The lung mycobiome in the next-generation sequencing era" (169)

asthma and COPD (39, 40). Given the ubiquity of fungi in the environment, the potential respiratory exposure to fungi, and the ability of fungi to trigger inflammation, the mycobiome may play a key role in shaping the respiratory immune response and contribute to lung damage.

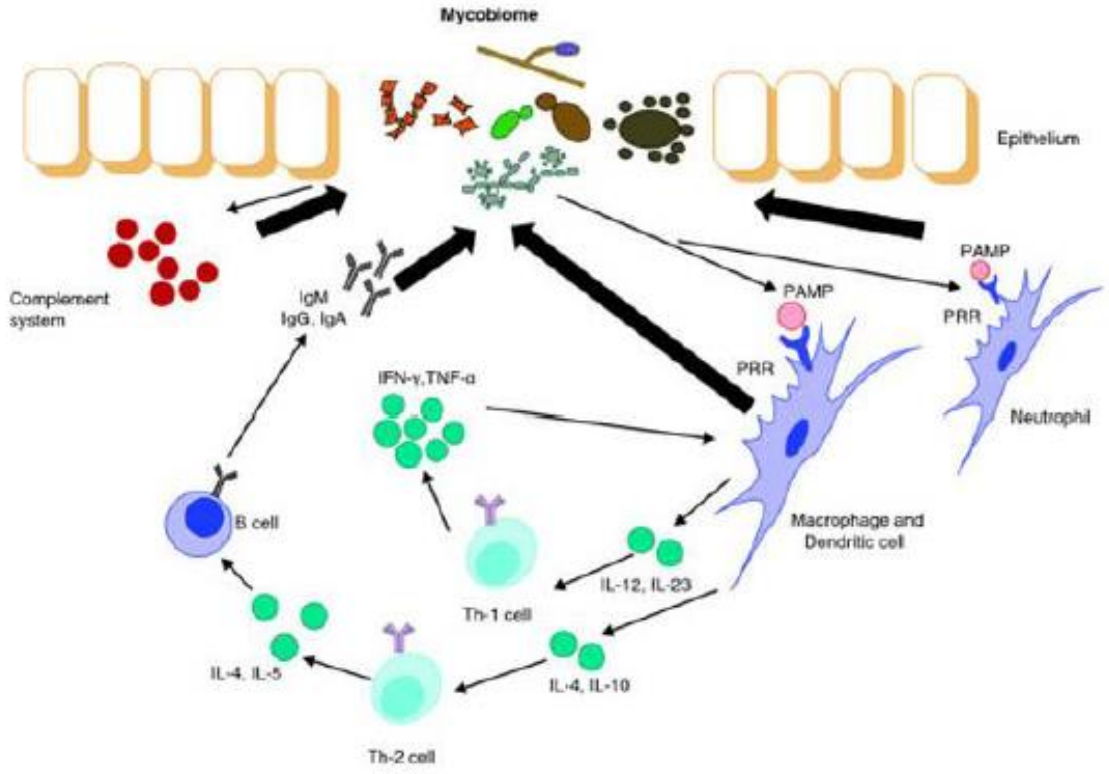

Figure 1.3 Interaction between the mycobiome and the immune system. When pattern recognition receptors (PRRs) recognize the pathogen-associated molecular patterns (PAMPs) on fungal cell walls, macrophages, T cells, and B cells are activated. The fate of the actived T cells is determined by the cytokines that are stimulated. INF, interferon; IL, interleukin; TNF, tumor necrosis factor. Figure from ref. (41)

1.2.2 What do we know about the lung mycobiome?

To date, there have been fewer than 10 NGS lung mycobiome papers published. Despite this low number, several themes emerge from the literature: (1) fungi are present in the human respiratory tract, even during health; (2) the fungi present in the respiratory tract are highly variable between individuals; and (3) many diseases are accompanied by decreased diversity of fungi in the lungs.

Fungi found in the human respiratory tract cover a range of phylogenies, but are predominantly from the Dikarya sub-kingdom, which is composed of the phyla Ascomycota and Basidiomycota. In fact, the most common taxa identified in healthy lung samples were the family Davidiellaceae, and the genera *Cladosporium*, *Eurotium*, *Penicillium*, and *Aspergillus* (25). Other genera found in healthy individuals include *Candida*, *Neosartorya*, *Malassezia*, *Hyphodontia*, *Kluyveromyces*, and *Pneumocystis* (20) (**Figure 1.4**). These eleven taxa cover the range of fungal growth patterns from filamentous, to yeast and yeast-like. Many of these genera, including *Cladosporium*, *Aspergillus*, *Candida*, *Malassezia*, and *Pneumocystis*, contain species that are either pathogenic to humans or cause allergic reactions (42–46). On the other end of the spectrum, the genera *Penicillium* includes producers of the antibiotic penicillin (47).

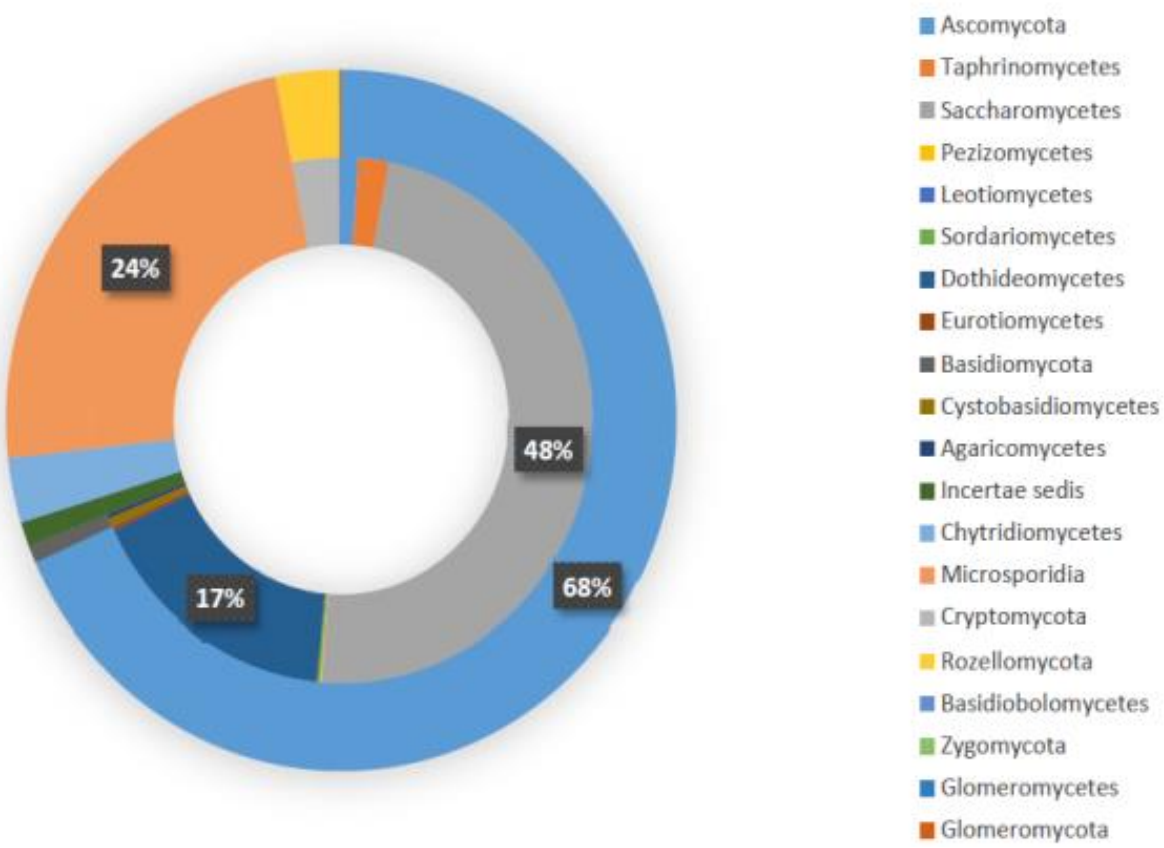

Figure 1.4 Distribution of fungal phyla in the sputum of healthy individuals(38). The inner ring displays the class of fungi while the outer ring displays the phyla. Any class or phyla that represented over 10% of the reads is labeled with its percentage of reads is labeled with its percentage of reads and classes or phyla below 0.1% are not represented.

The fungi present in the respiratory tract are also highly variable between individuals. Even in patients with the same disease, different patients have been shown to harbor distinct fungal communities (48). In our experience, the number of "private species", those present in only one individual, can be greater than the number of species shared across samples. Whether this difference is due to mis-identification of the fungi (perhaps due to a sequencing error) or a

patient's unique environmental exposures has yet to be determined. It has been proposed that the macromycetes (or macroscopic fungi commonly known as mushrooms) observed in a subset of samples represent the outdoor environment that a patient is exposed to as they often contain wood-inhabiting fungi and cereal grain pathogens (20). Even the level of fungal diversity in the lungs is highly variable between individuals. Compared to bacterial diversity in the lungs, average fungal diversity in the same samples is consistently lower (49) but has a higher coefficient of variation, or ratio of standard deviation to the mean. As an example, in a subset of 35 BAL samples from our study of the lung mycobiome in HIV-infected and uninfected individuals for which we have both 16S rRNA and ITS sequence data (40), the coefficient of variation is 22.9% for bacterial diversity, as measured by the Shannon diversity index, and is 73.9% for fungal diversity. Other factors, including patient health and environmental exposures, appear to have a greater impact on the diversity of the fungi than of the bacteria, the latter being considered relatively stable.

Only a limited number of diseases have been examined for their impact on, or association with, the lung mycobiome. Most diseases that have been studied, including CF, asthma, and COPD, as well as lung transplant, have been associated with decreases in fungal diversity (20, 25, 40, 50). Across these conditions, lower fungal diversity is correlated with lower respiratory function. The reduced diversity may be caused by an overgrowth of a single fungal species, or by the loss of rare species that comes with a reduction in overall fungal abundance.

CF has received the most attention with studies that range from correlating community characteristics with patient health indicators to comparing NGS and sequencing detection to community stability over time. Delhaes et al examined sputum of four CF patients, each sampled twice, and found that both bacterial and fungal community richness was positively correlated

with indicators of health and lung function (20), i.e. more fungal species were seen in the patients with the lowest disease severity scores, highest body-mass indices, highest forced vital capacity, and highest forced expiratory volume in 1 second. Harrison et al found that over 82% of the species identified by sequencing were not found by culture-based methods, which detected fungi in only 27% of the sputum samples from 55 CF patients compared to a 90% detection rate by sequencing (50). Willger and colleagues sought to compare sputum from six CF patients before and after antimicrobial therapy and found that the fungal communities were relatively stable (51). Similarly, a study of 89 sputum samples from 28 CF patients showed that the fungal communities were stable through clinical exacerbation and treatment (48). This study combined NGS of the mycobiome with phenotypic and genotypic analysis of *Candida* isolates from the samples to identify mutations leading to the filamentous phenotype in the presence of filamentation repressive cues from the bacteria *Pseudomonas aeruginosa* (48). It is the filamentous phenotype that is considered pathogenic and evading the repressive signals from other members of the microbiome could lead to *Candida* infection.

Lung transplantation could impact the mycobiome due to the immunosuppression and antibiotics received by recipients as well as structural changes in the lung. In general, lung transplant recipients have reduced fungal richness and increased fungal abundance compared to healthy controls. For example, Charlson et al found that combined bacterial and fungal community richness was reduced in BALs from 21 lung transplant patients compared to healthy controls and richness was lowest in patients who had a transplant due to CF (25). All transplant patients were receiving antibiotics in addition to immunosuppression at the time of sampling, making it difficult to attribute causality in these changes. In the four patients with high fungal amplification from BAL, the dominant species (*Candida albicans* in three samples and

Aspergillus fumigatus in one sample) was also found by culture methods, which were only able to identify four species: *C. albicans, A. fumigatus, Aspergillus flavus,* and *Paecilomyces lilacinus* (also known as *Paecilomyces lilacinus*). Expanding this dataset to include a total of 149 BAL samples from healthy subjects, HIV-infected subjects, subjects with mixed lung disease, and lung transplant recipients, Bittinger et al showed that fungal abundance increases from healthy subjects to lung transplant recipients with HIV-infected subjects and subjects with mixed lung disease falling in the middle (49). To ensure that they were counting species truly present in the lungs, they used DNA quantification to filter out any species that were likely to have come from contamination before calculating species abundances.

Asthma, COPD, and pneumonia have been less well-studied, with only a single paper each examining shifts in lung mycobiome communities. For asthma, a case-control study to compare induced sputum samples of 30 subjects with asthma to that of 13 control subjects found 90 species to be more abundant in asthma and 46 species to be more abundant in the controls (39). Species with more than a 5% increase in abundance between the asthma and control sample pools were *Psathyrella candolleana, Malassezia pachydermatis,* and *Termitomyces clypeatus*, none of which were seen in the control sample pool. Species with more than a 5% decrease in abundance between the asthma and control sample pools were *Eremothecium sinecaudum, Systenostrema alba, Cladosporium clasdosporioides,* and *Vanderwaltozyma polyspora*. We published the only paper on COPD where we first compared HIV-infected to HIV-uninfected individuals and then compared HIV-infected individuals with COPD to HIV-infected with normal lung function (40). We used an overlap of multiple methods to identify overrepresented species in the BAL of 32 HIV-infected individuals, 10 with and 22 without COPD, and 24 HIV-uninfected controls (40). We found *Pneumocystis jirovecii* to be the most distinguishing species

as it was overrepresented in both HIV and COPD. Finally, in the only published study on pneumonia, which is the largest lung mycobiome study to date, Krause et al compared BALs from 87 healthy controls, 18 patients with extrapulmonary infection on antibiotics, 8 intensive care unit patients without antibiotics, 23 intensive care unit patients with extrapulmonary infection on antibiotics, 34 intensive care unit patients with pneumonia on antibiotics, and 32 patients with candidemia (52). They focused on *Candida* and found that intensive care unit admission, but not antibiotic therapy, shifted the lung mycobiome to be dominated by *Candida*. Even this recent study still used culture-based fungal identification as the gold standard for fungal identification, as this is standard practice in a clinical setting.

1.3 OTHER MICROBES IN THE LUNGS

While a wide variety of viruses have been identified in healthy lungs, most are bacteriophages, viruses that infect bacteria. Across individuals, there appears to be a core functionality of this virome of the lung (53). However, if the lung micro- and myco-biomes are considered new fields, the lung virome is truly nascent. While the virome encompasses both DNA and RNA viruses, the studies that have been published to date examine only the DNA viruses and focus on the viruses that infect the human hosts (53, 54).

In addition to viruses, other non-bacterial, non-fungal microbes in the lungs consist of other eukaryotes. These include protists and helminths, both of which have been known to infect

the human lung (55). However, neither of these groups has been studied in the healthy human lung. Therefore, neither these eukaryotes nor viruses will be included in this work.

1.4 CHALLENGES TO STUDYING THE LUNG MICROBIOME[2]

Studies of the lung microbiome and mycobiome may be limited because of the numerous challenges that exist at every step. The challenges begin with sampling the lung and continue through sample processing. These are followed by tough choices with regards to amplification and sequencing and more challenges to process the sequencing data. Finally, the historical system of naming fungi that resulted in multiple names for a single species has created difficulties now that NGS is used to define and identify species. Because many of these challenges are applicable to all NGS microbiome and mycobiome studies, we have included only a brief overview of each one and its relevance to the lung communities.

The human lungs are difficult to access. The two most common means of sampling the lungs are induced sputum (IS) and BAL. Both methods run the risk of contamination from the upper respiratory tract. IS is obtained by having subjects cough after inhalation of hypertonic solution, potentially introducing mouth microbes during collection, and the bronchoscope may introduce upper respiratory microbes to the lungs during passage through the nose or mouth. However, it has been shown that both IS and BAL mycobiomes are distinct from the oral

[2] This work was published in *Virulence* as "The lung mycobiome in the next-generation sequencing era" (169)

mycobiome (40). We have shown that there are differences in the fungal communities of IS and BAL, likely because the two methods sample the lungs differently: IS samples from a greater anatomic region of the lung, while BAL samples from a subset of the alveoli. Different environmental conditions existing in different portions of the lungs, or microenvironments, will be indistinguishable in an IS sample, but may be missed entirely by a BAL. The choice of sampling method should be selected based on the question under investigation, or, in the case of pre-existing samples, the limits of the sampling method should be addressed to the extent possible.

Once a sample is obtained, DNA needs to be extracted. As with any NGS-based study of microbes, one of the first steps is to break open the cell. While this is relatively simple for bacteria, the fungal cell wall is composed of a combination of glucans and chitin, for which proportions vary by fungal growth patterns (56). The varying composition of the fungal cell wall leads to a range of tensile strengths, and there are a number of methods to break open the cell wall that vary in harshness. For the purposes of extracting DNA from both yeasts and filamentous fungi, mechanical disintegration has proven most effective (57); however, this method runs the risk of shearing the DNA and therefore must be carefully calibrated for the given sample composition.

The harsh mechanical treatment to break open the fungal cell walls also creates a challenge by releasing DNA from other cells present in the lung sample, both bacterial and human. The extra DNA released from non-target cells, along with any DNA found in the laboratory reagents (a recent study attempted to characterize the bacteria found in DNA extraction kits (58), but no equivalent study has been performed for fungi), necessitates careful primer design for amplification. Common targets for amplification include the gene encoding

one or more of the hypervariable regions of the 16S ribosomal RNA (rRNA) for bacteria (59), and the 18S rRNA or the internal transcribed spacer (ITS) region(s) located between the 18S and 26S rRNA genes (60) for fungi. Each of the fungal targets has its own benefits and drawbacks. Specifically, the 18S rRNA gene is conserved across all eukaryotes, so targeting this gene for amplification and sequencing of fungi will include non-fungal microbial eukaryotes. Because the 18S rRNA gene is conserved across all eukaryotes, amplifying this region of the genome can also amplify any human DNA present in the sample, depending on the specificity of the primers. Due to the low biomass of microbes in the lung, the amount of human DNA in the sample prior to targeted amplification is bound to be higher than the amount of fungal DNA. In contrast, the ITS region is more diverse across eukaryotes and primers have been designed specifically for the amplification of fungal DNA (61), to the exclusion of all other eukaryotes. Some of these primers are narrowly targeted such that they introduce bias towards particular fungal phyla, another issue worthy of careful consideration. The diversity of the ITS region and the specificity of the primers combine to allow a greater depth of taxonomic assignment, often down to the species level. It is this advantage that has led the ITS region to be the "official primary barcoding marker" for fungi (61). However, because it is a non-coding region, ITS sequences cannot be used to determine phylogenetic relationships between unidentified fungi.

One of the greatest challenges following sequencing is a lack of data, specifically a lack of reference genomes. For bacteria, the GreenGenes 16S rRNA gene database is a common reference database (62, 63), but it is missing references that may be clinically relevant to respiratory disease such as the *Trophyrema* genus, which has been shown to be enriched in the lungs of HIV-infected patients (14). For fungi, the UNITE database of fungal ITS sequences represents the largest collection of fungal sequences and (as of version 7) contains more than

64,500 “species hypotheses” at the 1% similarity threshold, where a species hypothesis is any group of sequences that are no more distant than the similarity threshold (64). Compared to the over 203,000 bacterial species hypotheses at the 1% similarity threshold in the May, 2013 release of the GreenGenes 16S rRNA gene database (63), the number of fungal reference species seems small. The sequences within the UNITE database are heavily biased (87% of species hypotheses) towards the Dikarya sub-kingdom (64) (**Figure 1.5**). While this bias may accurately represent the distribution of fungal species, or may simply stem from UNITE’s history as a database for plant root fungi (65), it certainly explains why the majority of species identified in the lungs belong to this sub-kingdom.

Even after the sequences are identified, there are still challenges to be overcome in mycobiome studies that are not as prevalent in microbiome studies. Despite years of expert mycologists pushing for each fungus to have a single species name, many fungi still have one name for their sexual reproductive stage (or teleomorph) and one name for their asexual reproductive stage (or anamorph) (66, 67). The problem with this dual naming system in the NGS era is two-fold. First, it can complicate a search for knowledge prior to the NGS era. Many older studies reference only one name and it can be unclear if the results apply to the opposite morph. There is no way to identify which morph is present in a sample based on its DNA. It can also be that the higher order taxonomic assignments, such as family and order, of the two morphs are different, leading to phylogenetic confusion about the placement of the species as a whole. Uncertain phylogenetic placement and phylogenetic restructuring result in taxonomic hierarchies that include *incertae sedis* (Latin for “of uncertain placement”), as seen in the taxonomies of members of the former phyla Zygomycota (68). The second problem with a dual naming system is that sample sequences may have two or more identical matches when a database has reference

sequences for both the teleomorph and anamorph causing ambiguous assignments. Curated databases such as Mycobank (69) can aid in the reduction of duplicate reference sequences, but similar curation is not readily available for pre-NGS knowledge.

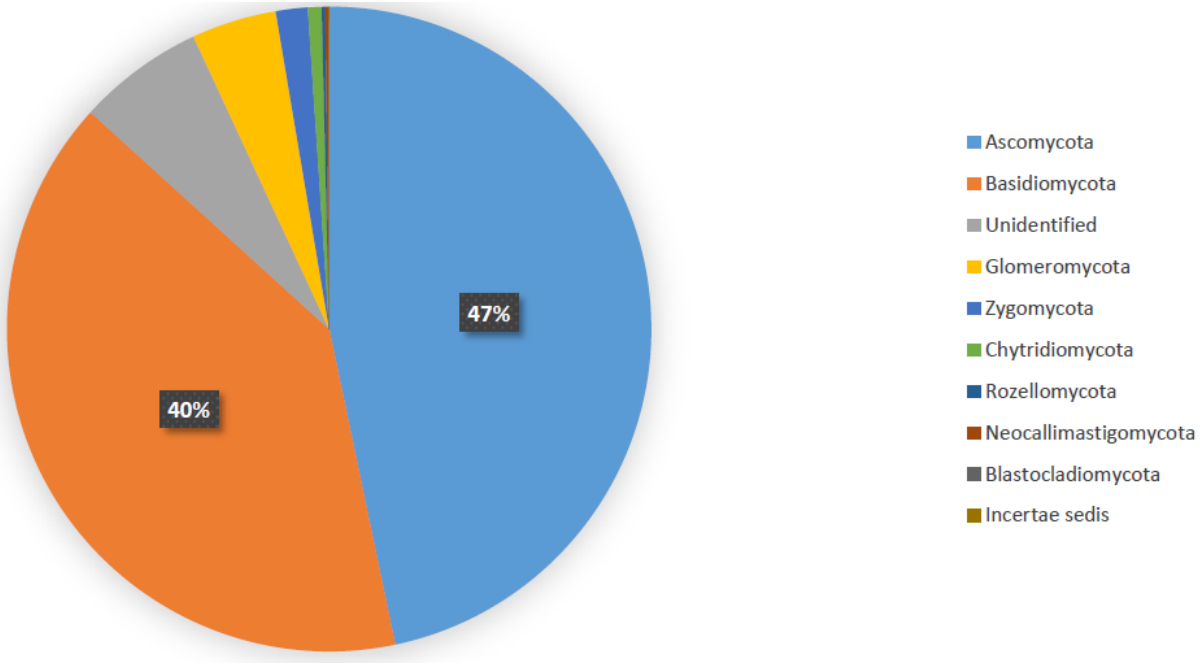

Figure 1.5 Distribution of fungal phyla in the UNITE database (64). The chart shows the breakdown of phyla of the 64,500 "species hypotheses" at the 1% similarity threshold found in the UNITE ITS database. Phyla that represent over 10% of the species hypotheses are labeled with its percentage of species hypotheses.

1.5 FUTURE OF LUNG MICROBIOME RESEARCH

Many of the challenges to studying the lung microbiome and mycobiome are unavoidable. There is likely never going to be easy access to the human lung that avoids the upper respiratory tract, and the microbes in the lung will always be low in biomass. However, improvement is possible in primer design, reference databases, and analytic methods. Going forward, many of the advances made in the study of the lung bacteria will aid in the study of the lung mycobiome. Once the sequencing data are collapsed into a "biom" file or taxa table (a table that displays the abundance of each taxonomic group for every sample) it makes little difference if the taxa are bacterial or fungal. All of the statistical methods to handle the abnormal distributions (70) and complex study designs associated with bacterial studies can be used on fungal studies with little or no modifications. Similarly, as bacterial studies shift from cross-sectional to longitudinal, so too should fungal studies. There have already been studies into the daily changes in bacterial communities that occur during CF and its exacerbations (21) but fungi were not examined. The tools, including sequencing capacity, that are developed to handle daily sampling of the bacteria can also be put to use to analyze the fungi present during the same time period.

When more studies include both bacterial and fungal amplicon sequences from target gene sequencing, we can begin to look at cross-kingdom interactions. Interactions between bacteria and fungi are important among oral microbiota (71) and identified as an emerging field across biology (72), so they will no doubt be important to the study of the lung microbiota. Looking farther into the future, as amplicon sequencing gives way to whole metagenome and whole metatranscriptome sequencing, these delineations between bacterial and fungal

communities will fall away. Both kingdoms will be sequenced simultaneously and their members' abundance and transcriptional activity, relative to each other, will be apparent.

Another avenue for future investigation will be the mechanisms of interactions between the microbiome, the mycobiome, and the host. As a part of the mucosal immune system, the lungs and the microbes within, play an important role in human health and disease. The impact of inflammation on the development of many lung diseases represents an area of active investigation, one in which the contribution from the lung microbiome or mycobiome could prove crucial to understanding.

In both the short- and long-term, the critical need for the lung microbiome and mycobiome is more data, in the forms of reference sequences and additional studies. Adding sequences to reference databases by sequencing more bacteria and fungi will help in identifying species that are currently unclassifiable. These sequences can come from culturing some of the estimated 99% of the world's bacteria (73, 74) and fungi (75) that have yet to be reliably grown in the lab, or from assembling genomes present in deeply sequenced metagenomes. The latter makes it possible to obtain sequences from unculturable microbes without the time and manpower required to optimize the culturing conditions of newly cultured organisms. The other, and perhaps more important, way to contribute to the knowledge of the lung microbiome and mycobiome is to perform more studies. Additional lung microbiome and mycobiome studies will provide more information about the changes in the bacteria and fungi present under health and disease conditions and will help to explain the role of microbes in influencing the respiratory immune response.

1.6 OVERVIEW

The introduction highlighted the importance of including longitudinal data, taking into account cross-kingdom interactions, and developing new analytical methods in our exploration of the lung microbiome. This thesis is presented as three separate sections that touch on each of those aspects. In the first part, the lasso-penalized generalized linear mixed model (LassoGLMM) for microbiomes is introduced. LassoGLMM is applied to a short time-course study of the human oral bacterial microbiome with standard blood chemical measurements. LassoGLMM is then applied to repeated measurements of the human lung bacterial microbiome and fungal mycobiome with local and systemic markers of inflammation.

In the second part, cross-domain interactions between bacteria and fungi are examined. Ecological interaction networks are inferred for the macaque lung, human lung, and human skin micro- and mycobiomes. In the human lung and human skin studies, networks limited to a single domain of life are compared with those that include both bacteria and fungi.

Finally, in the third section, the metabolism of the bacteria within the human lung is explored using three different "-omics" datasets. Each dataset—taxonomic assignments from 16S rRNA gene sequences, gene families from metatranscriptomic sequences, and mass-to-charge ratio (*m/z*) features from metabolomics—is explored for its associations with COPD and HIV. Then, correlations are examined between pairs of datasets and finally, all three datasets are

integrated to identify bacteria contributing the metabolic processes that may have otherwise gone unnoticed.

2.0 MEASURING ASSOCIATIONS BETWEEN THE MICROBIOTA AND REPEATED MEASURES OF CONTINUOUS CLINICAL VARIABLES USING A LASSO-PENALIZED GENERALIZED LINEAR MIXED MODEL[3]

2.1 BACKGROUND

Epidemiologic studies, ranging from clinical trials to observational studies, often include the collection of demographics, disease symptoms, treatment, diagnostic tests, and clinical laboratory information. Recent evidence that the human microbiome influences disease occurrence (31, 76) has led to interest in how the microbiome may more generally impact clinical and treatment outcomes, and the natural history of a disease. While continuous clinical measures are used to describe and to identify risk subgroups in the patient population, the relationship between these measures and the microbiome is rarely analyzed. This rarity is in part caused by methodologic limitations in applying current microbiome and analytic techniques to continuous clinical data.

[3] Paper under review.

One stumbling block to analyzing the microbiome in the context of clinical variables comes from repeated measurements, i.e. the same measurement taken at multiple time points or multiple measurements made at a single time point. Even in non-equilibrated communities, where variance between repeated measures is high, measurements of the microbial community are expected to be highly correlated with each other, thus presenting a problem for standard statistical methods. However, repeated measures can provide important data for processes that evolve or change over time. Techniques to analyze repeated measures would be of use to the microbiome field as they are often necessary to obtain a more complete understanding of a system of interest.

An additional challenge in analyzing clinical outcomes and biomarkers in light of the microbiome is that the outcomes are often continuous rather than dichotomous variables. Continuous variables are those that can take on any value within a given range, and when they are converted to a categorical or dichotomous format, in some instances, information is lost. In practice, count variables, although not technically continuous, are treated as continuous variables. These continuous variables, as opposed to categorical variables, have repeatedly been dichotomized in the microbiome literature (33, 49) with the potential for loss of nuance in the relationship between them and the microbiota.

Mixed models—both generalized linear mixed models (GLMMs) and linear mixed models—have been used in ecology at least as long as methods for microbiome studies have existed (77). These models incorporate both fixed effects that are the same for every observation or sample, and random effects that apply to select samples or groups of samples. Through the use of random effects, linear mixed models are designed to handle repeated measures and other complex study designs (77). In addition, generalized linear models (GLMs) attempt to model

data that do not follow a traditional normal distribution. The linear relationship between the outcome and predictors is redefined as the set of linear predictors and their relationship to the expected value of the outcome via a "link" function. This link function, along with the variance of the expected value of the outcome, are selected from the members of the exponential distribution family, which are well known.

We focused on the GLMM method because it is the only analysis method that handles both continuous variables and repeated measures. GLMMs have just recently been incorporated into microbiome studies (31, 78, 79). These early adopters of the GLMM methods primarily use the sample group (i.e. sample site, treatment, pregnant/non-pregnant) to explain species abundance. When combined with a penalty parameter—an additional term that eliminates extraneous explanatory variables—GLMMs can use species abundance to explain clinical laboratory measurements (including continuous measurements such as cholesterol, blood glucose, cytokines) and other clinical measures.

Penalized regression models have been used in genomics and metagenomics studies for several years (80). Of the two most basic penalty types, lasso and ridge (also known as L1 and L2, respectively), the lasso penalty has the advantage of performing variable selection by reducing some coefficients to zero. In comparison, the ridge penalty shrinks some coefficients towards but not all the way to zero. The elastic net penalty, which is the combination of the lasso and ridge penalties, reduces some coefficients to zero and shrinks others, thereby limiting its capacity to perform variable selection (81). Only the lasso penalty performs variable selection without having to decide on a coefficient size threshold to define association.

The lasso penalized generalized linear mixed model (LassoGLMM), originally developed in 2011 for sports statistics and human-computer interactions (82, 83), has many properties that

make it well-suited for microbiome applications. This model leverages the power gained by repeated measures and compensates for the large number of variables. The lasso penalty forces some coefficients to be equal to zero, leaving only those variables (or in our case, microbes) with the strongest associations with non-zero coefficients. This feature resolves the problem of having many more explanatory variables than observations. The mixed effects in the LassoGLMM also allow for repeated measures by including a random effect for each subject and repeated measurement.

We now present an application of the LassoGLMM to examine the relationships between the microbiome and continuous variables related to health and inflammation from clinical studies of the respiratory tract. We applied a LassoGLMM with a correlation-based variable screening step to two microbiome datasets: a 16S rRNA gene survey of the oral microbiota from the Oral Cyclosporine in Chronic Obstructive Pulmonary Disease study (OC-COPD; clinicaltrials.gov ID: NCT00974142, a randomized controlled clinical trial), and a combination bacterial 16S rRNA gene and fungal Internal Transcribed Spacer (ITS) survey of the bronchoalveolar lavage (BAL) for the Pittsburgh site of Lung HIV Microbiome Project (LHMP; clinical trials ID: NCT00870857, an observational cohort study). In the OC-COPD study, we sought to discover associations between the oral microbiota and laboratory values measured in peripheral blood. In the LHMP, we aimed to identify which bacteria and fungi were associated with increased inflammation both locally in the lungs and systemically in the blood.

2.2 METHODS

Multiple specimens including oral washes and BAL for microbiota characterization, and blood for chemistry, inflammatory markers, and other laboratory measurements were collected as part of the OC-COPD and the LHMP. The OC-COPD dataset included 15 samples from 8 individuals at pre-randomization (trial week 0) and at trial week 16 (one participant did not have a sample for the pre-randomization visit). These OC-COPD participants, who were sequentially enrolled from the parent trial, had advanced COPD but were free of active infections. Specific inclusion criteria included: between 45-80 years of age, having advanced COPD (defined as forced expiratory volume in 1 second, FEV1, between 25% and 60% predicted), and being non-responsive to traditional inhaler therapy. Once enrolled, participants were randomized to receive the test drug, cyclosporine (an immune suppressant), or placebo for 16 weeks. Additional eligibility requirements for the trial are described at clinicaltrials.gov, identifier NCT00974142. Laboratory outcomes include 32 blood measurements found in a typical blood chemistry panel with electrolytes. Clinical independent variables used were gender and treatment group (test drug or placebo).

The LHMP lung microbiome dataset contained 30 samples from 21 participants who had BAL performed on their right and left lungs at the same clinical visit. This group included both HIV-infected (HIV+; N=11) and HIV-uninfected (HIV-; N=10) individuals, and could be classified as current smokers (N=3), former smokers (defined as having quit more than 6 months prior to the study; N=3), and never smokers (defined as having smoked fewer than 100 cigarettes in a lifetime; N=15). Inclusion criteria included no use of antibiotics in the past three months and no evidence of acute respiratory disease for four weeks. The lung microbiome was sampled by BAL following an oral wash and gargle with antiseptic mouthwash. Specific inclusion criteria

and sampling procedures can be found in (4). The 16S and ITS rRNA sequence data are described in (4) and (40), respectively. Laboratory outcome variables include 12 cytokines measured in both the BAL and the blood; 6 cytokines that were detectible in less than 10% of the samples were excluded from further analysis. Clinical independent variables used were HIV status and smoking history category.

2.2.1 Sequence Data Processing

The sample processing procedures were performed as previously described in (4) and (40). In brief, all samples had DNA extracted using standard techniques with the PowerSoil® DNA Isolation Kit from MO BIO (Carlsbad, CA). For the OC-COPD, the bacterial V4 hyper-variable region of the 16S rRNA gene was amplified and sequenced on the Illumina MiSeq platform. For the LHMP, the hyper-variable regions 1 through 3 (V1-V3) were amplified and sequenced using the Roche 454 GS-FLX platform with Titanium chemistry. For fungal DNA sequencing, the ITS1 was amplified and sequenced on the Ion PGM^{TM} Sequencer using the 400 bp protocol (60). Sequences were processed using the QIIME pipeline version 1.7 (84) with default settings for de novo Operational Taxonomic Unit (OTU) picking. Bacterial 16S rRNA gene sequences were clustered at 97% similarity and fungal ITS sequences were clustered at 99% similarity. Additional processing and taxonomic assignment for the ITS sequences was performed using FHiTINGS (85). Samples with fewer than 1,000 16S rRNA bacterial reads, and samples with fewer than 100 ITS fungal reads were considered to have failed and were removed from further analysis.

After initial taxonomic assignments were made using the default settings in QIIME or FHiTINGS, OTUs were combined by taxonomic assignment at the genus level. For each kingdom, all genera counts were normalized using total sum scaling, also known as relative abundance. Any bacterial genus present in fewer than half of the samples was removed. Due to greater diversity between samples in the fungal genera, we reduced this cut off to remove genera present in fewer than 10% of the samples.

2.2.2 Variable Screening Step

The number of genera present is often at least an order of magnitude larger than the number of subjects sampled. When seeking to assess the relationship of microbiota components with clinical variables, the mismatch in number of subjects versus microbial variables presents an analytic challenge. We overcome this problem by preceding LassoGLMM regression with a variable screening step based on correlation. For each response-genera pair, Spearman correlations deemed significant ($p \leq 0.05$) without multiple testing correction were used as independent variables in the regression model. **Figure 2.1** shows an overview of this two-step method.

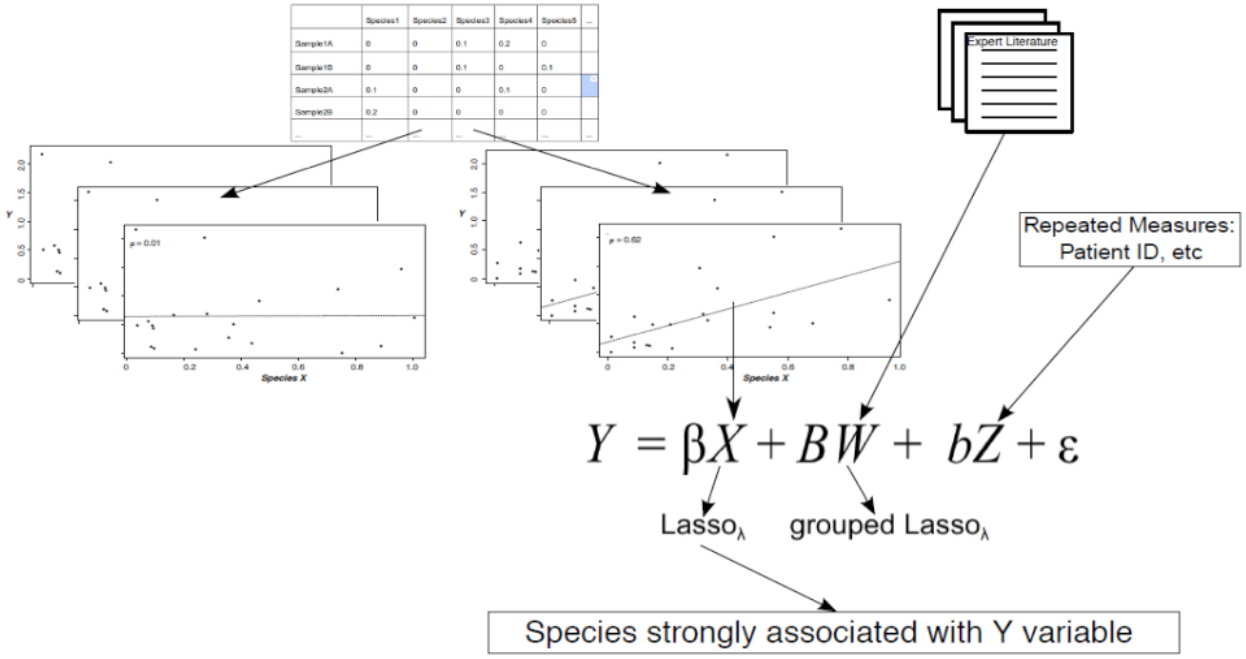

Figure 2.1 Overview of the two-step LassoGLMM model developed. Species (or OTUs or any other explanatory variables of interest) are divided into those that are correlated with the dependent continuous variable, Y, and those that are not. Species that are correlated are stored in a matrix X. Relevant categorical variables, found through a review of expert literature or other means, are stored in a matrix W. Indicators of repeated measures such as patient ID are stored in matrix Z. Matrices X, W, and Z are entered into a generalized linear mixed model to be regressed on outcome variable Y. Coefficient β for matrix X and coefficient B for W are subjected to the lasso penalty. Any species that retain non-zero coefficients are considered strongly associated with the dependent variable Y.

2.2.3 Lasso-Penalized Generalized Linear Mixed Model

The LassoGLMM combines variable selection with the flexibility to account for repeated measures and other random effects. It can be built up from the random-intercept linear mixed model:

$$Y = \beta X + bZ + \varepsilon, \qquad (3.1)$$

where Y is the response variable, or outcome of interest, X is the matrix of the fixed effects including genera abundances, β is the vector of coefficients associated with the fixed effects, Z is the matrix form of the random effects including patient, b is the vector of coefficients associated with the random effects, and ε is the random error. For example, we modeled the response variable, Y, of blood glucose on the relative abundance of bacterial genera in the mouth, formatted as a matrix X, while accounting for the individual participant as a random effect, Z. This traditional format highlights the breakdown of independent variables into two groups: the fixed effects and the random effects. Fixed effects are those that are the same for all observations or samples, for example genera abundances and disease status. Random effects are those that are unique to each observation or group of observations, for example the study participant or time point. The unit of repeated measurement, in our case the individual, is always considered a random effect. In the OC-COPD study, the same subject was sampled at a pre-randomization visit and 16 weeks later. Although we did not expect there to be high correlation between individuals, we included the visit as a random effect to account for any seasonal or batch processing effects. In the LHMP study, the right lung and the left lung were sampled in the same subject in the same visit in a randomized order (right first, or left first). We included an identifier

for the first and second sides to be sampled as a random effect to account for any order bias, including the possibility of higher contamination from the upper respiratory tract in the first side.

The fixed effects can be split again into continuous and categorical variables, and the resulting formula becomes:

$$Y = \beta X + BW + bZ + \varepsilon, \qquad (3.2)$$

where X now only contains the continuous fixed effects and W is the matrix form of the 'dummy' variables indicating each level among the categorical variables including disease status. This split is important for the penalization of the categorical variables described below. In our regression models we included the following categorical variables that are known to be associated with the outcomes (Y) of interest: gender (86) and treatment (drug or placebo) in the OC-COPD models; smoking (87) and HIV (88) status for the LHMP models.

By their nature, many of the variables (genera or OTUs) in microbiomes are highly correlated with each other. This correlation makes including all variables in the regression redundant and necessitates the use of the lasso or other penalty. During the maximal likelihood estimation of the β, B, and b coefficients, the lasso penalty is added to the log-likelihood approximation. The penalty parameter λ performs variable selection by forcing the smaller β and B coefficients to equal zero. All of the B values of one categorical variable are penalized together with a grouped Lasso penalty adapted from (89). Thus, either all possible statuses are included in the model, or none are included. For example, the LHMP smoking status 'current', 'former', and 'never' result in two dummy variables, one for 'current' and one for 'former'. The B coefficients for both dummy variables are either reduced to zero or included in the model. By increasing λ, more of the β and B coefficients will be forced to zero. It is important to note that

only the fixed effects coefficients are subject to the lasso penalty. Random effects are included in the model regardless of the size of λ.

We determined the optimal lasso penalty term (λ) for each model by scanning between 0 and 200 (by increments of 1) using the R package glmmLasso version 1.3.3 (90). The model with the lowest Bayesian Information Criteria (BIC) (90) was selected as optimal. When λ=0, if the Fisher matrix was not invertible (i.e. the regression could not be completed) we started the scan at λ=1. We considered those genera with non-zero coefficients in the model using the optimal penalty term to be strongly associated with the response variable. Following Groll's recommendation (82), we then ran a GLMM regression including only the strongly associated genera using the R package lme4 (91). This final regression step is related to the adaptive lasso penalty and is designed to compensate for the lack of oracle properties of the basic lasso penalty that we used here (83). These results indicate not only a strong association, but also if the association was positive (more microbes when the variable is high), or negative (more microbes when the variable is low).

2.2.4 Evaluating Models

We evaluated the fit for each of our mixed models using both the marginal and conditional R^2 coefficients of variation (92). Marginal R^2 represents the percent of variation explained by the fixed effects while conditional R^2 represents the variation explained by the entire mixed model. These values provide a more absolute measure of the goodness of fit for the model in question compared to the BIC that was used for penalty optimization. We also inspected the residual plots to ensure that the relationship between the microbes and clinical variables was linear. When a

relationship was found to be non-linear, we attempted to refit the model with a generalized model.

2.2.5 Dichotomous Methods

Because there is no consensus method to evaluate the association between microbiota abundance and a continuous variable, we compared our LassoGLMM method to the most basic dichotomous variable method, the Wilcoxon (or Mann-Whitney U) test (93). The Wilcoxon test is a non-parametric statistical test that determines if the genus tends to be more abundant in one group than in another based on ranks. To dichotomize our data, we divided samples into those above and below the sample average for the outcome of interest.

2.2.6 Ethics approval and consent to participate

Written informed consent was obtained from all participants in both studies following approval of human subjects' protection protocols from review boards of the University of Pittsburgh, University of California San Francisco, and the University of California Los Angeles.

2.2.7 Availability of data

The sequence data supporting the results of this study are available in NCBI sequence read archive (SRA) under accessions PRJNA308310 (OC-COPD), SRP065274 (LHMP 16S), and SRP040237 (LHMP ITS). The R code that was used to implement LassoGLMM is available at https://github.com/ghedin-lab/LassoGLMMforMicrobiomes and can be found in **Appendix A**.

2.3 RESULTS

2.3.1 Associations between Oral Bacteria and Laboratory Measurements

To identify associations between the easily accessible oral bacteria and laboratory values measured in blood, we characterized the microbiota in 15 oral wash samples from 8 individuals at two different time points, 16 weeks apart. A metabolic panel of 32 measurements, including electrolytes and cholesterol levels, was performed at each visit. In the 15 oral washes, we found a total of 95 bacterial genera present in at least half the samples. Each sample was dominated by *Streptococcus* (mean: 32.2%, standard deviation: 11.6), *Prevotella* (mean: 12.4%, SD: 6.5), *Rothia* (mean: 10.6%, SD: 6.5), *Fusobacterium* (mean: 6.2%, SD: 5.0), and *Veillonella* (mean: 5.6%, SD: 3.7).

We calculated Spearman correlations between every pair of bacterial genera and blood metabolic profile measurement. There were 202 correlations (out of 1,425 possible) that were nominally significant, $p < 0.05$ before correcting for multiple hypotheses testing. Each clinical

variable was significantly correlated with 1 to 20 genera, averaging 7.5 nominally significant correlations. Out of the 95 genera, 75 were nominally significantly correlated with 1 to 9 of the clinical variables.

The genera that had nominal significant correlations with a clinical variable were entered into a LassoGLMM as potential explanatory variables along with Cyclosporine/placebo treatment assignment and gender. All but 64 genera (out of 202) coefficients were forced to zero by the Lasso penalty. Coefficients that were not forced to zero are presented in **Table 2.1** and are considered strong associations. Ten laboratory measures were associated with bacterial genera since their models retained non-zero coefficients (see **Figure 2.2**): percent neutrophils (model O1), blood urea nitrogen (BUN) (model O2), immunoglobulin M (IGM; model O3), partial pressure of oxygen (model O4), SAT (model O5), alkaline phosphatase (model O6), serum glutamic oxaloacetic transaminase (SGOT; model O7), serum glutamic-pyruvic transaminase (SGPT; model O8), cholesterol (model O9), and glucose (model O10). Of these lab measures, BUN, IGM, partial pressure of oxygen, SAT, and SGPT (models O2, O3, O4, O5, and O9) were strongly associated with all correlated bacterial genera (optimal penalty parameters of 0). For the remaining 5 models, the optimal λ penalty parameter ranged from 2 to 144. The higher λ penalty parameters eliminated some bacterial genera in all 5 models but also eliminated drug treatment assignment in model O10, and gender in model O9.

Table 2.1: Laboratory measurements and their strongly associated bacteria in OC-COPD. Bacteria that could not be classified to the genus level are listed at the lowest taxonomic level that could be confidently identified. Bacteria in bold are negatively associated with the laboratory measurement, indicating that higher microbial abundance is associated with lower measurement level.

Percent Neutrophils (O1)	BUN (O2)	IGM (O3)	Partial Pressure of Oxygen (O4)	SAT (O5)
Bacteroidales (order)	**Aerococcaceae (family)**	*Pseudomonas*	***Bacillus***	***Bacillus***
S-24 (family in Bacteroidales order)	***Enterococcus***		*Pseudomonas*	*Pseudomonas*
Clostridiaceae	***Streptococcus***			
Oribacterium	Lachnospiraceae (family)			
Oscillospira				
Ruminococcus				
Phascolactobacterium				
Succinivibrio				

Table 2.1 Continued

Alkaline Phosphatase (O6)	SGOT (O7)	SGPT (O8)	Cholesterol (O9)	Glucose (O10)
Bifidobacteriaceae (family)	**Bacillales (order)**	***Rothia***	**Micrococcaceae (family)**	*Rothia*
***Weeksellaceae* (family)**	***Lachnospiraceae* (family)**	***Scardovia***	*Porphyromonas*	***Porphyromonas***
***Gemellales* (order)**	***Moryella***	*Clostridiales* (order)	*Prevotella*	***Tannerella***
***Gemellaceae* (family)**	*Oribacterium*	**Lachnospiraceae (family)**	*Catonella*	***Prevotella***
Granulicatella	*Peptostreptococcus*	***Moryella***	*Filifactor*	**Gemellaceae (family)**
Lactobacillus	***Eikenella***	*Oribacterium*	*Peptostreptococcus*	*Lactobacillus*
Eikenella	***Neisseria***	***Schwartzia***	Mogibacteriaceae (family)	**Peptostreptococcaceae (family)**
Neisseria	***Cardiobacterium***	*Succinivibrio*	TG-5 (member of Dethiosulfovibronaceae family)	***Peptostreptococcus***
Aggregatibacter			*TM-7.3*	*Veillonella*
			Mycoplasma	**Mogibacteriaceae (family)**
				TG-5 (member of Dethiosulfovibronaceae family)
				Mycoplasma

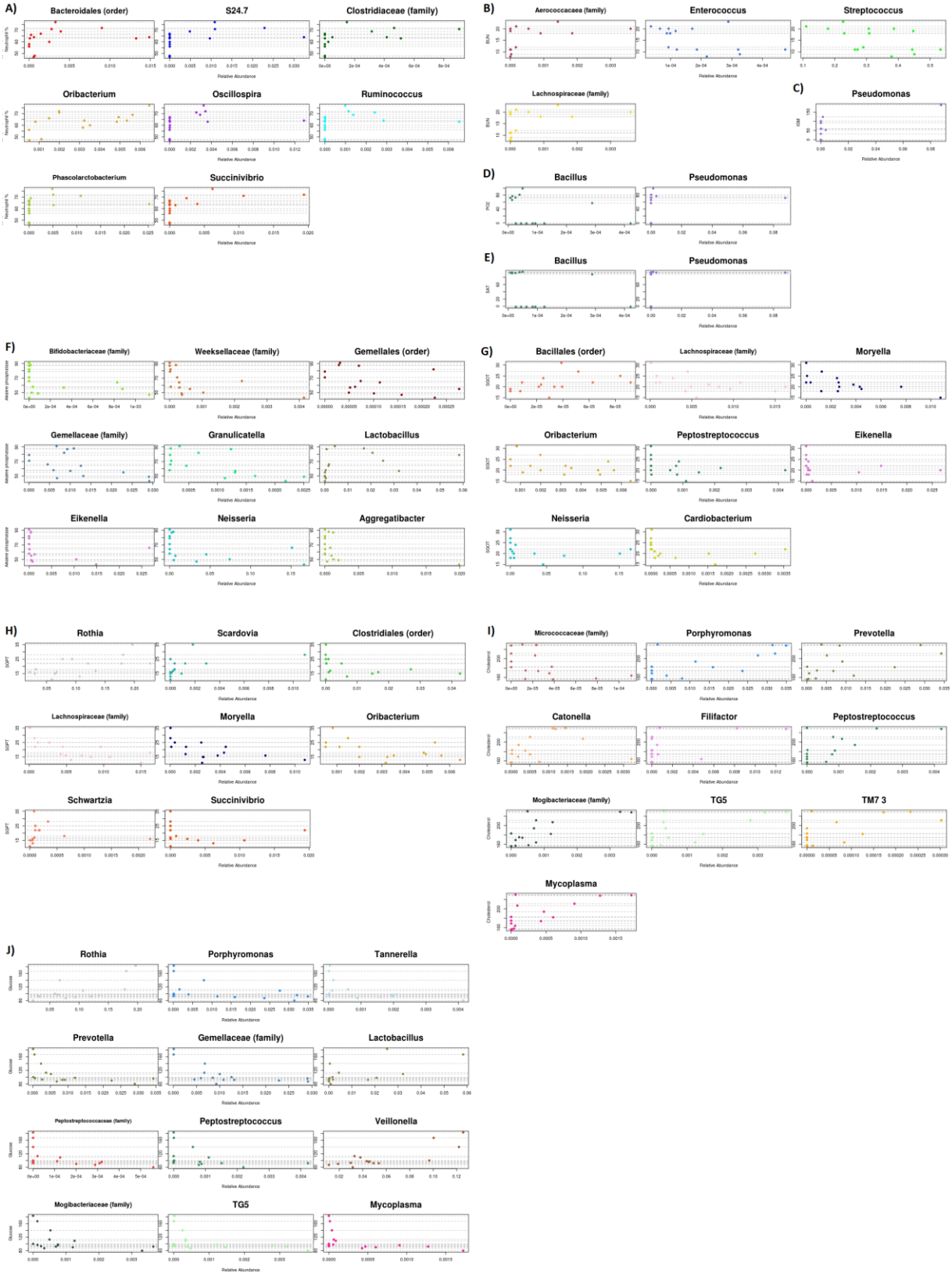
A)
Bacteroidales (order)
S24.7
Clostridiaceae (family)
Oribacterium
Oscillospira
Ruminococcus
Phascolarctobacterium
Succinivibrio
B)
Aerococcaceae (family)
Enterococcus
Streptococcus
Lachnospiraceae (family)
C)
Pseudomonas
D)
Bacillus
Pseudomonas
E)
Bacillus
Pseudomonas
F)
Bifidobacteriaceae (family)
Weeksellaceae (family)
Gemellales (order)
Gemellaceae (family)
Granulicatella
Lactobacillus
Eikenella
Neisseria
Aggregatibacter
G)
Bacillales (order)
Lachnospiraceae (family)
Moryella
Oribacterium
Peptostreptococcus
Eikenella
Neisseria
Cardiobacterium
H)
Rothia
Scardovia
Clostridiales (order)
Lachnospiraceae (family)
Moryella
Oribacterium
Schwartzia
Succinivibrio
I)
Micrococcaceae (family)
Porphyromonas
Prevotella
Catonella
Filifactor
Peptostreptococcus
Mogibacteriaceae (family)
TG5
TM7 3
Mycoplasma
J)
Rothia
Porphyromonas
Tannerella
Prevotella
Gemellaceae (family)
Lactobacillus
Peptostreptococcaceae (family)
Peptostreptococcus
Veillonella
Mogibacteriaceae (family)
TG5
Mycoplasma

Figure 2.2 OC-COPD associations between laboratory measurements and bacteria identified by LassoGLMM. Strong associations between bacteria and A) percent neutrophils (O1), B) BUN (O2), C) IGM (O3), D) partial pressure of oxygen (O4) E) SAT (O5), F) alkaline phosphatase (O6), G) SGOT (O7), H) SGPT (O8), I) cholesterol (O9), and J) glucose (O10). Each horizontal grey line represents an individual. Each colored line represents a microbe. When a colored circle is located on the grey line, it is the relative abundance of that microbe for that subject. Perfect positive association between clinical variable and bacteria would be a line from the bottom-left to the top-right of the figure and would have a highly positive β coefficient in the LassoGLMM. Perfect negative association would be a line from the top-left to the bottom-right of the figure and would have a highly negative β coefficient.

2.3.2 Associations of Lung Bacteria and Fungi with Cytokines

Using the LHMP dataset, we sought to identify associations between indicators of local or systemic inflammation and bacteria and/or fungi detected in BAL samples. We used bacterial and fungal surveys previously performed on 30 BAL samples from 21 individuals (4, 40). Across all samples 49 bacterial genera were found in at least half of the samples and 28 fungal genera were found in at least 10% of the samples. There were 106 correlations (out of 1,386 possible) that were nominally significant at $p < 0.05$. Each cytokine had between 2 and 9 nominally significant correlations with bacterial and fungal genera (average number of genera nominally correlated with each cytokine = 5.9). Conversely, of the 77 genera identified, 42 were nominally significantly correlated with 1 to 7 cytokines.

These bacterial and fungal genera were entered into the LassoGLMM along with HIV status and smoking status as potential explanatory variables. As in the oral microbiome models, most genera coefficients (103 out of 106) in the LHMP models were forced to zero by the Lasso-penalty. All fungal genera coefficients were forced to zero. The 3 bacterial genera that maintained non-zero coefficients are presented in **Table 2.2**. In 16 models assessing cytokine associations with genera and species, all genera/species coefficients were forced to zero, which indicates that increases in the cytokines are best explained by HIV and/or smoking status. The remaining 2 models with evidence of strong genera association retained non-zero coefficients (see **Figure 2.3**). These models were: BAL interleukin receptor antagonist (IL-ra) (model L1), and systemic IL-ra (model L2). BAL IL-ra (model L1) had an optimal penalty parameter of 0, indicating that both correlated bacteria were strongly associated with BAL IL-ra; no fungi were nominally correlated with BAL IL-ra. Conversely, systemic IL-ra (model L2) had an optimal penalty parameter of 13, retaining 1 bacterial genus as strongly associated and eliminating 7 others as well as HIV and smoking status.

Table 2.2: Cytokines and their strongly associated microbes in LHMP. Bacteria and fungi that could not be classified to the genus and species level, respectively, are listed at the lowest taxonomic level that could be identified. Microbes in bold are negatively associated with the cytokine, indicating that higher microbial abundance is associated with lower cytokine level.

BAL IL-ra (L1)	Systemic IL-ra (L2)
Clostridia (class)	*Leptotrichia*
Ralstonia	

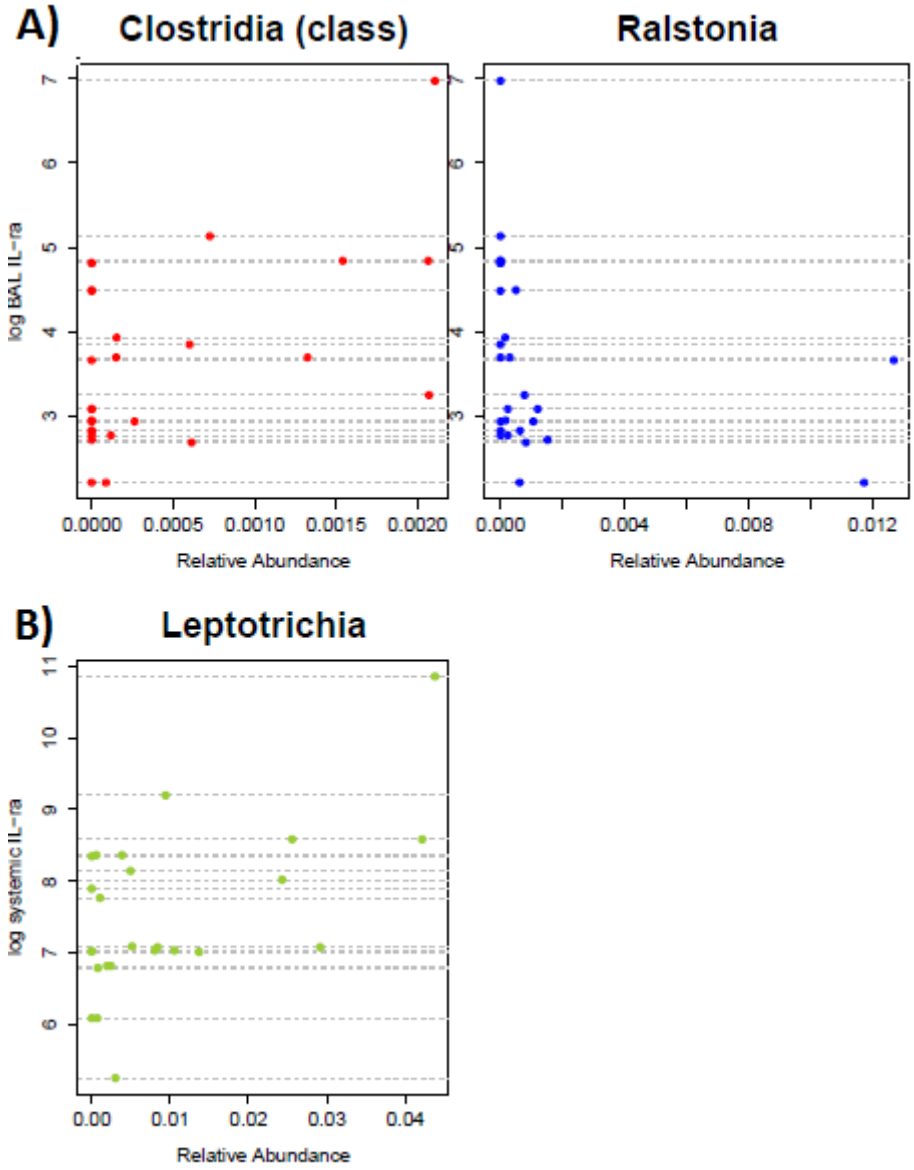

Figure 2.3 LHMP associations between cytokines and bacteria identified by LassoGLMM. Strong associations between bacteria and a) BAL IL-ra (L1) and b) systemic IL-ra (L2). Each horizontal grey line represents a subject. Each colored line represents a microbe. When a colored circle is located on the grey line, it is the relative abundance of that microbe for that individual. Perfect positive association between cytokine and bacteria would be a line from the bottom-left to the top-right of the figure and would have a highly positive β coefficient in the LassoGLMM. Perfect negative association would be a line from the top-left to the bottom-right of the figure and would have a highly negative β coefficient.

2.3.3 Model Evaluation

To evaluate our models, we used both marginal (fixed effects only) and conditional (whole model) coefficients of determination, or R^2 (92). For models O1-O10 we had an average marginal R^2 value of 0.44 (SD 0.32) and an average conditional R^2 value of 0.90 (SD 0.14; **Table 2.3**). These R^2 values demonstrate that our models explained, on average, 90% of the variation seen in the clinical variables and that 44% of the variation is explained by the bacteria that are strongly associated with the laboratory measurements, gender, and drug treatment. However, models O1, O4, O5, O6, and O10 were found to be over-fitting the data with conditional R^2 greater than 0.99. Both LHMP models, L1 and L2, were also found to be over-fitting the data with conditional R^2 equal to 1.00. The residuals from the remaining models indicated that the models fit the data reasonably well (**Figure 2.4**). The most notable exception is in model O3, for IGM, which has large residuals whose pattern indicates a non-linear relationship. We attempted to fit a generalized model to these data, as well as to models O2 and O7, but were unable to significantly improve the fit.

Table 2.3: Marginal and conditional coefficients of variation (R^2) for OC-COPD models and Lasso-penalized GLMM variants. The two-step LassoGLMM method, in columns 1 and 2, is presented here. The original LassoGLMM, in columns 3 and 4, omits the first step of correlation-based variable screening, adding all OTUs to the LassoGLMM. The GLMM with correlated genera, in columns 5 and 6, uses the correlation-based variable screening step, adding only those variables that are correlated with the outcome to the model, but modifies the second step to not include the Lasso penalty. Columns for each method contain the marginal and conditional R^2, which represent fit of the fixed effects and entire model, respectively.

	Two-step LassoGLMM		Original LassoGLMM		GLMM with correlated genera	
	Marginal R^2	Conditional R^2	Marginal R^2	Conditional R^2	Marginal R^2	Conditional R^2
BUN (O2)	0.58	0.60	No non zero coefficients		All correlated variables were in Two-step LassoGLMM	
IGM (O3)	0.19	0.89	No non zero coefficients		All correlated variables were in Two-step LassoGLMM	
SGOT (O7)	0.22	0.84	No non zero coefficients		0.50	0.59
SGPT (O8)	0.44	0.75	No non zero coefficients		All correlated variables were in Two-step LassoGLMM	
Cholesterol (O9)	0.80	0.93	0.95	0.98	0.99	1.00

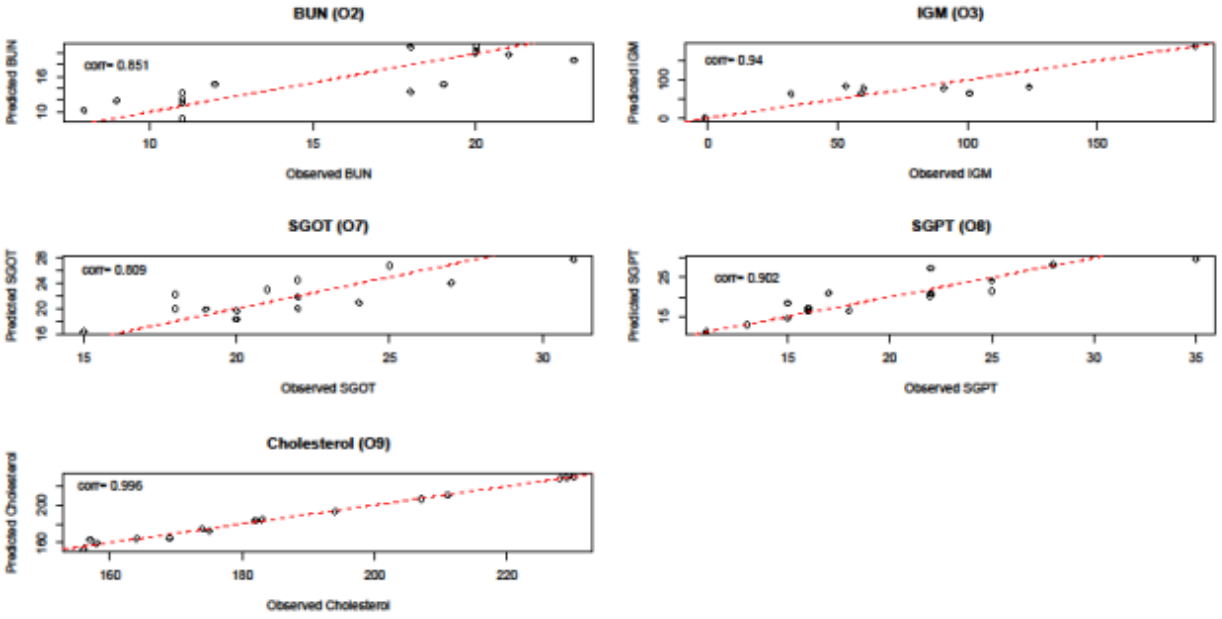

Figure 2.4 Observed vs predicted value plots evaluating the fit of the LassoGLMMs from the OC-COPD study. Each plot represents one LassoGLMM with non-zero coefficients that was not found to be over-fitting the data. The value observed (X-axis) is plotted against the value predicted by the LassoGLMM (Y-axis). Each point represents a sample. The red line indicates where the predicted value matches the observed value. For models that deviate from this line (O2, O3, and O7), we attempted to fit a generalized model but found no significant improvements in fit.

We then compared our models with LassoGLMMs, as originally described by Groll (82, 90), leaving out our first-step of variable screening, and to non-penalized GLMMs that include all correlated genera that passed variable screening, thus modifying the second step of our two-step LassoGLMM method. The GLMM with all correlated genera can also be thought of as a two-step LassoGLMM with a lambda penalty parameter of 0. When 0 is the optimal lambda for the two-step LassoGLMM method presented here, these two models are identical. The marginal

and conditional R^2 values for each model are included in **Table 2.3**. With the notable exception of model O9, we found that our two-step model performed at least as well as the original LassoGLMM without a variable screening step and applying a non-penalized GLMM after a variable screening step. By including both the variable screening step and the lasso penalty, our two-step method successfully found associations that would have been missed when the original LassoGLMM retained no non-zero coefficients. It is also capable of finding identical models to the non-penalized GLMM with all correlated variables.

2.3.4 Comparison to Categorical Methods

To evaluate the performance of our method as compared to the categorical methods that are used most frequently in the microbiome field, we dichotomized the continuous variables based on their average values; we then compared the microbiota between the two groups using a Wilcoxon test (93).

For the ten models with non-zero coefficients in the OC-COPD (models O1-O10), the Wilcoxon tests found between 1 and 12 (average 5.4) bacterial genera to be differentially abundant between above- and below-average outcome groups, before correcting for the large number of tests (**Figure 2.5**). For each of the 2 cytokines with non-zero coefficients in the LHMP (models L1-L2), the Wilcoxon test found 2 bacterial and 1 fungal genera to be differentially abundant between above- and below-average cytokine groups (**Figure 2.6**). When each outcome or cytokine was corrected for multiple hypotheses testing using the Benjamini-Hochberg false discovery rate (94), no genera were significantly differentially abundant. Before multiple hypotheses testing correction, the 60 significantly different genera across all 12 models

showed 52% overlap with the 67 genera identified as strongly associated with the outcome by the LassoGLMM. With one exception (Leptotricia in model L2), all genera identified by the LassoGLMM had a Wilcoxon test p-value no greater than 0.23, indicating that there is a difference between the samples with high and low outcomes that may be identified by a test with more statistical power or a much larger sample size.

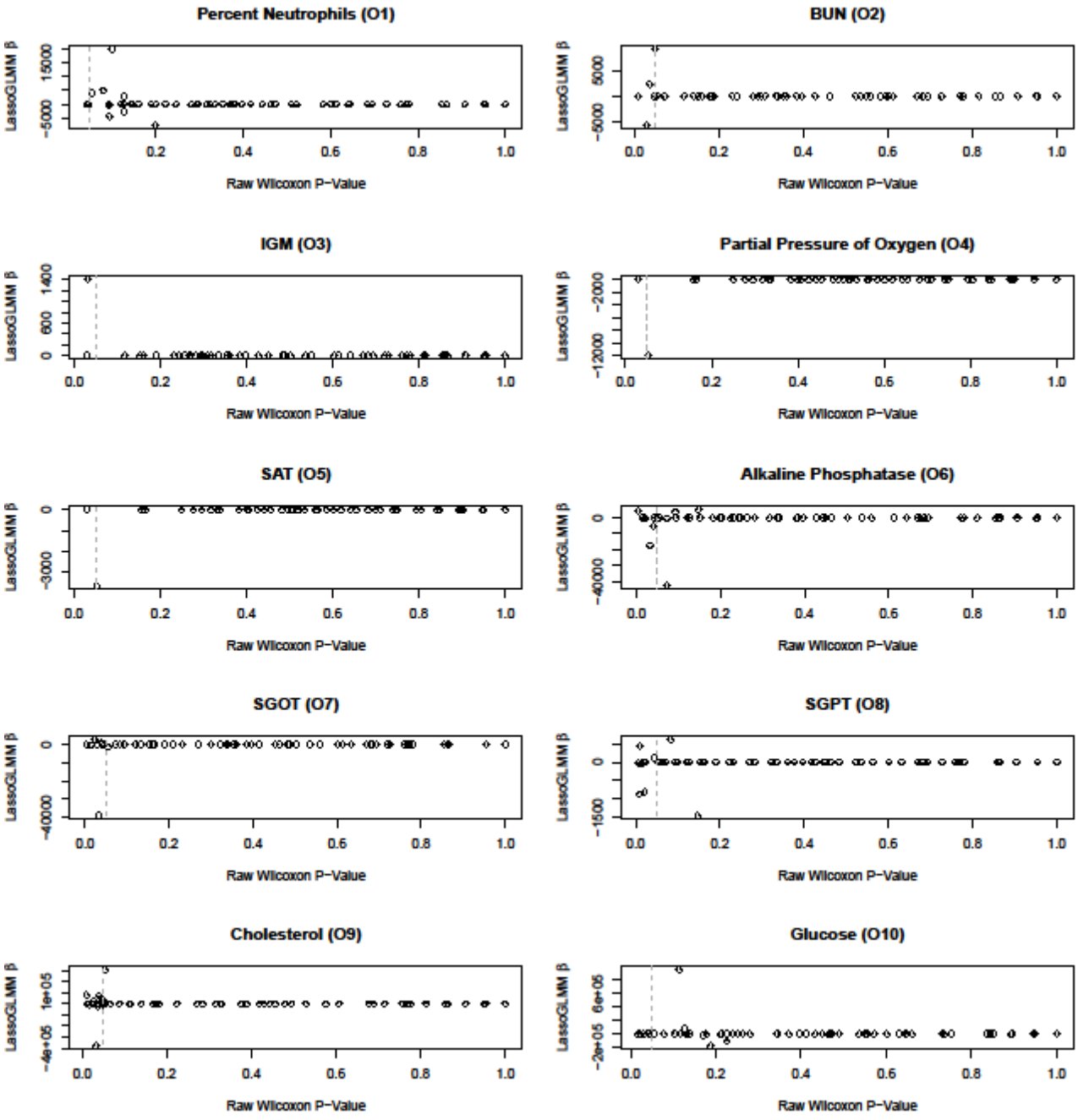

Figure 2.5 Wilcoxon P-values compared to LassoGLMM β coefficients for OC-COPD study. Each plot represents one LassoGLMM with non-zero coefficients. For each bacterial genus, the Wilcoxon P-value (before adjustment for multiple hypotheses testing) is plotted on the X-axis and the LassoGLMM β coefficient is plotted on the Y-axis. Most β coefficients are equal to zero. The dashed vertical line indicates nominal significance based on a Wilcoxon P-value of 0.05.

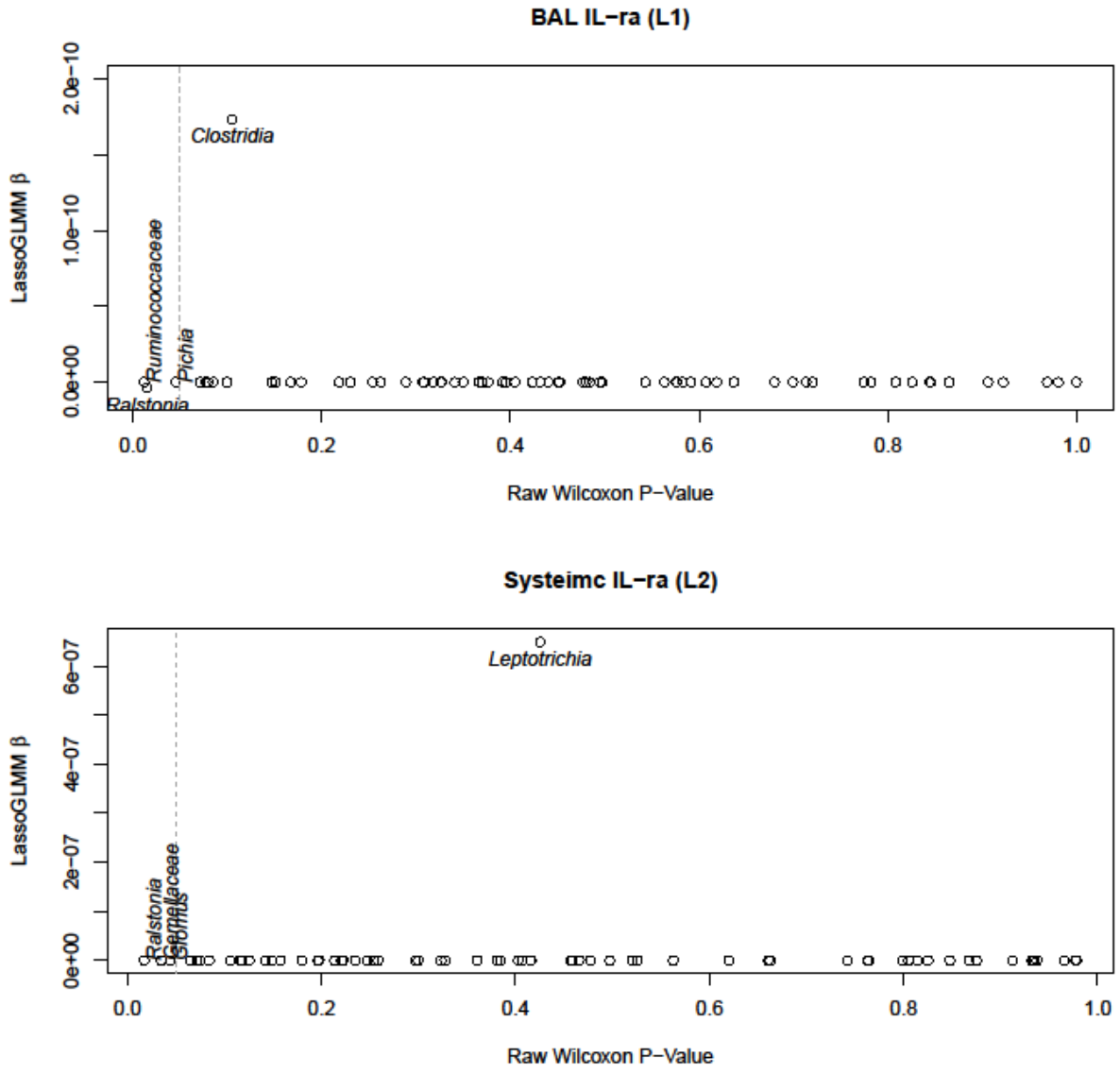

Figure 2.6 Wilcoxon P-values compared to LassoGLMM β coefficients for LHMP study. Each plot represents one LassoGLMM with non-zero coefficients. For each bacterial or fungal genus, the Wilcoxon P-value (before adjustment for multiple hypotheses testing) is plotted on the X-axis and the LassoGLMM β coefficient is plotted on the Y-axis. Most β coefficients are equal to 0; those that are not are labeled with their lowest taxonomic assignment appearing horizontally. The dashed vertical line indicates nominal significance based on a Wilcoxon P-value of 0.05. The nominally significant genera that have a β coefficient of 0 are labeled with their lowest taxonomic assignment appearing vertically.

2.4 DISCUSSION

We applied the LassoGLMM to two mucosal microbiome datasets to analyze the relationship of microbes and their abundances to continuous clinical variables with repeated measurements. We were able to computationally identify a number of associations between microbes and continuous clinical variables, including standard blood chemistries. To date, there is no other established approach to relate repeatedly measured continuous outcomes to microbes and their abundances and thus this method represents an important addition to the field.

Traditionally, associations between microbial abundance and continuous outcomes, with repeated measures or not, have been built on an arbitrary grouping of samples derived from the values found within the study itself. Often samples are grouped by whether their measurement is above or below the study average, as we did in our comparison of the LassoGLMM to the Wilcoxon test. The study-dependent splitting of a variable limits reproducibility and ignores natural variation in the larger population. Any association between microbial abundance and a repeated clinical measurement found by this type of test ignores the fact that repeated samples are not independent of each other. This limitation may explain why there was minimal overlap between the genera identified by the LassoGLMMs and the Wilcoxon test.

Repeated measurements taken in the clinic, such as ours, break a number of assumptions that are common among statistical tests, even those developed specifically for microbiome studies. Multivariate Association with Linear Models (MaAsLin) was recently developed to simultaneously find associations between microbes and multiple clinical outcomes, including continuous variables, through variable selection and linear modeling (95). These models make the assumption that all samples are independent and do not allow for the complex covariance structure that accompanies repeated measurements. The two-part zero-inflated Beta regression

model with random effects (ZIBR) is designed to handle repeated measurements through the use of random effects, but assumes that all subjects will have samples taken at the same time points with no missing measurements (96). With so many points of failure, from a missed appointment to failed amplification and sequencing, real-world clinical datasets rarely contain all time points for all subjects. Both MaAsLin and ZIBR use microbial abundances as the response variables and clinical measurements as the explanatory variables. This set up does not allow for correlations or interactions between microbial abundances beyond compositional effects.

One of the advantages of the LassoGLMM is the ability to find associations in small sample sizes. The statistical power of GLMMs is best calculated by simulations that account for the impact of the random effects (97), which are largely unknown in microbiome studies. Despite the lack of a power analysis, we were able to find associations in our studies, both of which have small sample sizes by any standard. However, many of the early microbiome studies also had small samples sizes and, with their data accessible in public repositories, are available for re-analysis with newly developed tools such as the LassoGLMM.

Another added advantage of the LassoGLMM is the ability to account for correlations between genera, which may be indicative of biological interactions. Too many interactions or correlations between genera can be problematic for the lasso penalty, as it may discard a biologically important genus while retaining a non-zero coefficient for a correlated but less biologically important genus. The number of interactions can be mitigated by reducing the number of genera entered into the LassoGLMM with a variable screening step. The “choices” that the lasso penalty makes highlight the need to study the relationships between the genera in addition to their relationships with the outcome variable. Genera whose coefficients are pushed to zero may be chemically or physically interacting with genera whose coefficients are non-zero.

Or, if negatively correlated with each other, may be performing the same function. This redundancy may stem from bacterial interactions or from competition to fill the same niche. Biological interactions between genera within a microbiome represent an area of active research and in the meantime, methods such as LassoGLMM that can account for these uncharacterized interactions should be better able to determine associations than methods that ignore them.

A separate area of active research that will likely lead to improved discovery of associations between clinical variables and the microbiome is penalization parameters. Here we used a single parameter lasso penalty, applying the same penalty to both the continuous and discrete fixed effects. In graphical models with combinations of continuous and discrete variables, separate penalty parameters have been shown to improve accuracy in graphical models (98), and may have a similar impact on other regression models.

2.5 CONCLUSIONS

The potential applications of the LassoGLMM are multiple and go beyond what we have used it for here. We took advantage of the ability to account for potentially confounding categorical variables, treatment assignment and gender in OC-COPD, and HIV status and smoking status in LHMP. This ability can be used to account for attributes that are known or suspected to influence the outcome variable, including host genotype. We made use of the ability to analyze repeated measurements from the same individual, over two time points in OC-COPD, and in two lung locations (right and left lungs were sampled separately) in the LHMP. The method can accommodate any number of repeated measurements, including long-term longitudinal studies,

even when the number of measurements per individual is not identical. The inclusion of the individual as a random effect also accounts for an uneven number of observations per subject, a common issue in the clinic where study participants can be followed for different lengths of time, can be “lost to follow-up”, may die, or may drop out of the study. The generalized nature of the LassoGLMM also allows for the analysis of variables that do not follow a normal distribution, including time-to-event and categorical outcomes. The lasso penalty allows for variable selection to select the strongest genera associations but the selection criteria may be influenced by the correlations between microbes inherent in relative abundance data. However, the LassoGLMM is not limited to relative abundance data and when a consensus is reached about the optimal normalization methods for microbiome data, this method will be able to handle that data and improve performance.

We have demonstrated that the lasso-penalized generalized linear mixed model can be applied to microbiome studies with continuous outcomes and repeated measures. This model works well with both 16S rRNA gene surveys and more complicated 16S/ITS combination studies. The method combines the well-established lasso penalty to account for the large number of variables with the mixed model to account for repeated sampling—including longitudinal studies—and other variables that are known to be associated with the outcome. The power of this method lies not only in its ability to identify known associations between microbes and continuous clinical variables, but in its ability to identify novel associations that can be used to test new potential biomarkers.

3.0 INFERRED CROSS-DOMAIN INTERACTIONS IN THE LUNG AND SKIN MICROBIOMES[4]

3.1 INTRODUCTION

Determining networks of microbial interactions that affect the fitness of individual species is relevant for the functional characterization of a microbial community. These interactions can vary across time and space, depending on both abiotic and biotic factors. Common abiotic factors include oxygen, temperature and pH, while biotic factors can include the presence or absence of other microbes. The ability to predict biological interactions between microbes based on next-generation sequencing data, particularly from targeted amplicon sequencing (TAS), has been a topic of increasing interest with the development of multiple statistical tools for inferring networks (99–101). Within a microbiome, interactions can be informative at both the species and at the community levels.

At the individual species level, knowledge of interactions could provide information relevant to the growth or the targeting of the microbe. Interactions between a microbe that is

[4] Paper in preparation

considered un-culturable and a well-studied microbe that can be grown in culture would increase the chances of successfully growing the un-culturable microbe in the lab either through co-culture, spent media, or the inclusion of metabolites secreted by the well-studied organism. Co-culture has, for example, enabled the cultivation and sequencing of a member of the candidate division TM7, called TM7x, from the human oral microbiome (102). TM7x is now known to be an obligate epibiont of *Actinomyces odontolyticus* and cannot be cultured without it or a related basibiont. On the other end of the spectrum, interactions between a drug-resistant pathogen and drug-susceptible microbes can lead to new treatment strategies targeting the easier to kill microbes to render the pathogen harmless. Fungal pathogens that are notoriously harder to target than bacteria due to their closer evolutionary relationship with their human hosts may be particularly suited for this method of treatment (103).

At the community level, an interaction network reveals useful information about the structure and stability of the community. The topology or graph structure of an interaction network can indicate evolutionary pressures on the community (104). This phenomenon is seen in "hubs," or highly connected nodes, in any ecological interaction graph can indicate keystone species that have a large impact on their environment and many direct and indirect interactions with other species in the community (often despite being present in low abundance) (105). One such keystone species has been identified in the human gut microbiome, *Ruminococcus bromii*, by its superior ability to degrade resistant starches and release nutrients to the direct or indirect benefit of the other members of the microbiome (106).

An alternative topography would be disjointed cluster graphs, where each cluster represents an ecological niche being filled. Such niche separation can be seen in the separation of the lung microbiome of SHIV-infected cynomolgous macaques into bacteria enriched in the

animals that developed chronic obstructive pulmonary disease (COPD) and bacteria that are enriched in the animals that retained normal lung function (107)[5]. This network demonstrated negative associations between two large groups of OTUs and positive associations within each group (**Figure 3.1A**); singleton OTUs (i.e. not associated with any other OTU) were removed. In the first group, we saw OTUs belonging to the genera that were enriched in animals that developed COPD *including Fusobacterium, Prevotella, Veillonella, Neisseria*, and *Porphyromonas* (**Figure 3.1B**). In the second group, we saw OTUs in the genera identified as enriched in non-COPD animals including *Uruburuella* and *Flavobacterium* (**Figure 3.1B**). Other OTUs in this group that were not identified by the other methods belonged to genera including *Kinesporia*, *Enterococcus*, and *Vibrio*. Most interestingly, OTUs belonging to the *Streptococcus* genus were found in both groups, highlighting its importance within both COPD and non-COPD bacterial communities. The separation of the two groups showed the difference in community composition that accompanied the development of COPD. The negative correlations between members of the two groups emphasize how shifts in one species can impact multiple other species, potentially resulting in disease.

[5] This paragraph extracted from paper published in *Microbiome* as "Longitudinal analysis of the lung microbiota of cynomolgous macaques during long-term SHIV infection" (107).

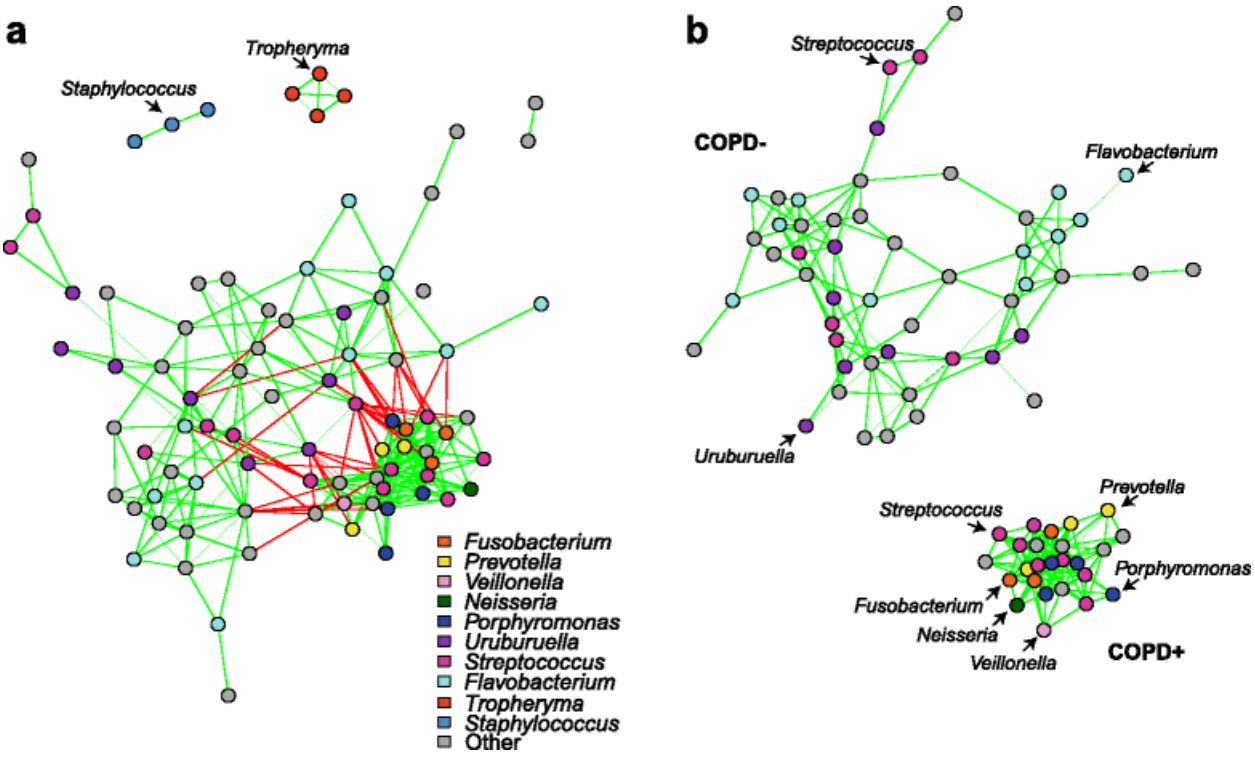

Figure 3.1 SPIEC-EASI Network for Cynomolgous Monkeys with SHIV Infection. The ecological correlation network shows two groups of OTUs that are negatively correlated with each other. One group includes OTUs identified as enriched in COPD animals and the other includes OTUs enriched in non-COPD animals. Each node represents an OTU and is colored by its assigned taxonomy. Green edges represent positive correlation between OTUs and red represent negative correlation. In the insert, negative edges have been removed to show that the two groups have no positive correlations with each other. Figure from (107)**.**

Evidence of community stability, or robustness to perturbation, can be seen in the level of connectedness of these interaction networks. If the interaction network is considered to be a system for passing metabolites as messages, and the network has a scale-free topology, then a more connected network is a more stable network (100, 108). Network and community stability

become important in microbiomes when species are wiped out by antibiotics, or by other means (109).

While many methods have been developed to infer interactions between microbes based on TAS, these inferences are often based on co-occurrence or correlations (99, 101, 110, 111). These methods were developed on (and, to our knowledge, have only been applied to) bacterial communities. Bacterial microbiome studies rely on TAS of the gene encoding the 16S subunit of the ribosomal RNA gene (16S rRNA gene). By targeting the 16S rRNA gene sequence, these studies ignore other important components of the community, such as viruses and eukaryotes, including fungi. Although present at significantly lower levels than bacteria, fungi play an important role in the microbial community and interactions between individual fungi and bacteria are well-documented (99, 112, 113), making these interactions of relevance for further study (72).

Here we present a statistically sound method for investigating cross-domain interactions, apply this method to the human lung and skin microbial communities, and validate three predicted interactions, including two cross-domain interactions. Sparse InversE Covariance estimation for Ecological Association Inference (SPIEC-EASI, pronounced “speak-easy”) identifies networks of interactions based on TAS data from a single domain (104). The included centered log ratio (CLR) transformation was designed specifically for the compositional nature of the relative abundances. By applying the CLR transformation separately to the independent compositions of bacteria and fungi, we maintain the statistically sound properties of the transformation. By using independent TAS studies of the 16S rRNA and Internal Transcribed Spacer (ITS) from the same samples, SPIEC-EASI is able to infer both within domain and cross-domain interactions.

In order to identify interactions that may be exploited in the future, we applied SPIEC-EASI to two microbiome studies that include both bacterial 16S rRNA and fungal ITS sequence data: the lung microbiome from the Pittsburgh cohort of the Lung HIV Microbiome Project (4, 40), and the skin microbiome (114, 115). We then validated by co-culture a subset of three predicted interactions from the skin microbiome, including two cross-domain interactions.

3.2 RESULTS

To highlight the variety and impact of cross-domain interactions on community stability, we analyzed the interaction networks of two microbiome communities. The first community was the lung microbiome from the Pittsburgh cohort of the Lung HIV Microbiome Project (4, 40). The cohort included both HIV-positive and HIV-negative individuals as well as individuals with normal lung function or chronic obstructive pulmonary disease (COPD). The cohort consisted of 25 individuals with a total of 35 bronchoalveolar lavage (BAL) samples. The second community was the skin microbiome from a National Human Genome Research Institute study (114, 115). This cohort consisted of 10 healthy individuals from whom 382 skin swab or nail clipping samples from 14 body sites were obtained. The sites were classified by the body region from which they originate (head, torso, arm, or foot) and also by what type of environment was present at the site (dry, moist, or sebaceous). Using SPIEC-EASI, we created three ecological networks for each microbiome: one of bacteria only, one of fungi only, and one of the combination of bacteria and fungi. We compared the connectedness, distances between nodes, and robustness of the three networks.

3.2.1 Lung Microbiome

In the "bacteria only" network derived from the lung microbiome dataset, we observed a network with a dense central cluster of well-connected nodes and several less-connected nodes on the periphery (**Figure 3.2A**). The majority of the operational taxonomic units (OTUs) (99.01%) created a single connected component with only 3 out of 302 (0.99%) OTUs with no connections to the main graph. This topography led to an average degree of nodes (number of adjacent edges) of 15.75 (SD: 10.70). There was also a high degree of assortativity, or clustering, by phyla, with the nominal assortativity coefficient (116) of the network by phyla measured at 0.518. Even the unclassified bacterial OTUs clustered together, with 61/66 (92.94%) forming a connected subcomponent with distances of no more than 3 edges between nodes. Overall, the network was highly connected, with an average normalized node betweenness centrality (a measure of the number of shortest paths through the node, where a lower number means a more connected network) of 0.007 (SD 0.010).

In the "fungi only" network of the lung microbiome, we saw a largely banded network that appears sparser than the bacteria only network (**Figure 3.2B**). This network contained one large connected component (83.33% of OTUs), 4 dyads (8.33%), and 8 singletons (8.33%). There was minimal assortativity within this network, with the nominal assortativity coefficient measuring 0.216. The low assortativity may be due to the low average degree of nodes (3.46; SD: 2.35). Yet the network remained well connected with an average normalized node betweenness centrality of 0.027 (SD: 0.032).

Surprisingly, the combined bacteria and fungi analysis of the lung microbiome provided a network that appeared more connected than either bacteria or fungi alone (**Figure 3.2C**). Only one fungal OTU out of 370 nodes (0.27%) remained outside the connected component. This

OTU was identified as *Candida dubliniesis* and in the "fungi only" network it was only connected to the fungus *Plicaturopsis crispa*, which did not appear in the combined dataset. The edges across domains, between bacteria and fungi, resulted in a higher average degree of the nodes (16.12; SD: 9.78) than either the "bacteria only" or "fungi only" networks. Assortativity of the combined network was somewhere between the "bacteria only" and "fungi only" networks, with the nominal assortativity coefficient measuring 0.320, and the bacteria forming a central cluster and the fungi being more peripheral. Within the central cluster there was no distinct clustering as there was in the "bacteria only" network. Overall, the network was very highly connected with an average normalized node betweenness centrally of 0.005 (SD: 0.005).

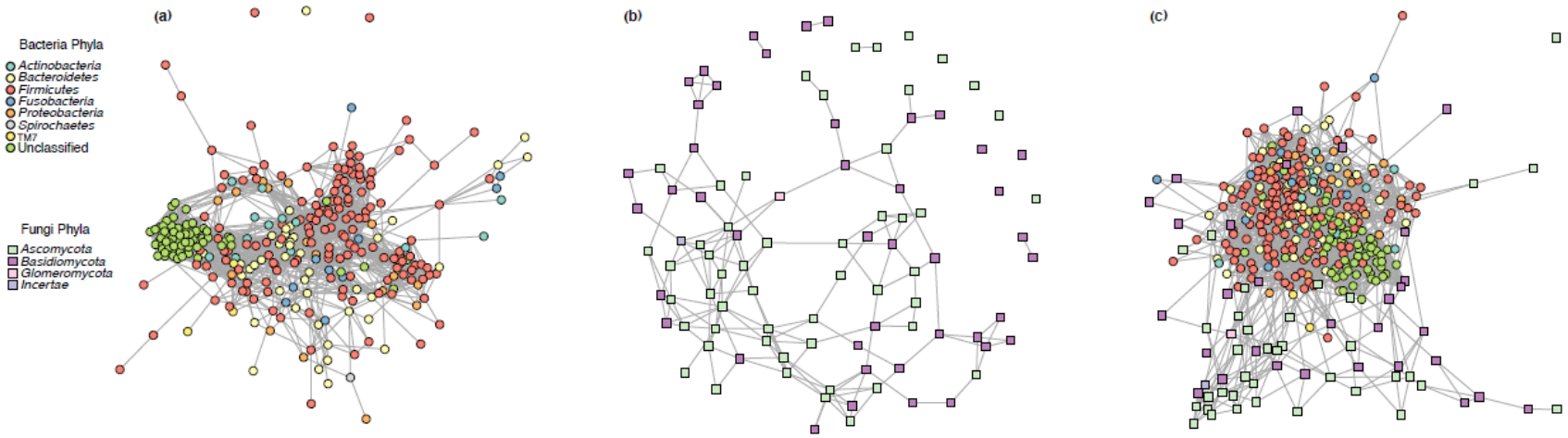

Figure 3.2 Lung microbiome networks. Networks inferred for the lung microbiome based on (A) bacteria only, (B) fungi only, and (C) combination of bacteria and fungi. In all three networks, bacterial nodes are circles and fungal nodes are squares. Each node is colored by phyla. Edges between nodes represented a predicted interaction, either positive or negative.

We then compared the combined network to the domain specific networks. We found that the distances between given node pairs were significantly shorter in the combined network (mean: 2.588; SD: 0.734 between bacterial nodes and mean: 3.604; SD: 1.508 between fungal nodes) than in the "bacteria only" (mean: 3.176; SD: 1.094, Welch t-test $p<0.0001$) or "fungi only" (mean: 4.549; SD: 2.203; Welch t-test $p<0.0001$) networks. Similarly, the node betweenness centrality for the bacterial nodes in the combined network was significantly lower than in the "bacteria only" network (Welch t-test $p=0.003$) while the decrease in node betweenness centrality for fungal nodes approached significance (Welch t-test $p=0.057$). This increased connectivity resulted in a larger percentage of nodes contained in the connected component (99.73% for the combined network vs 99.01% in "bacteria only" and 83.33% in "fungi only"). We measured the robustness of the networks by sequentially removing nodes and measuring the percent of the remaining nodes in the largest connected component (**Figure 3.3**). Nodes were removed in order of decreasing betweenness (**Figure 3.3A**), in order of decreasing degree (**Figure 3.3B),** or at random (**Figure 3.3C**), and in each case, the combined network was found to be slightly more robust than the "bacteria only" and greatly more robust than the "fungi only" network.

To determine the influence of HIV infection and COPD on our network, we examined the nodes present in only HIV-infected or HIV-uninfected individuals and those nodes present in only individuals that were COPD positive or those with normal lung function. For HIV status, 17 fungal nodes were present only in HIV-infected individuals while 1 bacterial and 5 fungal nodes were present only in HIV-uninfected individuals. All single HIV infection status nodes occurred around the periphery of the combined network and had a significantly lower average normalized node betweenness centrality (HIV-infected and HIV-uninfected exclusive nodes: 0.004, SD:

0.003 vs non-exclusive nodes: 0.005, SD: 0.005, Welch t-test p=0.011). For COPD status, 7 fungal nodes were present only in individuals with COPD while 8 bacterial and 10 fungal nodes were present only in individuals with normal lung function. Similar to the single HIV infection status nodes, the single COPD status nodes were located on the periphery with significantly lower average normalized node betweenness centrality (COPD positive and normal lung function exclusive nodes: 0.003, SD: 0.003 vs non-exclusive nodes: 0.005, SD: 0.004, Welch t-test p=0.006). All of these nodes were therefore unlikely to impact the connectedness or robustness of the network. We then examined each of the 4 "neighborhoods" surrounding the exclusive nodes, which consist of the exclusive nodes and their immediate neighbors (**Figure 3.4**). Comparing these neighborhoods, we saw that the HIV-infected neighborhood, containing 17 single-status nodes and 51 adjacent nodes, was larger and more connected than the HIV-uninfected neighborhood which had 6 single-status nodes and 32 adjacent nodes (**Figure 3.4A**). In contrast, it was the normal lung function neighborhood that was larger and more connected than the COPD positive neighborhood, 17 normal lung function nodes and 51 adjacent nodes compared to 7 COPD positive nodes and 33 adjacent nodes (**Figure 3.4B**).

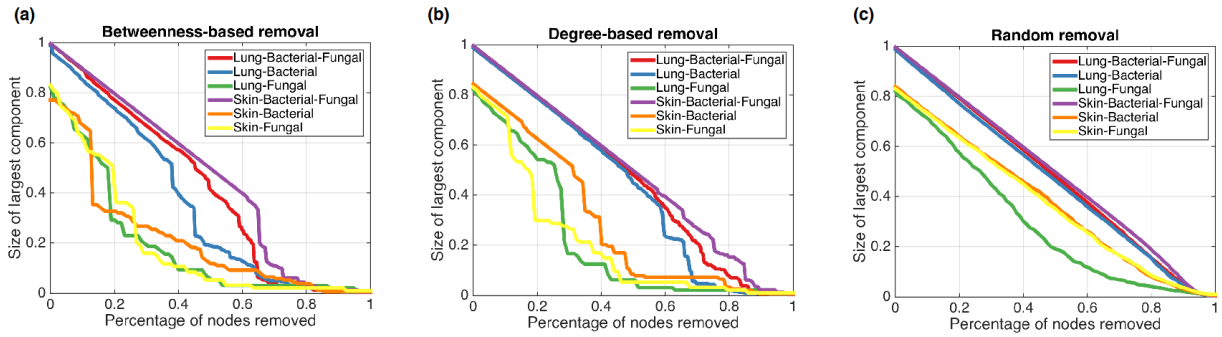

Figure 3.3 Robustness curves for all networks. Robustness of a network is measured by sequentially removing nodes based on the node's (A) betweenness, (B) degree, or (C) randomly selected and measuring the percentage of nodes that remain in the central connected component. Measurement of robustness was performed for each of our 6 networks and the results are plotted here with the percentage of nodes removed on the X axis and the percentage of remaining nodes in the central connected component on the Y axis. Each network is represented by a line on this graph. A fully connected, completely robust network would be a horizontal line at 1; the closer a line is to this horizontal, the more robust the network is.

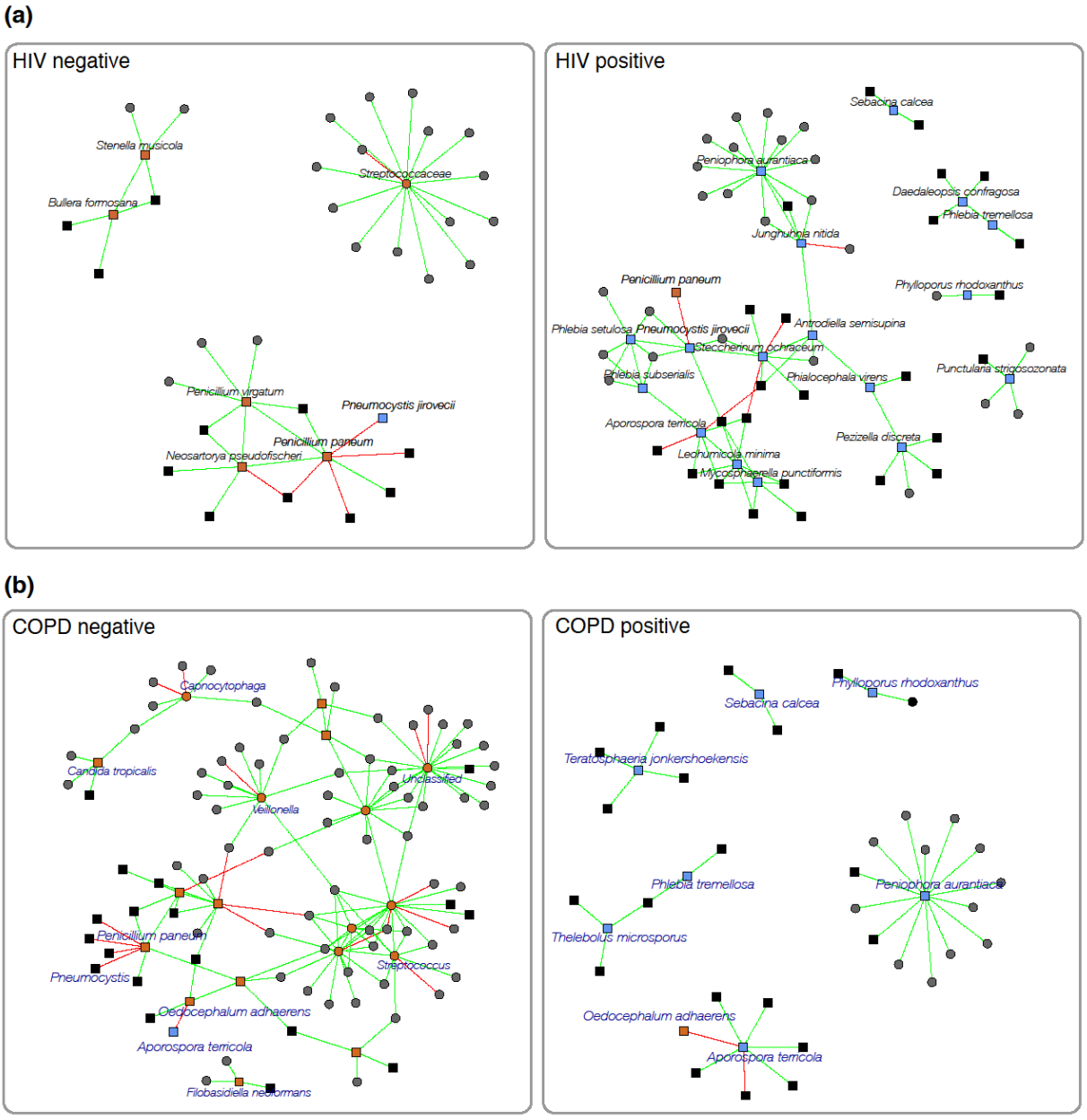

Figure 3.4 Lung microbiome neighborhoods for HIV infection and COPD status. Bacterial and fungal OTUs that occurred exclusively in (A) HIV-infected or HIV-uninfected, or (B) COPD negative or COPD positive samples and their adjacent nodes were isolated from the combined lung microbiome network (Figure 3.2C). Round nodes represent bacteria and square nodes represent fungi; green edges represent positive interactions and red edges represent negative interactions. The single-status nodes are colored orange if they are exclusive to the negative status and blue if they are exclusive to the positive status; nodes that are not exclusive are black or dark grey. Single-status nodes are labeled with their genus or species.

3.2.2 Skin Microbiome

In the “bacteria only” network of the skin microbiome, we saw a dense network similar to the “bacteria only” network in the lung microbiome (**Figure 3.5A**). The majority of OTUs, 130 out of 153 (84.97%), were in one large connected component, leaving 18 (11.76%) in a small connected component, and 5 (3.27%) singletons completely disconnected from the rest of the network. The resulting graph had an average degree of the nodes of 11.37 (SD: 7.63). Unlike the lung microbiome “bacteria only” network, there was minimal assortativity in the “bacteria only” network of the skin microbiome to the point of a negative nominal assortativity coefficient of -0.017. However, the network remained highly connected with a normalized betweenness centrality of 0.011 (SD: 0.021).

In the “fungi only” network of the skin microbiome, the network consisted of one large connected component containing 79 out of 94 (84.04%) nodes, a quintet (5.32%), a dyad (2.13%), and 8 singletons (8.51%) (**Figure 3.5B**). This topography resulted in a lower average degree of the nodes of 7.51 (SD: 6.40). Similar to the “fungi only” network for the lung microbiome, there was no obvious visual assortativity among the fungal phyla in the skin microbiome, although the nominal assortativity coefficient measured 0.438. Yet the network remained highly connected with a normalized node betweenness centrality of 0.014 (SD: 0.026).

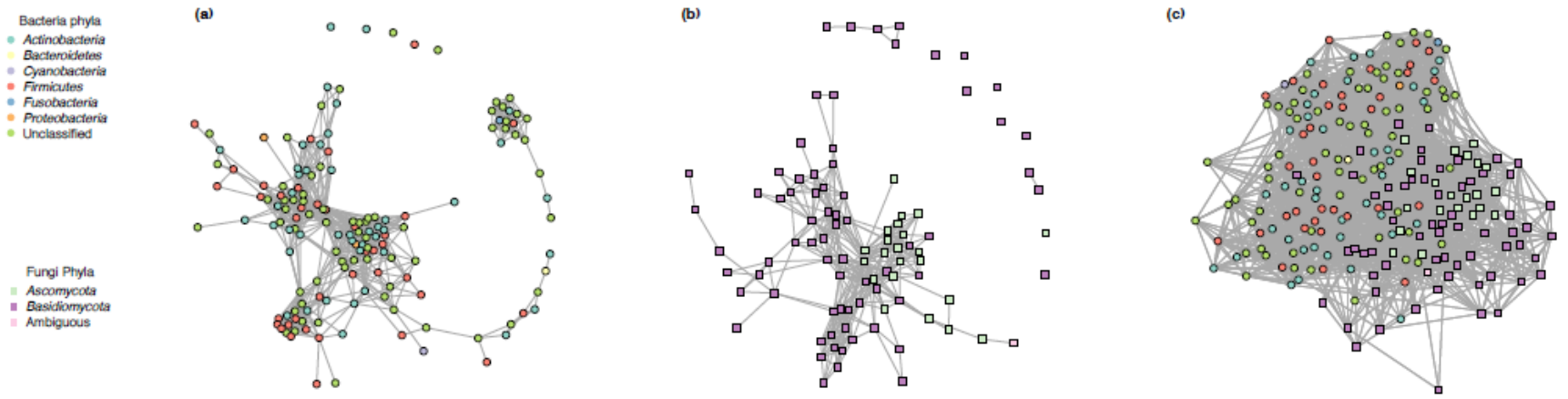

Figure 3.5 Skin microbiome networks. Networks inferred for the skin microbiome based on (A) bacteria only, (B) fungi only, and (C) combination of bacteria and fungi. In all three networks, bacterial nodes are circles and fungal nodes are squares. Each node is colored by phyla. Edges between nodes represented a predicted interaction, either positive or negative.

In the combined bacteria and fungi network of the skin microbiome, we saw our only fully connected network, indicating the inclusion of many cross-domain edges (**Figure 3.5C**). This topography resulted in a very high average degree of the nodes of 40.03 (SD: 13.77), and a low normalized node betweenness centrality of 0.006 (SD: 0.005). Assortativity remained relatively low in the combined network, with the nominal assortativity coefficient measuring 0.170, although visualization hinted at a separation between the bacteria and fungi.

When we compared the combined network to the domain specific networks, we found that it was the most connected of the three networks. A larger percentage of the nodes were contained in the connected component, but the node betweennness centrality was not significantly lower than the "bacteria only" (Welch t-test $p=0.111$) even though it was significantly lower than the "fungi only" networks (Welch t-test $p=0.004$). The reduction in distance between two connected nodes was highly significant in both the bacteria (from 3.05; SD: 1.31 in the "bacteria only" network to 2.10; SD: 0.67; Welch t-test $p<0.0001$) and fungi (from 2.75; SD: 1.23 in the "fungi only" network to 2.12; SD: 0.73; Welch t-test $p<0.0001$). The combined network for the skin microbiome was the most robust network of our study, regardless of the method used to select nodes for removal. The combined network was much more robust than either the "bacteria only" or "fungi only" networks for the skin microbiome and even more robust than the combined network for the lung microbiome (**Figure 3.3A-3.3C**).

To rule out any impact of the wide variety of skin sampling sites, we looked at nodes by sampling region (head, torso, arm, and foot) and by location physiology (dry, moist, and sebaceous). All nodes were found in at least 3 of the regions and across all physiologies, so sampling site had no impact on the overall network or the interactions within.

3.2.3 Co-culture Validation

To validate interactions across domains without support in the literature, we looked for a maximal clique (a fully connected component to which no other nodes can be added and still be fully connected) in the combined network of the skin microbiome that was limited to 3 nodes (281 cliques) and included at least 1 bacterium and 1 fungus (143 cliques). We further limited ourselves to medically relevant fungi that could be commercially obtained and to bacteria that could be identified to the species level and were commercially available. Seven cliques remained. Of these remaining cliques, 1 clique contained non-pathogenic microbes that grow on brain heart infusion (BHI). This clique contained a positive predicted interaction between the fungus *Emericella nidulans* (also called *Aspergillus nidulans*) and the bacterium *Propionibacterium acnes*. The third member of the clique was *Rothia dentocariosa*, which was predicted to have negative interactions with both *E. nidulans* and *P. acnes* (**Figure 3.6A**).

To measure growth, we established and compared growth curves for each bacterial species under uniform conditions (see methods), in pairs, and finally as a trio. Of the three predicted interactions being tested in duo-cultures, we were able to confirm two. As predicted, *P. acnes* appeared to grow better for the first 72 hours when in the presence of *E. nidulans* than when grown in monoculture. However, a Kolmogorov-Smirnov (KS) test revealed that the two growth curves were not significantly different (p=0.211; **Figure 3.6B**). As also predicted, *R. dentocariosa* grew significantly worse in the presence of *E. nidulans* than when grown in monoculture (KS p=0.003; **Figure 3.6C**). While a negative interaction was predicted between *P. acnes* and *R. dentocariosa*, the growth curve of the duo-culture was slightly above that of *R. dentocariosa* in monoculture or the sum of the two monocultures, but it was not significantly greater (KS p=0.299 and p=0.591, respectively; **Figure 3.6D**). Finally, we looked at bacterial

growth in a tri-culture with all three microbes. The overall effects were not predicted explicitly by the network, but we expected a negative effect on bacterial growth due to the greater number of negative interactions. The bacterial growth of the tri-culture curve could not be distinguished from either *R. dentocariosa* or *P. acnes* in monoculture (KS p=0.228 and p=0.925, respectively; **Figure 3.6E**). However, our bacterial growth curve measurements could not distinguish between *R. dentocariosa* and *P. acnes*, meaning that this curve could represent growth of one species and the complete death of the other.

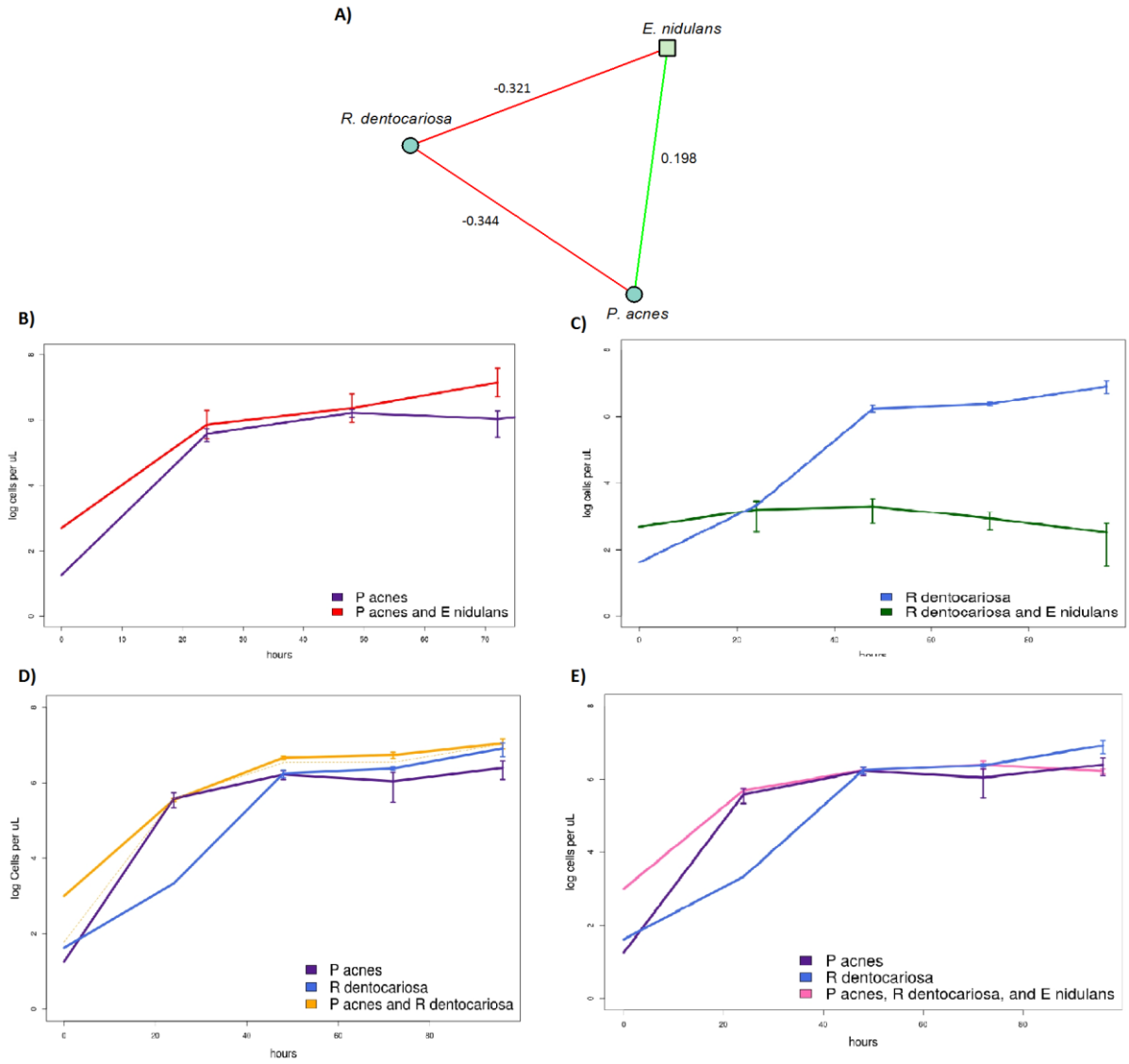

Figure 3.6 Growth curves for co-culture validation experiment. A maximal clique (A) was identified in the combined skin microbiome network that included a positive interaction (shown in green) between *Emericella nidulans* and *Propionibacterium acnes* and negative interactions (shown in red) between *Rothia dentocariosa* and both *E. nidulans* and *P. acnes*. The interaction edges are labeled with the optimal covariance between the two nodes. The microbes were grown in pairs and a trio, and the growth curves for the bacteria were compared to when they were grown in monoculture. Growth curves are based on the average of 3 biological replicates and the vertical lines indicate their standard deviations. (B) *P. acnes* grown with *E. nidulans* (red line) or alone (purple line); (C) *R. dentocariosa* grown with *E. nidulans* (green line) or alone (blue line); (D) *R. dentocariosa* grown with *P. acnes* (orange line) or alone (blue line); the sum of R. dentocariosa and P. acnes

monocultures is represented by the dashed orange line; (E) trio of all three organisms grown together (pink line) compared to *R. dentocariosa* alone (blue line) or *P. acnes* alone (pink line).

3.3 DISCUSSION

Based on SPIEC-EASI, we modified the CLR to investigate cross-domain interactions then applied this method to create three ecological networks each for the lung and skin microbiomes: one of bacteria only, one of fungi only, and one of the combination of bacteria and fungi. In the lung microbiome, we found all three networks to be well connected but the network that included both bacteria and fungi was the most well connected and robust. From this network we were able to isolate interactions specific to HIV infection and COPD. We found that the HIV-infected neighborhood was larger and more connected than the HIV-uninfected neighborhood, in part due to a higher number of exclusive nodes and to an increase in the number of fungal taxa in the lungs. If the exclusive nodes and their interactions build over time, it could be indicative of a fungal succession that occurs following HIV infection. In contrast, the neighborhood associated with normal lung function was more connected than the COPD neighborhood, indicating that several core interactions are lost when COPD develops.

Similar to the lung microbiome, all three ecological networks from the skin microbiome were well connected, but the network that included both the bacteria and fungi was the most connected and robust. From this network, we isolated a clique containing one model fungus and two common bacteria that could be cultured under the same conditions. By co-culturing the bacteria and fungus we saw growth curves in line with two of the three predicted interactions: a positive interaction between *E. nidulans* and *P. acnes*, and a negative interaction between *E.*

nidulans and *R. dentocariosa*. This represents the largest culture-based validation to date of microbial interactions predicted computationally.

In both microbiome communities, we saw increased connectivity in the combined networks. This increase in connectivity indicates cross-domain interactions between the bacteria and fungi. Cross-domain interactions made up 135 out of 2982 (4.53%) of all interactions in the lung microbiome and 480 out of 2292 (20.94%) of all interactions in the skin microbiome. The greater percentage of cross-domain interactions in the skin may be driven by the higher biomass located there or by the greater overlap of OTUs between the skin samples.

While many interactions have been previously identified between bacteria and common fungi, such as *Candida*, they appear to be ecosystem dependent. In supragingival plaque, *Candida albicans* interacts with *Streptococcus* to form "corncob" structures (117). In contrast, in the human gut microbiome, *Candida* correlates with *Prevotella* and *Ruminoccus* species (113) with no correlation with *Streptococcus*. Although *Candida* species and all 3 of these bacterial genera were present in the lung and skin microbiomes, we did not see evidence for these interactions in our datasets. Instead, we observed interactions in the lung microbiome between *Candida tropicalis* and *Capnocytophaga*, *Veillonella*, and *Streptococcus* and between *Candida parapsilosis* and *Neisseria* and an unclassifiable member of the Bacteroidales order. In the skin microbiome, the only *Candida* species detected, *Candida parapsilosis*, had 8 cross-domain interactions (out of 18 interactions): *Rothia dentocariosa*, *Propionibacterium granulosum*, *Streptococcus* spp, and 5 unclassified OTUs. Because of the ecosystem specific interactions seen elsewhere, it is not surprising that we identified different patterns of *Candida*-bacteria interactions in the lung and skin microbiomes.

Candida and other model fungi, including *E. nidulans*, have been studied in co-culture with bacteria in the laboratory to induce properties not produced in mono-cultures. Direct contact between *C. albicans* and *Fusobacterium nucleatum*, both oral microbiome commensals, leads to mutual attenuation of virulence, preventing *C. albicans* from transitioning to its pathogenic hyphal phase (118). Direct contact with the bacterium *Streptomyctes hygroscopicus* is required for *E. nidulans* to produce secondary metabolites, including polyketide synthase, often seen in nature, but not in the laboratory (119). These two model organisms highlight the variety of cross-domain interactions that can and do occur in microbiomes. However, most co-culture experiments, including those between *C. albicans* and *F. nucleatum* or between *E. nidulans* and *S. hygroscopicus*, originate from the knowledge that the common fungus grows in physical proximity to the bacteria rather than from computationally-predicted community ecological networks.

Interactions between bacteria and less common fungal species are more difficult to identify. Similar to the notion that only 1% of known bacteria can be grown in the laboratory (73, 74), fewer than 17% of known fungi are considered culturable, representing fewer than 1% of the estimated global fungal species (75). The global species estimate is based largely on the ratio of vascular plants to fungi, which demonstrates how little is known about fungal diversity. If a fungus is unknown or understudied, then little to nothing is known about its cross-domain interactions.

We have shown here that cross-domain interactions can be inferred computationally and in a statistically sound manner using SPIEC-EASI. As validation of some of the inferred interactions, we co-cultured a small subset of microbes with positive and negative predicted interactions. We considered this co-culture experiment to be a basic proof of principle; therefore,

it had severe limitations. We limited ourselves to commercially available strains of aerobic and aerotolerant species, which may or may not be representative of what was present on an individual's skin. The choices we made in setting up our experiments represent an oversimplification of the community: we seeded the cultures with approximately equal numbers of bacteria and an order of magnitude lower of fungi, which was not representative of their relative abundance in our samples. We also seeded all three organisms at the same time, but it is highly likely that colonization occurs in stages and not concurrently. We made no attempt to identify mechanisms such as shared metabolites and did not include or attempt to mimic the human skin on which these microbes would normally interact. Despite these limitations, the co-culture serves as a first step towards validating the interactions inferred with SPIEC-EASI and do indeed demonstrate the positive or negative effects these microbes have on each other's growth. These limitations highlight how complicated microbial interactions are likely to be and demonstrate how a tool such as SPIEC-EASI can help infer some of these interactions and provide biological insight.

In summary, we have devised a statistically sound method for predicting cross-domain interactions, applied this method to two human-associated microbiome datasets, and validated a subset of the predicted interactions. From these results, we can conclude that limiting studies of ecological interaction networks to a single domain fails to reveal the entirety and robustness of the network. In the future, we expect to see this approach being used to incorporate protists, archaea, and even viruses as well as an increase in culture-based validations and searches for interaction mechanisms of computationally predicted interactions.

3.4 METHODS

3.4.1 Adapting SPIEC-EASI for Two Domains

We adapted the SPIEC-EASI method to analyze microbiome networks across multiple microbial domains (104). The tables of absolute abundance of bacteria and fungi OTUs are stored in matrices $W \in \mathbb{N}_0^{n \times d}$ and $V \in \mathbb{N}_0^{n \times p}$, where $w^{(j)} = \left\{w_1^{(j)}, w_2^{(j)}, \dots, w_d^{(j)}\right\}$ and $v^{(j)} = \left\{v_1^{(j)}, v_2^{(j)}, \dots, v_p^{(j)}\right\}$ denote the *d*- and *p*-length row vectors of counts from the *j*th sample, and $\mathbb{N}_0$ denotes the set of natural numbers. We define the total cumulative counts for each domain as $M_{(j)} = \sum_{i=1}^{d} w_i^{(j)}$ and $N_{(j)} = \sum_{i=1}^{p} v_i^{(j)}$.

In a standard sequencing experiment, the true count data $w^{(j)}$ and $v^{(j)}$ are unknown, since absolute abundance information is not available. However, by dividing observed sequencing counts by the total library size, we get compositional data vectors, $x^{(j)} = \left\{x_1^{(j)}, x_2^{(j)}, \dots, x_d^{(j)}\right\}$ and $y^{(j)} = \left\{y_1^{(j)}, y_2^{(j)}, \dots, y_p^{(j)}\right\}$, with associated relative abundance matrices $X \in \mathbb{S}^{d \times n}$ and $Y \in \mathbb{S}^{p \times n}$, where $\mathbb{S}^{p \times n} = \left\{x \middle| x_i > 0, \sum_{i=1}^{p} x_i = 1\right\}$ is the *p*-dimensional unit simplex. It is well known that components of a composition are not independent due to the unit sum constraint, and covariance matrices of compositional data exhibit negative bias due to closure. It follows that, compositional data can be completely determined by the absolute abundance data it was

generated from (termed a basis), i.e. dividing by the total library size $x^{(j)} = \left\{x_1^{(j)}/M^{(j)}, x_2^{(j)}/M^{(j)}, \ldots, x_d^{(j)}/M^{(j)}\right\}$.

As noted by John Aitchison, the equivalence

$$\log\left[\frac{x_i}{x_j}\right] = \log\left[\frac{w_i/M}{w_j/M}\right] = \log\left[\frac{w_i}{w_j}\right], \qquad (3.1)$$

implies that statistical inferences drawn from the analysis of the log-ratios of compositions ($\log\left[\frac{x_i}{x_j}\right]$) are equivalent to those drawn from analysis of log-ratios of the basis components ($\log\left[\frac{w_i/M}{w_j/M}\right]$). This equivalence establishes the precedence of log-ratio transformations to study compositional data. The centered log-ratio (CLR) transformation,

$$\mathrm{CLR}(x) = \left\{\log\left[\frac{x_i}{g(x)}\right], \ldots, \log\left[\frac{x_d}{g(x)}\right]\right\}, \text{ where } g(x) = \left[\prod_{i=1}^{d} x_i\right]^{1/d},$$

is particularly useful, as it is symmetric and isometric (equal in dimension) with respect to the original composition (x). The CLR maps compositional data from the d-dimensional unit simplex to a $(d-1)$-hyperplane of d-dimensional Euclidean space. This mapping also applies to the population covariance matrix such that $\Gamma_X = \mathrm{Cov}[\mathrm{CLR}(X)]$. The matrix Γ_X is related to the population covariance of the log-transformed absolute abundances, $\Omega_W = \mathrm{Cov}[\log W]$ by

$$\Gamma_X = G^d \Omega_W G^d, \qquad (3.2)$$

where $G^d = I^d - \frac{1}{d}\mathbf{1}\mathbf{1}^T$, is the standard centering matrix, where I^d is the $d \times d$ identity matrix and $\mathbf{1}$ is a d-length vector of ones. Therefore, for high dimensional data, $d \gg 4$,

$$G^d \approx I^d, \qquad (3.3)$$

and $\Gamma_X \approx \Omega_W$ is a reasonable approximation. The sparsity conditions necessary to approximately identify the covariance structure Ω_W from Γ_X, have recently been shown, and have recovery guarantees based on sparsity, dimensionality, and sample size (120).

The equivalence in **Equation 3.1** and the ability to identify the population covariance structure from log-transformed absolute abundance are the foundation of SPIEC-EASI, which seeks to estimate a sparse inverse covariance (precision) matrix using the population covariance matrix as input. SPIEC-EASI uses the glasso method to solve the optimization problem,

$$\hat{\Omega}_W^{-1} = \underset{\hat{\Omega}_W^{-1} \in PD}{\operatorname{argmin}} - \log\det\left(\hat{\Omega}_W^{-1}\right) + \operatorname{tr}\left(\hat{\Omega}_W^{-1}\hat{\Gamma}x\right) + \lambda\left\|\hat{\Omega}_W^{-1}\right\|_1 \qquad (3.4)$$

where $\hat{\Gamma}_X$ is the empirical covariance estimate of $\mathrm{CLR}(X)$ and PD is the set of all positive definite matrices. Solving E**quation 3.4** ensures that the penalized estimator is full rank, with a sparsity pattern that depends on the value of λ, since the L1 norm, $\|\cdot\|_1$, penalizes the absolute values of the row sums of the symmetric inverse covariance matrix ($\hat{\Omega}_W^{-1}$). In the Gaussian case, the network, or graphical model, is specified from the non-zero entries of $\hat{\Omega}_W^{-1}$.

It is apparent that because they are amplifying and sequencing different marker genes that do not compete for reads, cross-domain studies generate technically independent compositions. Therefore, a naive application of **Equation 3.4** directly to the combined dataset $[X\ Y]$, an $n \times (d + p)$ matrix generated from a simple concatenation of two compositional datasets, would be inappropriate.

To illustrate this, consider that the log-ratio

$$\log\left[\frac{x_i}{y_j}\right] = \log\left[\frac{w_i/M}{v_j/N}\right] \neq \log\left[\frac{w_i}{v_j}\right] \qquad (3.5)$$

does not satisfy the scale-invariance property of **Equation 3.1**. Similarly, **Approximation 3.3** does not hold between cross-compositional pairs.

We instead consider the data matrix $Z = [\ \mathrm{CLR}(X)\ \mathrm{CLR}(Y)\ ]$, generated by concatenating independently transformed compositions. The matrix $\Gamma_Z = \mathrm{Cov}[Z]$ now has the following relation to the basis covariances:

$$\Gamma_Z = \begin{bmatrix} G^d\Omega_W G^d & G^d\Omega_{WV} G^p \\ G^p\Omega_{VW} G^d & G^p\Omega_V G^p \end{bmatrix} \qquad (3.6)$$

where $\Omega_{WV} = \mathrm{Cov}[\log W, \log V\ ]$, the cross-covariance matrix between the two log-transformed basis datasets, and $\Omega_{VW} = (\Omega_{WV})^T$. In other words, the $(d + p) \times (d + p)$ combined covariance structure Γ_Z is decomposable into blocks where A**pproximation 3.3** holds. If $p, d \gg 4$, then the approximation

$$\Gamma_Z \approx \Omega_Z = \begin{bmatrix} \Omega_W & \Omega_{WV} \\ \Omega_{VW} & \Omega_V \end{bmatrix} \qquad (3.7)$$

allows us to use $\hat{\Gamma}_Z$ as the input to **Equation 3.4** to get a penalized estimator $\hat{\Omega}^{-1}_Z$. This estimate is interpretable as an intra- and cross-domain interaction network, using the standard SPIEC-EASI pipeline. Going beyond two domains follows directly from this and is left to the reader.

3.4.2 Datasets

In this study, we analyzed two previously published microbiota datasets that included both bacterial and fungal sequences. The first was from BALs collected as part of the Lung HIV Microbiome Project, as published in (4) and (40). It contained 35 samples that were subjected to

16S rRNA gene and ITS sequencing. The BAL samples originated from the right middle lobe or the left upper lobe of the lungs from 25 individuals, of whom 14 were HIV-infected and 11 were HIV-uninfected. Of the 35 samples, 17 came from individuals with normal spirometry and 18 from individuals with COPD (diffusing capacity of the lungs from carbon monoxide (DLCO) < 80% or forced expiratory volume in 1 second (FEV1) < 70%). The demographics of the cohort analyzed here can be found in **Table 3.1**. No significant differences were found between HIV-infected and HIV-uninfected or between individuals with COPD and those with normal lung function.

The second dataset was from a skin microbiome study at the National Human Genome Research Institute, as published in (114) and (115). It includes 382 samples from 14 body sites on 10 healthy adults. Ten body sites were repeated on the left and right sides, and some of the healthy volunteers underwent repeat sampling 1-3 months after their initial visits.

Table 3.1: Demographics of the lung microbiome cohort. Values are presented as mean (SD) except for those that are the percentage of the subset denoted with (%). P-values are from Welch t-tests for continuous variables and from Fisher's exact tests for percentages.

	Cohort	HIV+	HIV-	p-value	COPD+	COPD-	p-value
N	25	14	11	-	13	12	-
Age (yrs)	51.5 (7.7)	51.2 (8.3)	51.9 (7.4)	0.8472	49.4 (8.1)	53.6 (7.1)	0.2032
Male (%)	88.0	92.9	81.8	0.5648	92.3	83.3	0.5930
White (%)	56.0	50.0	63.6	0.6887	53.8	58.3	1.0000
Current Smokers (%)	20.0	28.6	9.1	0.4913*	30.8	8.3	0.4671*
Former Smokers (%)	12.0	14.3	9.1		7.7	16.7	
BMI	25.9 (5.3)	24.2 (4.2)	28.1 (5.9)	0.0792	24.4 (5.4)	27.6 (4.8)	0.1426
Viral Load	-	1476.9 (2849.5)	-	-	2053.5 (3230.5; N=10)	35.5 (18.2; N=4)	0.0746
CD4 count	-	645.7 (305.3)	-	-	620.2 (326.2; N=10)	701.8 (278.8; N=4)	0.6195

*Smoking status p-value calculated using an ANOVA test.

Table 3.1 Continued

FEV1/FVC	79.0 (11.5)	80.1 (8.8)	77.7 (14.6)	0.6435	75.8 (15.0)	82.5 (4.1)	0.1463
DLCO	77.2 (15.3)	73.3 (16.0)	82.1 (13.5)	0.1520	66.8 (13.9)	88.4 (6.0)	<0.0001

3.4.3 Sample and sequence processing

Sample processing procedures for the lung microbiome have been previously described (4, 40). In brief, all samples had DNA extracted using standard techniques of the PowerSoil® DNA Isolation Kit from MO BIO (Carlsbad, CA). For bacterial DNA sequencing, the hyper-variable regions 1 through 3 (V1-V3) were amplified and sequenced using the Roche 454 GS-FLX Titanium platform. For fungal DNA sequencing, the ITS1 was amplified and sequenced on the Ion PGM™ Sequencer using the 400 bp protocol (60).

The sample processing procedures for the skin microbiome were previously described (114, 115). In brief, samples were lysed using the MasterPure™ Yeast DNA Purification Kit, cell walls were mechanically disrupted using a Tissuelyser (Qiagen, Valencia, CA), and DNA was extracted using the Invitrogen PureLink Genomic DNA Kit (Invitrogen, Carlsbad, CA). For bacteria DNA sequencing, the V1-V3 regions were amplified and for fungal DNA sequencing the ITS1 region was amplified. Both bacterial and fungal DNA was sequenced on the Roche 454 GS20/FLX platform with Titanium chemistry (Roche, Branford, CT). We analyzed the resulting

sequences in a manner consistent with the lung microbiome, which was different from that used in the original publications.

All sequences from both the lung and skin microbiomes were processed using the QIIME pipeline version 1.7 (84) with default settings for de novo Operational Taxonomic Unit (OTU) picking at 97% similarity for bacteria and 99% similarity for fungi. Additional processing for the ITS sequences was performed using FHiTINGS (85). Samples with fewer than 1,000 16S bacterial reads (N=12 for the lung microbiome; N=12 for the skin microbiome) and samples with fewer than 50 ITS fungal reads (N=11 for the lung mycobiome; N=3 for the skin microbiome) were considered to have failed and were removed from further analysis. Bacterial taxonomic assignments were made using the Green Genes 12.10 reference database (62) and fungal taxonomic assignments were made using the FHiTINGS version of the Index Fungorum (http://www.indexfungorum.org/) reference database (85).

We removed OTUs present in fewer than 1/3 of the samples (20 lung samples or 120 skin samples) as well as any OTUs represented by single reads in every sample. The number of samples, bacterial and fungal OTUs of each resulting network dataset are presented in **Table 3.2**. A pseudo count of 1 read was added to every OTU in every sample to eliminate zeros in samples where OTUs were absent. All OTU counts were normalized using total sum scaling (also known as relative abundance) followed by centered log ratio scaling (121), as described above.

Table 3.2 Dataset sizes for each network constructed. Amplification of target genes and sequencing were not successful for all samples resulting in variable node counts in the combined networks.

Network	Samples	Bacteria OTU Nodes	Fungi OTU Nodes
Lung bacteria only	77	302	--
Lung fungi only	48	--	96
Lung combined	35	302	68
Skin bacteria only	360	153	--
Skin fungi only	375	--	94
Skin combined	353	144	85

3.4.4 Constructing Networks

All networks were constructed using the *SpiecEasi* package version 0.1 in R (https://github.com/zdk123/SpiecEasi). We used the glasso estimation method to build the initial networks and selected the optimal sparsity parameter based on the stability approach to regularization selection (STARS) criteria (122). The STARS variability threshold was set to 0.1 for all networks.

3.4.5 Evaluating and Comparing Networks

Networks were evaluated using functions of the R package *igraph* version 1.0.1 (123). We evaluated node degree (i.e. the count of edges a node has) as a measure of sparsity. A complete network would have an average node degree equal to the number of nodes minus 1; a lower degree indicates a more sparse network. We measured assortativity by phyla with the nominal assortativity coefficient, which is designed to measure clustering by categorical variables. Higher coefficients indicate more clustering within categories. To evaluate connectedness of the networks, we used normalized node betweenness centrality for undirected graphs. Normalized node betweenness centrality measures the proportion of the shortest paths in the network that pass through the node. A lower average betweenness centrality number indicates a more connected network, either because of more shortest-paths, or because fewer of the shortest paths travel through each node. These metrics, as well as distance between nodes, were used to compare the networks using Welch's unequal variances t-tests (124).

3.4.6 Microbial Co-cultures

All organisms were purchased from ATCC and grown under their recommended conditions (**Table 3.3**) to establish stocks. From these stocks, uniform condition stocks were inoculated in brain heart infusion (BHI) broth and incubated at 37°C under aerobic conditions. These same uniform conditions were used to grow mono, dual, and tri organism co-cultures, each started with the same number of cells of each organism (10 million bacterial cells or 1 million fungal spores). Growth was measured by cellular density every 24 hours for 5 days, and curves were fit by connecting the average of 3 biological replicates. To ensure that cells were maintaining

viability, aliquots were plated on BHI agar at each time point and colony forming units (CFUs) were counted after a 24-hour incubation period. A complete standard operating procedure for microbial co-cultures is located in **Appendix B**.

Table 3.3: Organisms and their recommended growing conditions. Each of the three microbes used in co-culture validation experiments was purchased from ATCC, rehydrated, and grown under their recommended conditions before co-culturing began. The ATCC catalogue numbers and recommended growing conditions are presented here. Because *P. acnes* is an anaerobe, it was first grown in a homemade anaerobic jar inside the incubator.

Organism	**ATCC catalogue number**	**Recommended Temperature**	**Recommended Media**
Emericella nidulans	96921	24°C	Malt extract agar
Propionibacterium acnes	6919	37°C	Tryptic soy agar with 5% sheep blood
Rothia dentocariosa	17931	37°C	Brain-heart infusion agar

3.4.7 Accession Numbers

The sequencing data from the lung microbiome used in this study are available in the Sequence Read Archive (SRA) under the following accession numbers: SRP065274 for 16S and SRP040237 for ITS. The sequencing data from the skin microbiome used here are available in GenBank under accession numbers GQ000001 to GQ116391 for 16S and KC669797 to KC675175 for ITS.

4.0 MULTI-OMICS INVESTIGATION OF THE LUNG MICROBIOME[6]

4.1 BACKGROUND

The composition of the lung microbial community differs under conditions of chronic obstructive pulmonary disease (COPD) (3, 23, 24, 125), cystic fibrosis (20), lung transplant (25), and other diseased states. Yet the functional role of the lung microbiome in health and disease has not yet been clearly defined. It has been hypothesized that the inflammation seen in COPD is a result of the host's immune response to the bacteria present in the lungs (22). This hypothesis is supported in part by the increased abundance of opportunistic pathogens during COPD (126, 127) but is hindered by inconsistent findings of shifts in the sputum microbiome community during COPD exacerbations. One study found increased abundance of COPD-related pathogens, including *Haemophilus influenzae* (127), while another study found no significant shifts in the community composition (128). These discrepancies highlight the limitations of taxonomy-based studies and the need to perform functional assessments of the microbial community.

[6] Paper in preparation.

One population where functional differences in the lung microbiome community may be especially relevant is the HIV-infected population. HIV is an independent risk factor for COPD, regardless of smoking history (129–131). Opportunistic pathogens associated with COPD may have more opportunities to infect patients with HIV given the subtle immune deficits observed, even in treated HIV infection. These subtle immune deficits have not been shown to have an impact on the lung microbiota (27) but this study did not include patients with impaired lung function.

To investigate the metabolism of the lung microbiome, we examined datasets from 16S rRNA target gene sequencing, metatranscriptome analysis, and metabolomic mass-to-charge ratio (*m/z*) features found in the bronchoalveolar lavage (BAL) of subjects with COPD and HIV. First, we looked at each dataset alone for overt differences. As observed in other studies, we found minimal differences in 16S rRNA gene-based community composition for both COPD and HIV (3, 23, 27). We also found minimal differences in gene family expression levels and global *m/z* feature abundance for both COPD and HIV. We looked at correlations between dataset pairs, and compared the information that could be learned from each data set independently, including taxonomic compositions and metabolic functions defined by Kyoto Encyclopedia of Genes and Genomes (KEGG) ontology (KO) terms (132, 133). While we found many differences in the taxonomic assignments and metabolic functions, we saw overlap in metabolic functions associated with HIV infection and COPD. Finally, we integrated all three datasets using a sparse multi-block partial least squares (sMBPLS) regression (134) to identify blocks of associated operational taxonomic units (OTUs) from the 16S rRNA gene sequences, gene families from the metatranscriptome sequences, and *m/z* features from the metabolomics. We examined each block for pathway enrichment among the *m/z* features and corresponding gene families in the same

pathways. Analysis of the OTUs associated with the enriched metabolic pathways indicated that important pathways were encoded and expressed by bacteria not considered high producers of the pathway products, or that certain pathways were completed across multiple bacterial species.

4.2 RESULTS

4.2.1 Single datasets

For each of our 25 samples, we have three "-omics" datasets: the 16S rRNA gene sequences, which were clustered into 1,142 OTUs and used for taxonomic community composition; the metatranscriptomic sequences, which were mapped to 145,574 UniRef50 gene families; and the metabolomic *m/z* features, of which there were 5,868. We tested each of the three datasets for differences in abundance or expression based on COPD status and HIV infection using Wilcoxon tests.

Among the OTUs, we found that the community composition could not distinguish between COPD and normal lung function, nor between uninfected and HIV infected patients (**Figure 4.1A**). These comparisons had adonis PERMANOVA p-values of 0.051 and 0.053, respectively. We tested each OTU for differential abundance and found 87 to be nominally differentially abundant by COPD status, and 57 to be nominally differentially abundant by HIV infection. No OTUs were significantly differentially abundant following correction for multiple hypotheses testing. Of the OTUs that were nominally differentially abundant by COPD, 10 were over-abundant and 48 were under-abundant in COPD patients, as measured by median relative abundance (**Table 4.1**). When comparing HIV infection, 56 OTUs were under-abundant in HIV

infection, based on median relative abundance (**Table 4.2**). The single OTU that was over-abundant in HIV infection is classified as a member of the S24-7 family, an uncultured member of the Bacterodales order predicted to thrive in low-oxygen environments including mammalian guts (135).

We were able to detect a significant difference in community composition between COPD and normal lung function (adonis p-value 0.030) when we removed samples dominated by environmental bacteria. Specifically, the six samples that contained more than 50% relative abundance of the family Microbacteriaceae were removed (samples 8, 9, 10, 14, 20, and 22 in **Figure 4.3**). This bacterial family has previously been found to be a contaminant in DNA extraction kits (58) and the over-abundance of this family in these samples lead to their exclusion. Of these six samples, five originated from participants with COPD and one from a participant with normal lung function. All six samples originated from participants that were HIV-infected but when they were removed there remained no significant difference in community composition between HIV-infected and uninfected samples (adonis p-value 0.027).

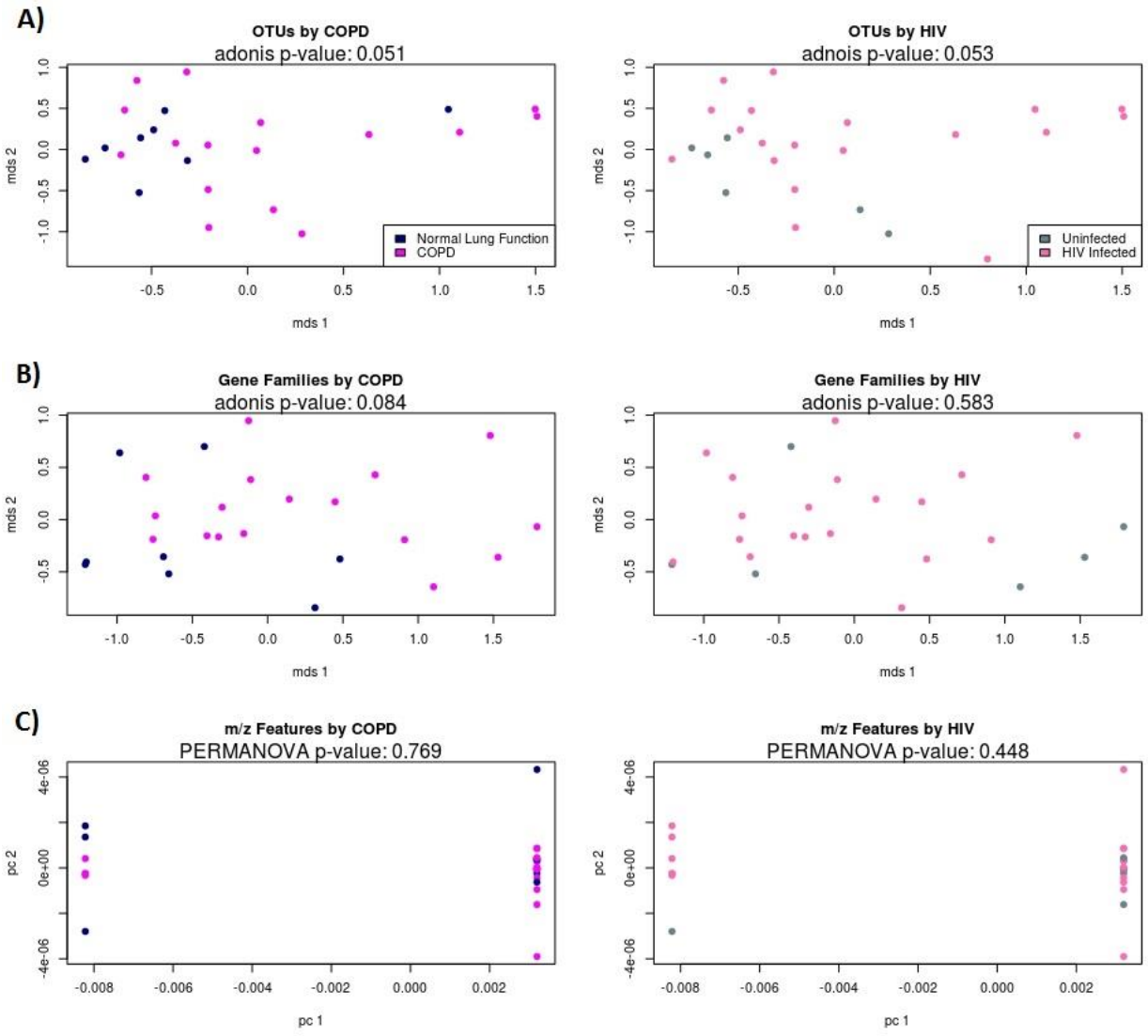

Figure 4.1 Ordination plots for COPD (left column) and HIV (right column) comparisons. (A) Principle Coordinates Analysis (PCoA) of OTUs based on Bray-Curtis distance. (B) PCoA of gene families based on Bray-Curtis distance. (C) Principle Component Analysis (PCA) of m/z features.

Table 4.1: List of OTUs differentially abundant in COPD by their lowest taxonomic assignments. OTUs with equal median relative abundance are not listed (N=29).

Taxonomy	**Number of OTUs**	**Over/Under Abundant in COPD**
Acinetobacter	1	Under
Bradyrhizobium	1	Under
Campylobacter	1	Under
Catonella	1	Under
Gemella	1	Under
Granulicatealla	1	Under
Fusobacterium	1	Under
Haemophilus	1	Under
Lactobacillus	1	Under
Leptotrichia	1	Under
Micrococcus	1	Over
Mycoplasma	1	Under
Neisseria	1	Under
Oribacterium	1	Under
Prevotella	2	Over
Prevotella	5	Under
Streptococcus	4	Under
Treponema	1	Under

Table 4.1 Continued

Trichoccus	1	Under
Comamonadaceae (family)	1	Under
Lachnospiraceae (family)	1	Under
Mogibacteriaceae (family)	1	Under
Neisseriaceae (family)	1	Over
S24-7 (family)	1	Under
Weeksellaceae (family)	1	Under
Bacillales (order)	1	Under
Clostridiales (order)	1	Under
Streptophyta (order)	1	Over
SR1 (phylum)	1	Over
Unclassified	4	Over
Unclassified	17	Under

Table 4.2: List of OTUs differentially abundant in HIV by their lowest taxonomic assignments.

Taxonomy	**Number of OTUs**	**Over/Under Abundant in HIV**
Actinomyces	1	Over
Bradyrhizobium	1	Over
Campylobacter	1	Over
Corynebacterium	1	Over
Enterococcus	1	Over
Fusobacterium	1	Over
Gemella	1	Over
Haemophilus	2	Over
Leptotrichia	1	Over
Micrococcus	1	Over
Moryella	1	Over
Neisseria	1	Over
Oribacterium	1	Over
Prevotella	4	Over
Rothia	1	Over
Streptococcus	7	Over
Trichococcus	1	Over
Aeromonadaceae (family)	1	Over
Comamondaceae (family)	1	Over

Table 4.2 Continued

Enterobacteriaceae (family)	1	Over
Lachnospiraceae (family)	1	Over
Mogibacteriaceae (family)	1	Over
Neisseriaceae (family)	1	Over
S24-7 (family)	1	Under
Bacillales (order)	1	Over
SR1 (phylum)	1	Over
Unassigned	21	Over

Similar to the OTU community abundance, from the metatranscriptome data we saw no difference in the composition of the gene families between COPD and normal lung function, nor between uninfected and HIV infected subjects (**Figure 4.1B**). These comparisons had adonis PERMANOVA p-values of 0.084 and 0.583, respectively. Each gene family was tested for differential expression in COPD and HIV infection. In COPD, 477 gene families were nominally significantly over-expressed and 9,230 gene families were nominally significantly under-expressed, based on median reads per kilobase (RPKs). Additionally, 9,598 gene families were considered nominally significantly different, but had the same median RPKs. In HIV infection, 499 gene families were nominally significantly over-expressed, 894 were nominally significantly under-expressed, and 312 gene families were considered significantly differentially expressed but had the same median RPKs. While these numbers may seem large in absolutes, the majority

of gene families being expressed are not significantly differentially expressed; 96.8% of gene families are equally expressed in COPD and normal lung function, and 99.8% are equally expressed in HIV-infected and uninfected subjects.

Finally, among the composition of all *m/z* features, we saw no significant difference in either COPD or HIV infection (**Figure 4.1C**). The PERMANOVA p-values for these comparisons were 0.448 and 0.051, respectively. In COPD, 88 *m/z* features were nominally significantly over-abundant and 80 *m/z* features were nominally significantly under-abundant, based on median values. None of these *m/z* features could be mapped to unique metabolites. Only two of the 5,868 *m/z* features could be mapped to unique metabolites: *m/z* 268.1907 mapped to 2-(3-AMINO-4-CYCLOHEXYL-2-HYDROXY-BUTYL)-PENT-4-YNOIC ACID, part of an archetypical dehalogenase, and *m/z* 999.3511 mapped to Sialyllacto-N-tetraose b, an oligosaccharide found in human breast milk. When we ran the *m/z* features that were nominally significantly different in COPD through *mummichog*, a pipeline that identifies pathway enrichment from *m/z* features bypassing assignments to unique metabolites (136), no pathways were found to have an overlap of more than two predicted metabolites. In HIV, 90 *m/z* features were nominally significantly over-abundant and 58 were nominally significantly under-abundant, based on median values. Again, none of these *m/z* features could be mapped to unique metabolites. When we ran the nominally significant *m/z* features through *mummichog*, the top pathways included valine, leucine, and isoleucine degradation (overlap of 5 out of 34 predicted metabolites, p-value = 0.0014), butanoate metabolism (overlap of 4 out of 23 predicted metabolites, p-value = 0.0015), keratin sulfate degradation (overlap of 2 out of 3 predicted metabolites, p-value = 0.0018), and tryptophan metabolism (overlap of 5 out of 39 predicted

metabolites, p-value = 0.0018). Although not as extreme as in COPD, the overlap of predicted metabolites is low in all pathways enriched in HIV-infection.

4.2.2 Two datasets

4.2.2.1 Correlations

For every pair of datasets, we calculated the Spearman (non-linear) correlations between all features at relevant abundances (greater than 0.1% average abundance). Overall, the correlations were skewed slightly positive (53.5%) but ranged from $\rho = -0.77$, between an OTU classified as *Prevotella* and the gene family A0A009RUM8: putative outer membrane protein (fragment), to $\rho = 0.81$ between an OTU classified as *Actinobacillus* and an *m/z* feature of 132.00, which could not be mapped to any unique metabolite.

Between OTUs and gene families, out of 3,440 correlations, 1,839 (53.5%) were positive and 1,601 (46.5%) were negative (**Figure 4.2 upper left**). Of these, 151 of the positive correlations and 295 of the negative correlations were statistically significant ($\rho > 0.4$ or < -0.4). Between OTUs and *m/z* features, out of 8,514 correlations, 4,359 (51.2%) were positive with 463 being significantly positive, and 4,155 (48.8%) were negative with 280 being significantly negative (**Figure 4.2 upper right**). This pair of datasets had some of the strongest correlations, many of which were with an OTU classified as *Actinobacillus,* a member of the healthy respiratory tract microbiome (137, 138). This OTU had an average relative abundance of 0.13% (sd 0.31) and was present in 12 of the 25 samples, all of which originated from patients with HIV infection. Between gene families and *m/z* features, out of 3,960 correlations, 2,309 (58.3%) are positive and 1,646 (41.6%) were negative (**Figure 4.2 lower right**). Of these, 182 were significantly positive and 67 were significantly negative.

Figure 4.2 Correlations between pairs of datasets. Each dot represents a pair of features, sized according to the magnitude of the correlation, and colored by the Spearman correlation between them (red is negative correlation and blue is positive correlation); darker dots indicate significant correlations.

4.2.2.2 Taxonomic Composition Comparison

We calculated the taxonomic compositions of the BAL samples based on the 16S rRNA gene amplicons and metatranscriptome reads; metatranscriptome-based taxonomic assignments could not be completed for three of the samples thus a comparison across data sets was done for 22/25 samples available. For the metatranscriptome data, taxonomic assignments are based on matching *k*-mers in the sequencing reads to sequenced whole genomes and were collapsed to the genus level since 16S rRNA gene data are generally not sensitive enough to allow taxonomic assignments down to the species. Genera that were found in only 1 or 2 samples were removed, leading to a total of 153 genera identified in the 16S rRNA gene data and 553 in the metatranscriptome data. Relative abundance plots show how different the taxonomic profiling appears in the 16S gene data and the metatranscriptome (**Figure 4.3**). In addition to non-bacterial genera, including the eukaryotic genus *Toxoplasma* and viral 'genus' *Enterovirus*, the metatranscriptome measured bacterial genera that were not detected by 16S rRNA gene sequencing. This occurred when either the genera were not in the reference database, (for example, *Tropheryma*), or because the 16S rRNA gene sequences are not appropriately targeted by the primers. Of the 71 genera that could be detected by both platforms, the Pearson (linear) correlation of abundance ranged from -0.42 to 0.93 (average 0.09; sd 0.28). Only 4 of the 71 correlations were significant, after adjusting for multiple hypothesis testing: *Prevotella* (r = 0.67), *Megasphaera* (r = 0.69), *Filifactor* (r = 0.93), and *Nesterenkonia* (r = 0.93). All of these genera except *Nesterenkonia* are common members of the human oral microbiome (139). These correlated genera ranged in average relative abundance from 10.0% (sd 9.7) for the 16S rRNA gene assignments and 6.8% (sd 14.1) for the metatranscriptome read assignments for *Prevotella*,

down to 0.02% (sd 0.07) for the 16S rRNA gene assignments and 0.007% (sd 0.03) for the metatranscriptome read assignments for *Nesterenkonia*.

When collapsed to the taxonomic level of class, the 16S rRNA gene assignments and metatranscriptome read assignments look more similar (**Figure 4.4**). Beyond the inclusion of Picornaviridea viruses and Coccidia protozoa, the metatranscriptome read assignments include the class Negativicutes. These members of Firmicutes phylum include the genera *Veillonella*, which is sometimes placed in the neighboring Clostridia class (62, 63). Other differences may reflect dead or inactive bacteria whose DNA is still being detected in the 16S rRNA gene assignments.

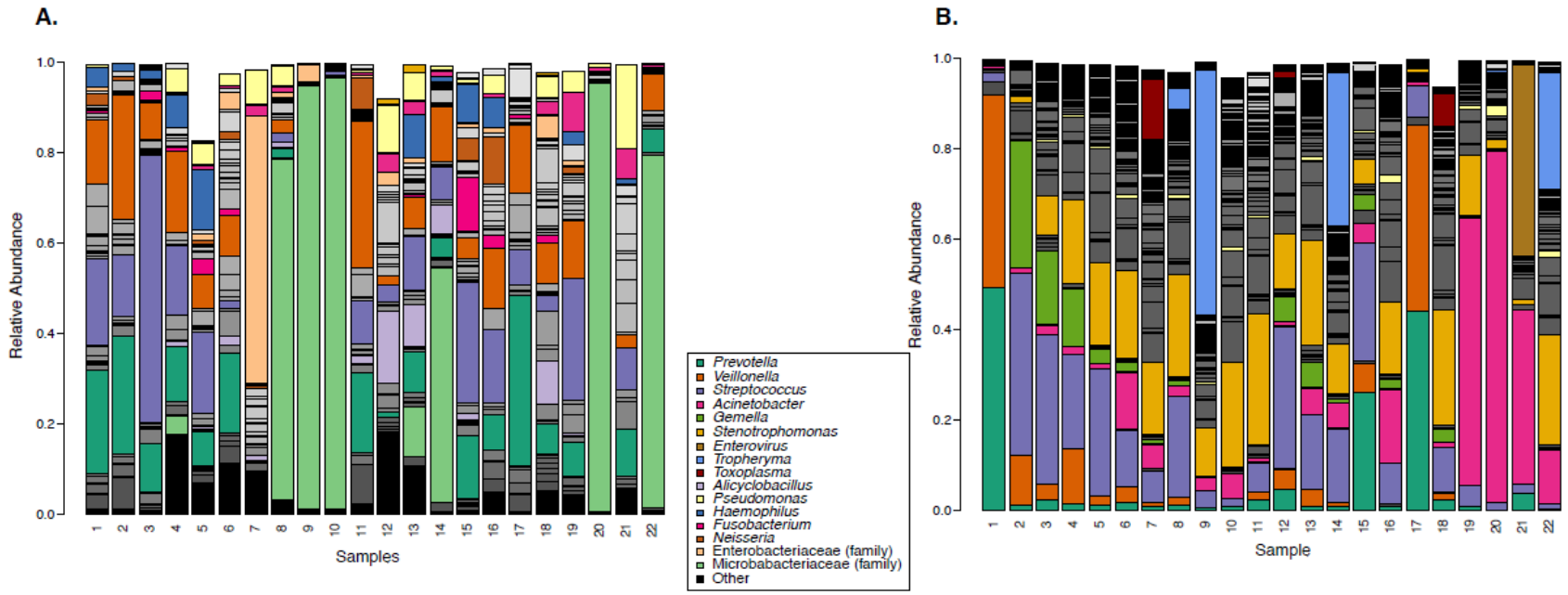

Figure 4.3 Relative abundance plots for assigned taxonomies at the genus level. Highlighted genera are present at greater than 10% abundance in at least one sample from one platform. Taxonomic assignments from (A) 16S rRNA gene sequences and (B) metatranscriptomic sequences.

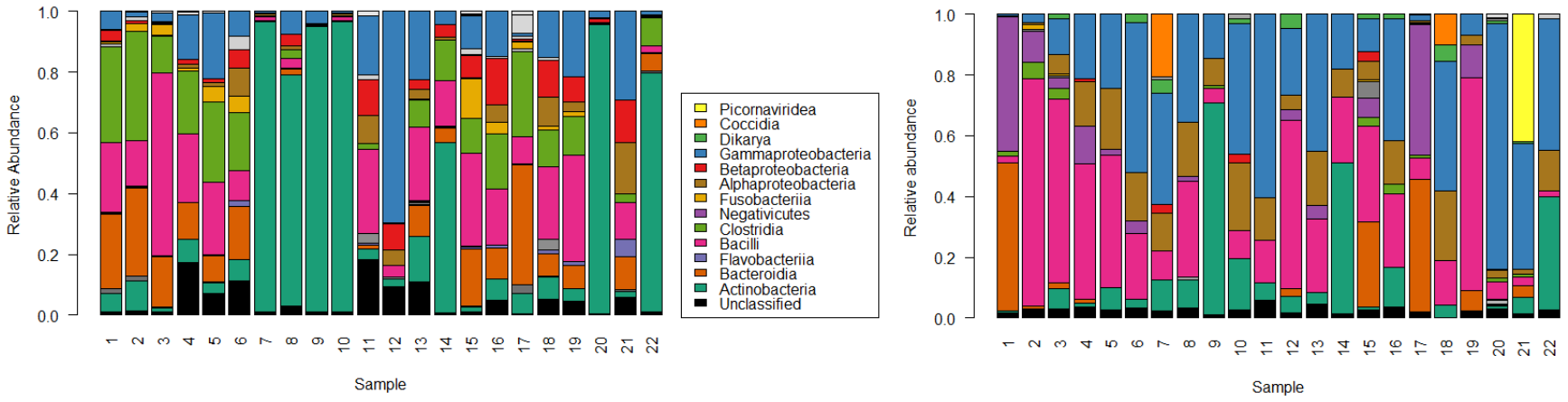

Figure 4.4 Relative abundance plots for assigned taxonomy at the class level. Highlighted classes represent the top 10 classes based on average relative abundance in either (A) 16S rRNA gene sequences or (B) metatranscriptome sequences.

4.2.2.3 KEGG Ontology Comparison

To examine the functional profile of the BAL community, we used PICRUSt (140) to predict the metagenome and metabolic potential from the 16S rRNA gene data, and HUMAnN2 (141) to calculate the expression of transcribed metabolic functions from the metatranscriptomic reads. The predicted metagenome contained 6,909 KO terms based on 749,878 (53%) of the 16S rRNA gene sequences across all samples. The remaining 47% of the sequences were removed for this analysis because they were not within 97% similarity to the reference database. The average nearest sequenced taxon index (NSTI) distance was 0.03 (sd 0.02). This indicates that the predicted metagenome is based on sequences that are, on average, from the same (97% similar) OTUs as our samples. The expressed metabolic functions included 3,490 KO terms based on 6.6 million metatranscriptome reads, across all samples, normalized on metagenomic reads from the same samples. This normalization resulted in the inclusion at low levels of KO terms present but not expressed and reduced the expression levels of highly abundant KO terms.

We then compared the abundance and expression levels of each KO term using Pearson correlations. We were able to match 3,490 KO terms between the PICRUSt predicted metagenome and the metatranscriptome. Correlations between these two datasets ranged from -0.3 to 1.0, with 325 (9.3%) KO terms being nominally significantly correlated (**Figure 4.4**), where nominal significance is defined as a Pearson correlation test with a p-value < 0.05, equivalent to a Pearson correlation coefficient of 0.4. These strong positive correlations indicate functions that are being expressed proportionally to their predicted abundance and thus the metatranscriptome and predicted metagenome data reveal the same trends. While no negative correlations reach significance, 2,157 (62%) of the KO terms were negatively correlated between the predicted metagenome and the metatranscriptome. Even though they are not significant, these

negative correlations indicate functions that are either highly expressed despite low predicted abundance or are slightly expressed despite high predicted abundance.

The KO terms identified were used to look for differential abundance and differential expression between HIV infected individuals (N=6) and HIV uninfected individuals (N=19). We compared the list of KO terms identified as differentially abundant or expressed between the predicted metagenome and the metatranscriptome data, as well as the direction (over abundant/expressed or under abundant/expressed) of all KO terms. Even when KO terms were identified as differentially abundant and differentially expressed for both, the direction was not always the same. The predicted metagenome and metatranscriptome both called four KO terms differentially abundant and differentially expressed (**Table 4.3)**. On these 4 terms, there was agreement on the direction for 3 of them (75%). Overall, there was 54% agreement on direction (**Figure 4.5A**).

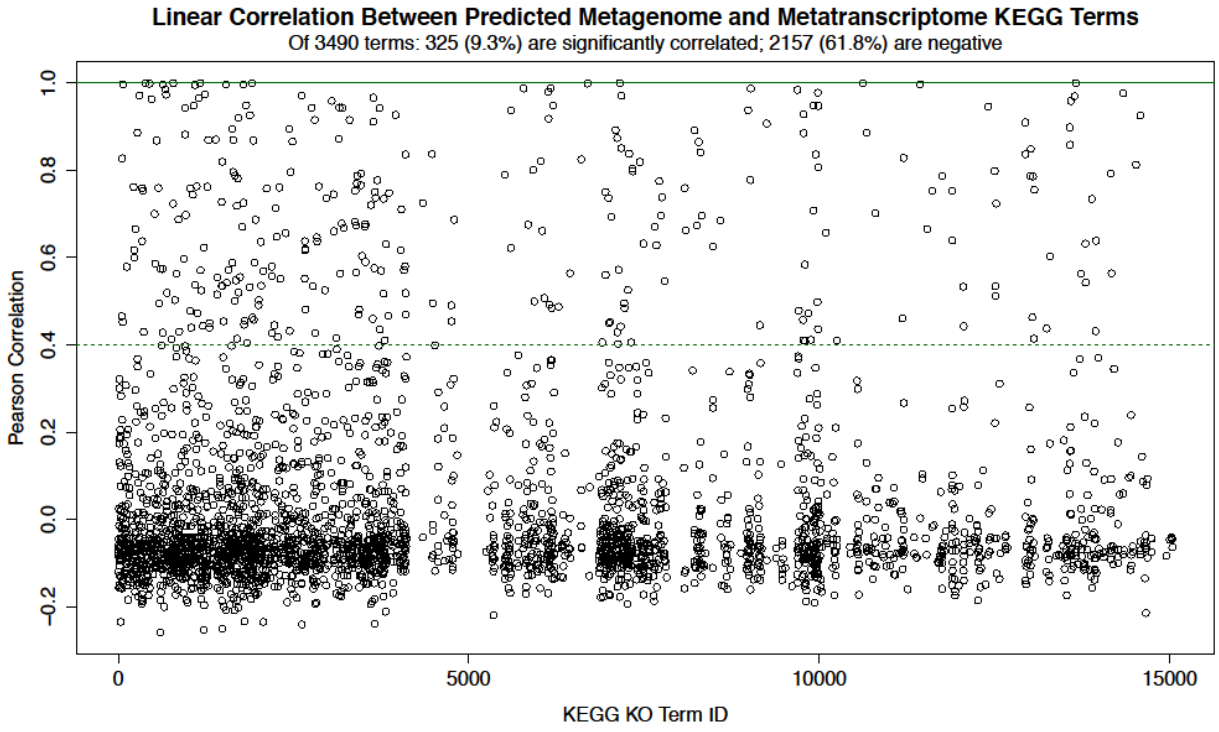

Figure 4.5 Comparison of KO terms between the predicted metabolic functions from the 16S rRNA gene sequences and the metatranscriptomic gene family assignments. Shows the linear correlation coefficient (on Y-axis) between the two platforms for each KO term that was identified in both datasets. The solid line at a Pearson Correlation of 1 represents perfect correlation, achieved only by K01909: long-chain-fatty-acid—[acyl-carrier-protein] ligase. The dashed line at a Pearson Correlation of 0.4 represents nominally significant correlations, of which there are 325 (9.3%).

Table 4.3: KEGG Ontology terms determined to be differentially abundant/expressed in HIV by both the predicted metagenome and the metatranscriptome. The direction of differential abundance is presented in column 3, "Over abundant" indicates that the KEGG term is more abundant in HIV-infected samples than uninfected samples and "Under abundant" indicates that it is more abundant in HIV-uninfected samples than infected samples. The direction of differential expression is presented in column 4, "Over expressed" indicates that the term is transcribed or expressed more in HIV-infected samples than uninfected samples.

KEGG ID	Definition	Predicted Metagenome Differential Abundance in HIV Direction	Metatranscriptome Differential Expression in HIV Direction
K00067	dTDP-4-dehydrorhamnose reductase	Over abundant	Over expressed
K00163	pyruvate dehydrogenase E1 component	Over abundant	Over expressed
K00845	glucokinase	Over abundant	Over expressed
K14205	phosphatidylglycerol lysyltransferase	Under abundant	Over expressed

Similarly, we looked for differential abundance and differential expression between individuals with normal lung function (N=8) and those with COPD (N=17), where COPD was defined as having diffusing capacity of the lungs from carbon monoxide (DLCO) < 80% or forced expiratory volume in 1 second (FEV1) < 70%. As in the HIV comparison, the direction of the differential abundance and differential expression was not the same between the predicted metagenome and the metatranscriptome, even if both platforms identified the KO term as being differentially abundant and differentially expressed. The predicted metagenome and the metatranscriptome both called 11 KO terms differentially abundant and differentially expressed, presented in **Table 4.4**. Of these terms, the two platforms agree on the direction of 7 of them (64%). Over all KO terms, they had 51% agreement on direction (**Figure 4.5B**).

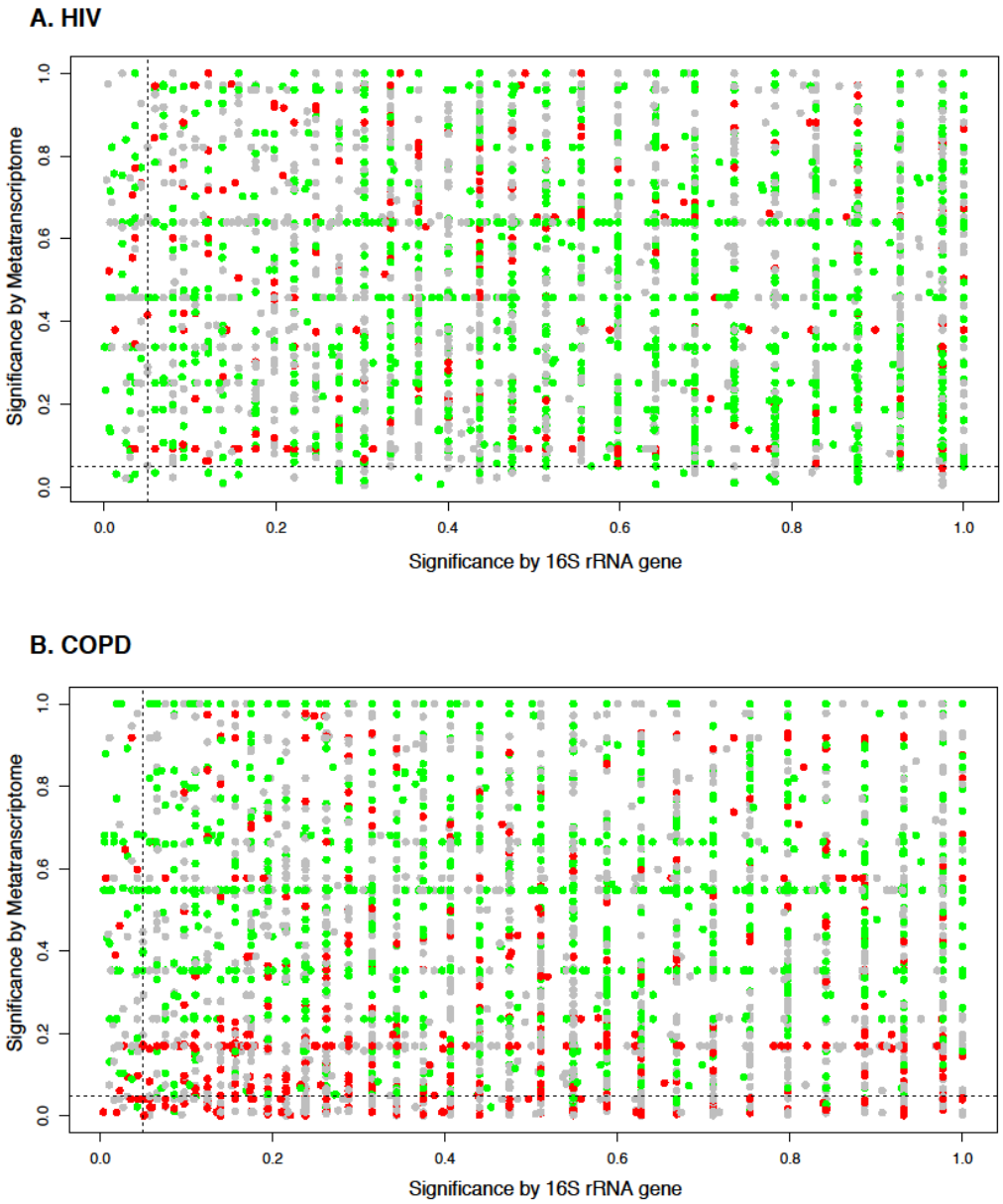

Figure 4.6 Significance of differential abundance/expression of KO terms. Plotted are the p-values resulting from Wilcoxon tests between (A) HIV-infected and uninfected patients and(B) between patients with COPD and with normal lung function. In these plots, green dots represent KO terms that are over-abundant and over-expressed in HIV and COPD, red dots represent KO terms that are under-abundant and under-expressed in HIV and COPD, and grey dots represent KO terms for which the direction of abundance and expression do not match. Dashed lines are included at p=0.05 to indicate nominal statistical significance. No attempt was made to correct for multiple hypotheses testing.

Table 4.4: KEGG Ontology terms determined to be differentially abundant/expressed in COPD by both the predicted metagenome and the metatranscriptome. The direction of differential abundance is presented in column 3, "Over abundant" indicates that the KEGG term is more abundant in COPD positive samples than samples with normal lung function and "Under abundant" indicates that it is more abundant in samples with normal lung function than COPD positive samples. The direction of differential expression is presented in column 4, "Over expressed" indicates that the term is transcribed or expressed more in COPD positive samples than samples with normal lung function and "Under expression" indicates that the term is transcribed or expressed more in samples with normal lung function than those with COPD.

KEGG ID	Definition	Predicted Metagenome Differential Abundance in COPD Direction	Metatranscriptome Differential Expression in COPD Direction
K00174	2-oxoglutarate/2-oxoacid ferredoxin oxidoreductase subunit alpha	Over abundant	Over expressed
K00610	aspartate carbamoyltransferase regulatory subunit	Under abundant	Under expressed
K00851	Gluconokinase	Over abundant	Under expressed
K01571	oxaloacetate decarboxylase, alpha subunit	Over abundant	Under expressed

Table 4.4 Continued

K01968	3-methylcrotonyl-CoA carboxylase alpha subunit	Over abundant	Under expressed
K03146	thiamine thiazole synthase	Under abundant	Under expressed
K03955	NADH dehydrogenase (ubiquinone) 1 alpha/beta subcomplex 1	Under abundant	Under expressed
K05565	multicomponent Na+:H+ antiporter subunit A	Under abundant	Under expressed
K09002	CRISPR-associated protein Csm3	Under abundant	Under expressed
K10793	D-proline reductase (dithiol) PrdA	Under abundant	Under expressed
K11903	type VI secretion system secreted protein Hcp	Over abundant	Under expressed

4.2.3 Three Datasets

4.2.3.1 Block Identification

To generate hypotheses about what OTUs were transcribing which gene families to produce specific metabolites, we integrated the datasets from the 16S rRNA target gene sequences, the metatranscriptome, and the metabolome. We used sparse multi-block partial least squares (sMBPLS) regression (134) to identify blocks of OTUs that are associated with gene families and are producing specific *m/z* features. The regression produced 93 blocks that contained 127 to 575 OTUs, 29,320 to 90,570 gene families, and 1,438 to 3,335 *m/z* features. Block 15 is shown as an example in **Figure 4.6**. This block contains 437 OTUs, 36,778 gene families, and 1,438 *m/z* features. Within each block, we searched the metabolites for pathway enrichment with *mummichog*. The most commonly enriched pathway, i.e. the top pathway in 70 (75.3%) of the blocks, was aspartate and asparagine metabolism (**Table 4.5**). Aspartate and asparagine metabolism was represented by 27 to 42 out of 55 metabolites, depending on the block (p-values between 0.0002 and 0.034).

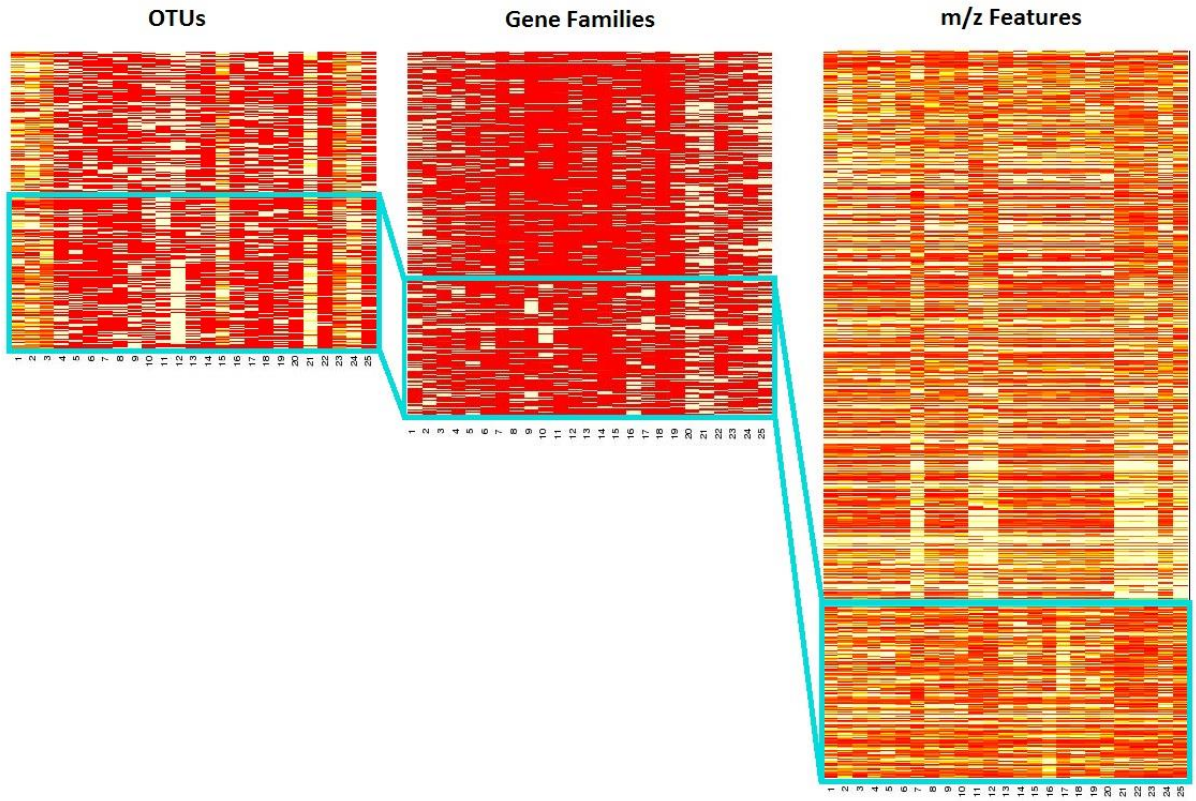

Figure 4.7 Heatmaps of example block identified by sMBPLS, block 15. Each heatmap is broken into two sections separated by a cyan box which encircles the features included in the block. The section outside the box contains the features not included in the block. The components of block 15 include (A) 437 OTUs, (B) 36,778 gene families, and (C) 1,438 m/z features.

Table 4.5: Top pathway identified for each block. Top pathway is defined as the pathway having the lowest adjusted p-value (see column 5) when comparing the m/z features in the cluster to all m/z features in the dataset with mummichog. Pathway names are from mummichog and a list of all possible pathways can be found at http://metafishnet.appspot.com/hbrowse.

BlockID	TopPathway	Overlap	RawP	AdjP
1	Drug metabolism – cytochrome P450	27/30	0.02211	0.00156
2	Linoleate metabolism	42722	0.03782	0.00311
3	Aspartate and asparagine metabolism	34/55	0.00263	0.00097
4	de novo fatty acid biosynthesis	19/30	0.00267	0.00444
5	Aspartate and asparagine metabolism	27/55	0.09785	0.00481
6	Linoleate metabolism	14/18	0.00265	0.00112
7	Lysine metabolism	15/20	0.00422	0.00385
8	Linoleate metabolism	15/18	0.00111	0.0036
9	Linoleate metabolism	15/18	0.00007	0.00486
10	Linoleate metabolism	15/18	0.00047	0.00496
11	Omega-3 faty acid metabolism	13/16	0.00215	0.01344
12	de novo fatty acid biosynthesis	20/30	0.00146	0.01937
13	purine metabolism	21/35	0.04213	0.00738
14	Aspartate and asparagine metabolism	41/55	0	0.00492
15	drug metabolism - other enzymes	42719	0.00061	0.0043
16	Linoleate metabolism	14/18	0.00176	0.00074
17	Linoleate metabolism	42722	0.01222	0.0039
18	Linoleate metabolism	14/18	0.00065	0.00201
19	de novo fatty acid biosynthesis	18/30	0.0412	0.00042
20	purine metabolism	23/35	0.00985	0.00161
21	de novo fatty acid biosynthesis	21/30	0.00083	0.00228
22	Linoleate metabolism	17/18	0	0.00095
23	Di-unsaturated fatty acid beta-oxidation	42592	0.01975	0.00166
24	Aspartate and asparagine metabolism	37/55	0.00061	0.00182
25	Linoleate metabolism	16/18	0.00438	0.01027
26	Aspartate and asparagine metabolism	40/55	0	0.00078

Table 4.5 Continued

27	Aspartate and asparagine metabolism	41/55	0	0.00059
28	Aspartate and asparagine metabolism	40/55	0	0.00189
29	Aspartate and asparagine metabolism	40/55	0	0.0003
30	Aspartate and asparagine metabolism	41/55	0	0.00285
31	Aspartate and asparagine metabolism	41/55	0	0.00102
32	Aspartate and asparagine metabolism	40/55	0	0.00682
33	Aspartate and asparagine metabolism	40/55	0	0.00321
34	Aspartate and asparagine metabolism	41/55	0	0.00347
35	Aspartate and asparagine metabolism	38/55	0.00001	0.00507
36	Aspartate and asparagine metabolism	40/55	0	0.00096
37	Aspartate and asparagine metabolism	40/55	0	0.00197
38	Aspartate and asparagine metabolism	40/55	0	0.03319
39	Aspartate and asparagine metabolism	41/55	0	0.00185
40	Aspartate and asparagine metabolism	40/55	0	0.0068
41	Aspartate and asparagine metabolism	41/55	0	0.00225
42	Aspartate and asparagine metabolism	41/55	0	0.00324
43	Aspartate and asparagine metabolism	41/55	0	0.00232
44	Aspartate and asparagine metabolism	41/55	0	0.03446
45	Aspartate and asparagine metabolism	41/55	0	0.00368
46	Aspartate and asparagine metabolism	41/55	0	0.00117

Table 4.5 Continued

47	Aspartate and asparagine metabolism	40/55	0	0.00069
48	Aspartate and asparagine metabolism	40/55	0	0.00076
49	Aspartate and asparagine metabolism	41/55	0	0.00075
50	Aspartate and asparagine metabolism	41/55	0	0.00231
51	Aspartate and asparagine metabolism	41/55	0	0.00348
52	Aspartate and asparagine metabolism	41/55	0	0.01767
53	Aspartate and asparagine metabolism	41/55	0	0.00034
54	Aspartate and asparagine metabolism	41/55	0	0.00096
55	Aspartate and asparagine metabolism	41/55	0	0.00668
56	Aspartate and asparagine metabolism	41/55	0	0.00244
57	Aspartate and asparagine metabolism	41/55	0	0.00039
58	Aspartate and asparagine metabolism	41/55	0	0.00626
59	Aspartate and asparagine metabolism	41/55	0	0.00035
60	Aspartate and asparagine metabolism	41/55	0	0.00559
61	Aspartate and asparagine metabolism	41/55	0	0.01828
62	Aspartate and asparagine metabolism	41/55	0	0.00068
63	Aspartate and asparagine metabolism	41/55	0	0.0009
64	Aspartate and asparagine metabolism	41/55	0	0.00065
65	Aspartate and asparagine metabolism	41/55	0	0.00077
66	Aspartate and asparagine metabolism	41/55	0	0.00338

Table 4.5 Continued

67	Aspartate and asparagine metabolism	41/55	0	0.01674
68	Aspartate and asparagine metabolism	41/55	0	0.00344
69	Aspartate and asparagine metabolism	41/55	0	0.00121
70	Aspartate and asparagine metabolism	38/55	0.00003	0.008
71	Aspartate and asparagine metabolism	41/55	0	0.00187
72	Aspartate and asparagine metabolism	41/55	0	0.0006
73	Aspartate and asparagine metabolism	39/55	0.00002	0.0021
74	Aspartate and asparagine metabolism	39/55	0.00002	0.00257
75	Aspartate and asparagine metabolism	41/55	0	0.00137
76	Aspartate and asparagine metabolism	39/55	0.00002	0.00015
77	Aspartate and asparagine metabolism	41/55	0	0.00026
78	Aspartate and asparagine metabolism	41/55	0	0.00077
79	Aspartate and asparagine metabolism	41/55	0	0.00329
80	Aspartate and asparagine metabolism	41/55	0	0.00136
81	Aspartate and asparagine metabolism	41/55	0	0.00284
82	Aspartate and asparagine metabolism	41/55	0	0.00071
83	Aspartate and asparagine metabolism	41/55	0	0.0036
84	Aspartate and asparagine metabolism	41/55	0	0.00482
85	Aspartate and asparagine metabolism	41/55	0	0.00258
86	Aspartate and asparagine metabolism	41/55	0	0.00067

Table 4.5 Continued

87	Aspartate and asparagine metabolism	41/55	0	0.00067
88	Aspartate and asparagine metabolism	42/55	0	0.0011
89	Aspartate and asparagine metabolism	41/55	0	0.00193
90	Aspartate and asparagine metabolism	41/55	0	0.00337
91	Aspartate and asparagine metabolism	41/55	0	0.00315
92	Aspartate and asparagine metabolism	41/55	0	0.00025
93	Aspartate and asparagine metabolism	39/55	0.00002	0.00161

To prove that these blocks had biological meaning, we looked at pathways completely covered by the *m/z* features (all metabolites present), verified that all the gene families were being transcribed within the block, and then looked to see if the OTUs belonged to bacteria that are known to use the pathway under investigation. Nitrogen metabolism is a small pathway that contains four metabolites and four gene families; while used by many bacteria, only a limited number of bacterial genera are known to be especially high nitrogen reducers. The pathway was completely covered by the metabolic features in 70 out of the 93 blocks. These blocks were more likely to contain members of all 4 gene families than blocks that did not completely cover the metabolic pathway (Fisher's exact test p-value = 0.01). We defined high nitrogen reducers as members of the following genera: *Rothia, Leptotrichia, Gemella, Treponema, Prevotella, Parvimonas, Staphylococcus, Streptococcus, Veillonella, Actinomyces, Neisseria, Heomophilus,* and *Granulicatella* (142–147). Each of these genera were found in our data and all but

Heomophilus are common members of the human oral microbiome (139). Given this list, the 70 blocks that cover the nitrogen metabolism pathway contained significantly more OTUs that are high nitrogen reducers than blocks that did not completely cover the pathway (Wilcoxon test p-value = 0.021). We confirmed biological meaning of the blocks identified by sMBPLS with the blocks that cover the *m/z* features in the nitrogen metabolism pathway, showing enrichment of nitrogen reducing bacteria and of the gene families in the nitrogen metabolism pathways.

Once the biological relevance of our identified blocks was established, we turned to a less ubiquitous metabolic pathway. A single block, block 4, completely covered the pathway for vitamin B2 (riboflavin) metabolism. This pathway contains only 3 metabolites. For gene families, block 4 is in the 94th percentile (ranked 87 out of 93, containing 51 gene families) for the number of gene families that contain the word "Riboflavin". These gene families include riboflavin biosynthesis proteins RibA, ribAB, RibBA, PYRD, RibC, RibD, RibF, Riboflavin transporter RibU, Riboflavin kinase, and Riboflavin synthase. However, block 4 is one of only 6 blocks that contains no members of either *Corynebacterium* or *Micrococcus* genera, which are known to be high riboflavin producers (148, 149). Instead, the block contained 1 OTU assigned to *Lactobacillus* and 13 OTUs assigned to *Streptococcus*, both of which are capable of synthesizing riboflavin *in vitro* (150) but not at high levels.

4.3 DISCUSSION

We set out to describe the metabolism of the lung microbiome in health, and examine shifts during COPD and HIV-infection. In the individual datasets, we found no overt shifts in OTU

community composition, gene family expression levels, or *m/z* feature abundance. This result was different from, but not contradictory to our previous findings where we saw no shifts in the OTU community composition in HIV-infection but identified a signature subset of 12 *m/z* features that could distinguish between HIV-infected and uninfected subjects (151). In the current work, after correcting for multiple hypotheses testing with the Benjamini-Hochberg procedure, no OTUs, gene families, or *m/z* features were significantly differentially abundant or expressed at a q-value threshold of 0.05. While this is, in part, due to the high number of OTUs, gene families, and *m/z* features, it may also be impacted by the health of our cohort. We defined COPD to include moderate lung function impairment, and significant differences have only been shown in severe COPD (3, 22, 31, 33, 34). The HIV infected patients in our cohort were well-managed and all were on antiretrovirals at the time of sampling. Of the differences that we did see, the gene families are the most readily interpretable. In COPD, the top three under-expressed gene families were UPI0003497762: sodium:glutamate symporter ($p = 3.9x10^{-5}$), UPI00037382D8: ABC transporter permease ($p = 6.5x10^{-6}$), and UPI00047C6BBB: uracil transporter ($p = 3.9x10^{-5}$), all of which had median RPK expression levels of 0. All of these gene families are involved in transporting nutrients into and within cells and the loss of their expression may indicate a shift in the nutrients available to the cell.

By looking at pairs of datasets, we learned more about what each dataset can tell us than about the metabolism of the lung microbiome. The fact that we see negative correlations between pairs of datasets, even when they have been processed to provide the same information, indicated that the information from one dataset cannot make up for another. The negative correlations when comparing taxonomic assignments by 16S rRNA gene sequences and metatranscriptome sequences was not unexpected as the two approaches measure different things – presence of

DNA for the 16S rRNA gene sequences and expression of RNA for the metatranscriptome sequences. The negative correlations lend credence to the theory that genes and proteins that are important to the functioning of the ecosystem often originate in rare members of the community. This theory is also supported by the disagreement in direction of abundance or expression in nearly half of the KO terms (46% when comparing HIV infection and 49% when comparing COPD status).

When the differential abundance or expression was significant, and the direction was the same in the predicted metagenome and the metatranscriptome, special attention was paid to these KO terms because they were so rare. Among these significantly different KO terms, we found terms known to be associated with the conditions in question. Pyruvate dehydrogenase, of which K00163 is a subcomponent (**Table 4.2**), has been shown to be over-expressed in HIV-1 infected cells *in vitro* (152). Similarly, glucokinase (K00845; **Table 4.2**) is required for HIV replication within cells (153). In contrast, we saw NADH dehydrogenase (ubiquinone) 1 alpha/beta subcomplex 1 (K03955; **Table 4.3**) under-abundant and under-expressed in COPD, but it has been shown to be over expressed in lung tumors from COPD patients compared to tumors from those with normal function (154). However, this study was examining human cells rather than the microbial community with which the cells are in contact.

When we integrated all three datasets, we were able to gain new information about the metabolism of the lung microbiome. By integrating metabolomics with the 16S rRNA gene sequence-based OTUs and metatranscriptome sequence-based gene families, we were able to identify blocks that were enriched for metabolic functions. Among these functions was riboflavin metabolism, which was not identified as enriched in the lung microbial community by any of the datasets individually. While humans are capable of processing riboflavin, they are unable to

produce it. Through our block identification, we show the possibility that riboflavin is being produced locally in the lungs by the bacteria present. This has potential implications for lung injuries, as riboflavin administration has been shown in rats to protect against lung injury (155, 156). COPD is the result of architectural damage to the lung (157) so expression of Riboflavin could potentially relieve the injury in this case too.

While this is one of the first studies to integrate data from 16S rRNA gene sequences, metatranscriptomics, and metabolomics for the assessment of the lung microbiome, the current study suffers from a number of limitations. Due to uneven BAL fluid volumes and sequencing failures, our sample size was limited to 25. A larger sample size would lead to greater power to detect differences in COPD status and HIV infection. The mixture of HIV-infected and -uninfected individuals and those with COPD and with normal lung function may also be seen as a limitation to our study. The heterogeneity of lung function and immune status may mask the metabolic functions of the community present in the healthy human lung. However, without this heterogeneity, we would not have been able to look for differentially abundant and expressed KO terms.

The choices in reference databases place unmeasurable limitations on this and other -omics studies. The limitations of reference databases are especially evident in our comparison of taxonomic assignments. For example, the genus *Tropheryma* is not in the GreenGenes database that is the default taxonomic reference for many analysis pipelines and was used for initial 16S rRNA gene sequence taxonomy assignments. This genus was seen in our metatranscriptome sequence taxonomy assignments and was confirmed as present in these samples with qPCR (data not shown). *Tropheryma* is highly relevant to the current study as it has been previously shown to be enriched in the lungs of HIV-infected patients (14). Another reference database limitation

is inherent in metabolomics studies. Only 1% of all possible *m/z* features can be mapped to metabolites in the current databases (158). This lack of identifiability is why we used *mummichog* to look at pathway enrichment among our blocks rather than metabolites. If the mapping of *m/z* features to metabolites were to improve, we could include metabolomics in more direct comparisons to the other -omics platforms.

4.4 METHODS

4.4.1 Patient Population

To compare and integrate three -omics datasets, we identified a subset of samples for which we had 16S rRNA gene sequences, metatranscriptomic, and metabolomics data. The 25 samples used for this analysis originated from the Pittsburgh cohort of the Lung HIV Microbiome Project (LHMP). The larger cohort has been described in (4). Briefly, eligibility requirements included no use of antibiotics in the past three months and no evidence of acute respiratory disease for four weeks. The subset of 25 samples analyzed here originated from participants with the following characteristics: 19 HIV-infected and 6 HIV uninfected, 17 with COPD and 8 with normal lung function. Written informed consent was obtained from all participants after approval of human subjects' protection protocols from review boards of the University of Pittsburgh, the University of California San Francisco, the University of California Los Angeles.

4.4.2 Sample and Sequence Processing

The lung microbiome was sampled by bronchoalveolar lavage (BAL) following an oral wash and gargle with antiseptic mouthwash. BAL fluid was collected from patients and split into multiple aliquots that were stored at –80°C until further processing. One aliquot was used for 16S rRNA gene amplicon sequencing, one for metatranscriptome sequencing, and one for metabolomics profiling, as described below.

For 16S rRNA gene amplicon sequencing, samples had DNA extracted using standard techniques with the PowerSoil® DNA Isolation Kit from MO BIO (Carlsbad, CA). The V4 hypervariable region was amplified and sequenced on the Illumina MiSeq platform using the Caporaso protocol (159). The resulting 1.4 million high quality sequences were processed using QIIME version 1.9 (84). Sequences were clustered at 97% similarity using uclust (160) to form *de novo* operational taxonomic units (OTUs). The OTUs were assigned to taxonomies using the uclust method and the Greengenes database (62, 63). To predict the metagenomics potential of each sample based on the 16S rRNA sequences, we used QIIME to perform closed-reference OTU picking and the PICRUSt software (140) to assign KO term abundances (132, 133).

For metatranscriptome sequencing, samples had RNA extracted using a modified version of Qiagen's RNeasy Micro (Hilden, Germany) protocol. Each aliquot was centrifuged at 4500 x *g* for 5' at 4°C. The supernatant was discarded and the cellular pellet was resuspended in 700μL QIAzol (Qiagen) before transfer to 2mL MP Biomedicals's Lysing Matrix B tubes (Santa Anna, California, USA). Samples were homogenized on a FastPrep-24 (MP Biomedicals) with two rounds at 6.0 m/s for 40s, then centrifuged at 10k x *g* for 2' at 4°C. QIAzol reagent was added to the supernatant to bring the final volume back to 700μL. For phase separation and cleanup, we

followed Qiagen's protocol, including an on-column DNase I treatment. RNA was eluted in 30 µl RNase-free water and checked for sample concentration and integrity before cDNA synthesis using the Nugen Ovation RNA-seq FFPE System (San Carlos, California, USA), according to the manufacturer's protocol. Resulting cDNA samples were purified and size-adjusted using the Zymo Select-a-Size DNA Clean & Concentrator kit (Irvine, California, USA) to remove fragments below 200 base pairs. Sequencing libraries were prepared from 100ng of cDNA with New England Biolab's NEBNext Ultra DNA Library Prep Kit for Illumina (Ipswitch, Massachusetts, USA). Individual libraries were prepared for multiplexing using the NEBNext Multiplex Oligos (Dual Index Primer Set 1, New England Biolabs) and were subjected to 8 cycles of PCR amplification. Libraries were pooled in an equimolar ratio, diluted to 2nM, and 5% PhiX was spiked-in to ensure sequence diversity. The library pool was split evenly across an Illumina HiSeq (San Diego, California, USA) flowcell using TruSeq SBS v3 chemistry for 2x100bp read lengths and run on a HiSeq 2500 in high output mode.

Prior to extracting RNA, we aliquoted 500µl of each sample for DNA metagenome sequencing and stored this at –80°C. DNA was extracted using the PowerSoil DNA Isolation Kit from MO BIO following the manufacturer's protocol with the following exception: after the addition of Solution C1, each tube was incubated at 65°C for 10' and run in a FastPrep-24 at 6.0 m/s for 60s before continuing through the protocol. DNA was eluted in 60µL 10mM Tris. Samples were diluted to 200 pg/µL before library preparation with the Illumina Nextera XT DNA Library Preparation Kit according to the manufacturer's protocol. Individual libraries were prepared for multiplexing using the Illumina Nextera XT index kit. Libraries were purified using 0.5X volumes of Beckman Coulter Life Sciences AMPure XP beads (Indianapolis, Indiana, USA) before inspection and quantification. Libraries were combined in an equimolar ratio into

two pools to avoid barcode overlap, diluted to 4nM, and 5% PhiX was spiked-in to ensure sequence diversity. Each library pool was clustered per lane of an Illumina HiSeq flowcell using a HiSeq Rapid v2 SBS Kit for 2x250bp read lengths and run on a HiSeq 2500 in rapid run mode.

The metatranscriptome and metagenome sequences were filtered to remove human and mitochondrial sequences using custom perl scripts available at https://github.com/ghedin-lab/human-16s-phix-filter. The remaining 6.5 million metatranscriptome sequences total were processed using the HUMAnN2 pipeline (141), normalizing the RNA expression by the DNA abundance. UniRef50 transcript expression tables for each sample were joined and regrouped by KEGG terms. Taxonomic assignments were made using the Livermore Metagenomics Analysis Tool (LMAT), based on k-mers of size 30 (161).

The metabolomics profiling of these samples was described in (151). Briefly, samples were analyzed by liquid chromatography-high-resolution mass spectrometry (LC-FTMS). Mass-to-charge ratios (*m/z* features) were collected from *m/z* 85 to 1275 over 10 minutes. Adaptive processing software package (apLCMS) with xMSanalyzer was used for peak extraction, noise removal, and quantification of ion intensities (162). These data represent *m/z* features, not definitively identified metabolites. Using the *mummichog* (136) pipeline we assigned these *m/z* features to pathways using 100 permutations, including KEGG pathway terms. Where possible, we assigned these *m/z* features to metabolites using MetaboSearch (163), mapping to the Madison Metabolomics Consortium Database (MMCD) (164) and LipidMaps (165) databases with a match threshold of 1 ppm. The metabolite and KEGG identifications with the smallest mass difference were used.

4.4.3 Differential Abundance/Expression

Differential abundance or expression was evaluated for both COPD and HIV in each dataset. For this comparison, COPD was defined as diffusing capacity of the lungs from carbon monoxide (DLCO) < 80% or forced expiratory volume in 1 second (FEV1) < 70% and compared to those with normal lung function. Similarly, individuals from which samples were obtained with HIV infections were compared to those who are HIV uninfected. All comparisons were made using Wilcoxon rank-sum test (166) and corrected for multiple hypotheses testing using the Benjamini-Hochberg correction (94).

We compared the differential abundance or expression of each KEGG term identified in each dataset. The KEGG terms from each dataset were normalized to the scale of 0 to 1. We then used a Wilcoxon rank-sum test to look for differential abundance and expression between HIV infected individuals and HIV uninfected individuals as well as between individuals with and without COPD. We then compared the list of KEGG terms identified as differentially abundant or expressed between datasets, as well as the direction (over abundant/expressed or under abundant/expressed) of all KEGG terms.

4.4.4 Correlations

We looked at Spearman correlations between pairs of datasets for associations between OTUs and gene families, between OTUs and *m/z* features, and between gene families and *m/z* features. Because OTUs were measured in relative abundances, correlations with this dataset were adjusted to partial correlations using the *pcorr* R package (167).

4.4.5 Block Identification

In an attempt to determine associations across datasets, we ran a sparse multi-block partial least squares (sMBPLS) regression (134). This form of regression was developed to study gene regulation and expression based on multiple genomic datasets (including copy number variation, methylation, and microRNA expression levels). The sMBPLS regression method seeks to identify multi-dimensional blocks, blocks that include all types of datasets included in the regression, that are enriched for functional activity. The sMBPLS regression was performed using the R package *msma* (168). We used relative OTU abundance and relative gene family (UniRef50) expression as independent variables (*X*) to explain the *m/z* feature dependent variables (*Y*). The *m/z* features in resulting blocks were run through *mummichog* to identify functional pathway enrichment within each block (136).

5.0 CONCLUSIONS

The work presented aimed to provide a quantitative assessment of the lung microbiome. Instead of listing the microbes detected within the lungs under different disease conditions, we ran sophisticated regression methods, inferred networks among the microbes, and characterized metabolic functions. The results enhance both our knowledge of the lung microbiome and the methodology available to analyze other host-associated microbiomes.

First, we used a LassoGLMM to look for associations between microbes and continuous clinical variables. While we found no surprising associations between oral microbes and clinical blood measurements, nor between lung microbes and inflammation, we were able to demonstrate the effective and flexible capabilities of the LassoGLMM. This regression method can handle repeated measurements from the same individual, whether over a time course or multiple source locations, and the continuous nature of many clinical co-variables.

Then, we looked at cross-domain interactions between bacteria and fungi found in the lungs and on the skin. By expanding the SPIEC-EASI method, we were able to do this in a statistically sound manner. We found that including cross-domain interactions creates more connected and robust networks than either the bacteria or fungal domains alone. The topography of these cross-domain networks can shed light on the community history and stability, including

robustness against perturbations such as antibiotics, that may not be apparent when examining a single domain of life.

Finally, we examined the metabolism of the lung microbiome using three "-omics" technology datasets. We found that taxonomic assignments and predicted metabolic functions from 16S rRNA target gene sequencing are not in agreement with the taxonomic assignments and metabolic functions from metatranscriptome sequences. However, when we integrated these datasets with metabolomics, we were able to uncover enrichment for metabolic functions that would not have been discovered by any one platform alone. Thus, we provided a complete characterization of the metabolism of the lung bacterial microbiome.

Future directions for each of these three sections include applying the methods to new microbiomes, testing hypotheses generated by the methods, and improving based on areas of active research. The LassoGLMM can be applied to new microbiome studies as longitudinal studies become more common and may be improved by incorporating more advanced penalty parameters and the option to include interactions between microbes. Any associations predicted between microbes and host characteristics are likely to be difficult to validate but bioreactors that imitate full ecosystems represent a possible testing environment. Cross-domain SPIEC-EASI networks can be built on any dataset that contains targeted amplicon sequencing of two or more domains, and, as we have shown, the predicted interactions can be validated by co-culturing experiments. Discussions are already underway to expand the SPIEC-EASI network framework to other -omics datasets such as metagenomics, which are cross-domain by nature. The description of the metabolism of the lung microbiome will continue to improve as the -omics technologies, as well as methods to analyze and integrate them, improve. Observations about the

lung microbiome metabolism that are generated by computational methods can then be tested in bioreactors and other laboratory set-ups that mimic the human lung environment.

Overall, we were able to adapt and develop tools to examine host-associated microbiomes in quantitative and inferential ways. By applying these tools to the lung microbiome, we confirmed that the human lung contains an active microbial community, complete with interactions with its host and between its own members. The activity of this microbial community has the potential to impact the human immune system and respiratory health. Additionally, the results in the preceding sections, and from other applications of the tools we adapted and developed, can be used to generate testable hypotheses about the impact of the microbiome on human health.

APPENDIX A

LASSOGLMMFORMICROBIOMES.R

```
### LassoGLMM for Microbiome Studies ###
### Written by Laura Tipton         ###
### Last edited: Jan 11, 2016       ###

## Data should be in the following formats:
# MBdat: 16S/ITS relative abundance data in 1 matrix, samples in rows and OTUs/species in columns with identifiable names
# dat: continuous response varibles in 1 matrix, samples in rows and variables in columns with identifiable names
# demos: categorical explanatory variables in 1 matrix, samples in rows and variables in columns with identifiable names
# ids: identification random effect variables in 1 matrix, samples in rows and variables in columns with identifiable names
# all rows in the above 4 matrices should be in the same order

## Data cleanup, skip if data is already clean
# relabun = function to calculate relative abundance, if not already in this format
relabun <- function(x){
  sums <- apply(x, 1, sum, na.rm=TRUE)
```

```
  y <- x/sums
  return(y)
}

# gt0 = function to count samples that contain OTUs
gt0 <- function(vec){
  v <- as.numeric(vec)
  s <- sum(v>0)
  return(s)
}

# remove OTUs in less than 2 samples by applying gt0 function
MBdat.gt0 <- as.matrix(apply(MBdat, 2, gt0))
MBdat2 <- MBdat[,-which(dat.gt0<2)]

## Variable screening step based on Pearson Correlations
# corrpairs = function to calculate Pearson correlations between all OTU-response variable pairs
corrpairs <- function(ys, xs, fName="Correlations.csv", useQ=FALSE){
  sums <- apply(xs, 2, gt0)
  res <- vector("list", length(ncol(ys)))
  #names(res) <- colnames(ys)
  for(i in 1:ncol(ys)){
    res[[i]] <- vector("list")
    for(j in 1:ncol(xs)){
      c <- cor.test(ys[,i], xs[,j], na.rm=TRUE)
      p <- c$p.value
      q <- c$p.value*(ncol(ys)*ncol(xs))
      if (!is.na(p)){
```

```
      c2 <- c(colnames(ys)[i], colnames(xs)[j], as.numeric(c$estimate), c$conf.in, p, q, sums[[j]])
      write(c2, file=fName, ncolumns=8, append=TRUE, sep=",")
      if (useQ){
        if (q < 0.05){
          res[[i]] <- append(res[[i]], j)
        }
      }
      else {
        if (p < 0.05){
          res[[i]] <- append(res[[i]], j)
        }
      }
    }
  }
  print(paste("Completed y variable ", i, ", ", colnames(ys)[i]))
 }
 return(res)
}

# calculate correlations, in order to move on, assign this to a variable (corrs)
corr <- corrpairs(dat, MBdat2, fName="Correlations.csv")

## Perform LassoGLMM
# penGLMM = function to regress MBdat on dat accounting for demos and ids
penGLMM <- function(ys, xs, corrs, randE, fName='Regression.txt', lam=seq(0,200,1), strat=NULL,
rtrnmod=FALSE){
 require(glmmLasso)
 for (i in 1:ncol(ys)){
```

```
    vars <- unlist(corrs[[i]])
    vars2 <- colnames(xs)[vars]
    vars3 <- ''
    for (j in 1:length(vars2)){
      vars3 <- paste(vars3, vars2[j], sep="+")
    }
    tmp <- data.frame(na.omit(cbind(ys[,i], xs, randE)))
    colnames(tmp) <- c(colnames(ys)[i], colnames(xs), colnames(randE))
    ranEf <- list()
    for(k in 1:ncol(randE)){ ranEf <- append(ranEf, as.formula(paste(colnames(randE)[k], "=~1")));
names(ranEf)[[k]] <- colnames(randE)[k]}
    if (!is.null(strat)){
      tmp <- data.frame(na.omit(cbind(ys[,i], xs, randE, strat)))
      colnames(tmp) <- c(colnames(ys)[i], colnames(xs), colnames(randE), colnames(strat))
      for (k in 1:ncol(strat)){
        vars3 <- paste(vars3, paste0("as.factor(",colnames(strat)[k],")"), sep="+")
      }
    }
    tmp[,c(1:ncol(xs)+1)] <- apply(tmp[,c(1:ncol(xs)+1)],2, as.numeric)
    tmp[,c((ncol(xs)+2):(ncol(xs)+ncol(randE)+1))] <- apply(tmp[,c((ncol(xs)+2):(ncol(xs)+ncol(randE)+1))], 2,
as.factor)
    min <- Inf
    lamb <- 0
    minmod <- list()
    minmod$coefficients <- 0
    minmod$ranef <- 0
    for (l in lam){
      try({
```

```
      mod <- glmmLasso(fix=as.formula(paste("tmp[,1]~", substr(vars3,2,nchar(vars3)))), rnd=ranEf,
data=data.frame(tmp), lambda=l, control=list(q_start=diag(0.1, ncol(randE))))
      if (mod$bic < min){
        minmod <- mod
        min <- mod$bic
        lamb <- l
       }
     }, silent=T)
   }
   write(paste("Y =", colnames(ys)[i]), file=fName, append=T)
   write("Fixed Effects:", file=fName, append=T)
   write.table(as.matrix(minmod$coefficients[abs(minmod$coefficients)>0]), file=fName, append=T)
   write("Random Effects:", file=fName, append=T)
   write.table(as.matrix(minmod$ranef), file=fName, append=T)
   write(paste("Optimal Lamba: ", lamb), file=fName, append=T)
   write(paste("Minimum BIC: ", min), file=fName, append=T)
   write(paste(""), file=fName, append=T)
   print(paste("Completed y variable ", i, ", ", colnames(ys)[i]))
   if (rtrnmod){ return(minmod) }
  }
}

# apply LassoGLMM, this does not need to be assigned to a variable
penGLMM(dat, MBdat2, corr, ids, strat=demos)

## Plotting example
require(car)
par(family="sans")
```

```
# create a temporary dataset sorted by response variable of interest (using 1 in this example and assuming OTU-1 is strongly associated)
tmpplot <- cbind(MBdat2[order(dat[,1]),], dat[order(dat[,1]),1])

# plot a "none" plot to set axes and labels
matplot(log(tmpplot[-which(tmpplot[,1]==0),1]), tmpplot[-which(tmpplot[,1]==0),ncol(tmpplot)], pch=19, type="n", ylab="Response Variable", xlab="log relative abundance", main="OTU-1")

# plot grey dashed lines for all responses
for(i in 1:nrow(tmpplot)){ lines(c(-20,20), c(tmpplot[i,ncol(tmpplot)], tmpplot[i,ncol(tmpplot)]), lty=2, col="grey")}

# finally plot abundances in red
matplot(log(tmpplot[-which(tmpplot[,1]==0),1]), tmpplot[-which(tmpplot[,1]==0),ncol(tmpplot)], pch=19, type="o", add=TRUE, col="red")
```

APPENDIX B

SOP FOR CO-CULTURING MICROBES

1. PURPOSE

1.1. To examine how microbes (both bacteria and fungi) grow together compared to separately.

1.2. To validate the following interactions predicted using the SPIEC-EASI software:

1.2.1. *Emericella nidulans* and *Propionibacterium acnes* (positive)

1.2.2. *Emericella nidulans* and *Rothia dentocariosa* (negative)

1.2.3. *Propionibacterium acnes* and *Rothia dentocariosa* (negative)

1.2.4. *Emericella nidulans, Propionibacterium acnes*, and *Rothia dentocariosa* (negative)

2. REQUIREMENTS

Microbes

Ethanol and bleach spray bottles

2 mL serological pipette tips and aid

Bunsen burner, striker, and gas line	Malt extract agar
Sterile water	Untreated culture plates
10 mL Falcon/culture tubes	Hot plate with stir bar
Tweezers	Disposable inoculum loops
Camera	Hemocytometer and cover slips
Microscope	P100 pipetteman and tips
Brain-Heart Infusion (BHI) media mix	LB media mix
Cryotubes	Parafilm
Trypticase soy agar pre-poured plates with 5% defibrinated sheep's blood (TSA)	
74 mm^2 untreated Nest culture flasks	Dry ice
Black light lamp	Agar

3. NOTES

3.1. Protocol v1.01 is written for those microbes under investigation in April 2016 and will need to be modified for any future test of predicted interactions.

3.2. Microbes under investigation in April 2016 are:

3.2.1 *Emericella nidulans* (aka *Aspergillus nidulans*)

3.2.2 *Propionibacterium acnes*

3.2.3 *Rothia dentocariosa*

4. Rehydrate and Grow Stock – *E nidulans*

When opening the vial, wear goggles and work above a tray to catch glass fragments.

4.0 Clean and disinfect biosafety cabinet (BSC). Flame sterilize and fill 10 mL Falcon tube with 6 mL of sterile water. Set up and light Bunsen burner.

4.1 Heat tip of *E nidulans* vial in Bunsen burner flame. Turn off Bunsen burner!

4.2 Squirt a few drops of water on the hot tip to crack glass.

4.3 Strike hot tip with file or pen to remove tip – make sure fragments go in tray!

4.4 Remove insulation and inner vial with tweezers. Gently raise cotton plug with flame sterilized and cooled tweezers.

4.5 Add .75 mL sterile water (from MilliQ spout) to inner vial, stir to form a suspension.

4.6 Draw up entire contents into pipette and transfer to 10 mL Falcon tube of sterile water.

4.7 Sterilize empty vials and fragments with ethanol prior to disposal in sharps bin.

4.8 Rehydrate at room temperature overnight.

4.9 (Next day, can be performed on the "bacteria" bench) Mix 12.5 g malt extract agar, 7 g glucose, and 250 mL distilled water. Bring to a boil on a hot plate, using a stir bar.

4.10 Autoclave media at 115°C for 10 minutes, let cool until it can be handled without burning hands.

4.11 Pour media into 3-4 culture plates. Wait for plates to solidify.

4.12 Mix rehydrated fungus well with pipette.

4.13 Drop “several” drops totaling about 1 mL of rehydrated fungus onto each of 3 malt extract agar plates.

4.14 Smear drops over plate using fresh, sterile inoculum loop.

4.15 Incubate at 24°C for 72 hours, periodically check for growth by visual inspection. Store remaining rehydrated fungus at 4°C until growth is confirmed.

4.16 After 72 hours of incubation, assuming good growth, photograph plates, inspect cells under microscope.

4.17 Count cells on hemocytometer to determine concentration and practice hemocytometer technique:

4.17.1 Clean hemocytometer and cover slip with ethanol, moisten coverslip with water and affix to hemocytometer.

4.17.2 From 1 plate scrape 1 “colony” with inoculum loop into sterile water, mix well.

4.17.3 Pipette 100 uL of water and cell mixture into loading port on hemocytometer.

4.17.4 Place hemocytometer under microscope, count cells in appropriate squares, photograph each square, and multiply the average cell count by 10^4 to determine cells/mL.

5. Freezing and Storing – *E nidulans*

5.1 Seal 1 plate with parafilm and store upside down in 4°C.

5.2 Mix 5 g LB mix in 250 mL sterile water, autoclave for 15 minutes at 121°C, let cool.

5.3 Mix 9.25 g Brain-Heart Infusion mix in 250 mL sterile water, heat to boil with a stir bar, and autoclave for 15 minutes at 121°C, let cool.

5.4 From 1 plate, scrape each colony (being sure to get the edge of the colony where new growth is happening) into labeled cryotubes containing .5 mL LB and pipette up and down to resuspend.

5.6 From the last plate, scrape each colony (being sure to get the edge of the colony where new growth is happening) into labeled cryotubes containing .5 mL BHI and pipette up and down to resuspend.

5.7 Add .5 mL 15% glycerol to each cyrotube and shake to mix. Freeze at -80°C.

6. Rehydrate and Create Stocks – *P. acnes*

Note that P acnes is an aerotolerant bacteria and all efforts should be made to reduce the oxygen exposure. This means that work should be done near a flame and, if possible, an anaerobic gas mixture or carbon dioxide gas should be blown over the tubes and plates to reduce the oxygen content in the headspaces.

6.0 Clean and disinfect bacterial bench. Set up and light Bunsen burner. Flame sterilize and fill 10 mL Falcon tube with 6 mL of BHI media.

6.1 Set 3 TSA plates around Bunsen burner.

6.2 Flame sterilize and cool tweezers and the top of the vial containing *P acnes.*

6.3 Carefully open vial and remove rubber stopper with tweezers.

6.4 Add .75 mL BHI to vial, stir, without creating air bubbles, to form a suspension.

6.5 Draw up entire contents into pipette and transfer to 10 mL Falcon tube of BHI.

6.6 Drop several drops totaling about 1 mL of rehydrated bacteria onto each of the 3 TSA plates.

6.7 Smear drops over plate using fresh, sterile inoculum loop. Immediately close plate.

6.8 Shut off Bunsen burner. Seal remaining rehydrated bacteria in Falcon tube with parafilm and store at 4°C until growth is confirmed.

6.9 After sufficient drying time, seal plates with parafilm.

6.10 Incubate at 37°C for 48-72 hours, periodically check for growth and contamination by visual inspection.

6.11 After 72 hours of incubation, assuming good growth, photograph plates, inspect cells under microscope. Check for orange glow under black light.

6.12 Count cells on hemocytometer to determine concentration and practice hemocytometer technique (see step 4.17).

7. Freezing and Storing – *P acnes*

7.1 Seal 1 plate with parafilm and store upside down in 4°C.

7.2 Make more LB and/or BHI broth if needed (see sections 5.2 and 5.3).

7.3 From 1 plate, scrape each colony (being sure to get the edge of the colony where new growth is happening) into labeled cryotubes containing .5 mL LB and pipette up and down to resuspend.

7.4 From the last plate, scrape each colony (being sure to get the edge of the colony where new growth is happening) into labeled cryotubes containing .5 mL BHI and pipette up and down to resuspend.

7.5 Add .5 mL 50% glycerol to each cyrotube and shake to mix. Freeze at -80°C.

8. Rehydrate and Create Stocks – *R. dentocariosa*

8.0 Clean and disinfect bacteria bench. Flame sterilize and fill 10 mL Falcon tube of BHI broth (make more if necessary).

8.1 Mix 9.25 g BHI media, 3.75 g agar, and 250 mL distilled water. Bring to a boil on a hot plate using a stir bar.

8.2 Autoclave media at 115°C for 10 minutes, let cool until it can be handled without burning hands.

8.3 Pour media into 3-4 culture plates. Wait for plates to solidify.

8.4 Open *R dentocariosa* vial according to ATCC instructions (see sections 4.1-4 or 6.3).

8.5 Add .75 mL BHI to vial, stir to form a suspension.

8.6 Draw up entire contents into pipette and transfer to 10 mL Falcon tube of BHI.

8.7 Drop "several" drops totaling about 1 mL of rehydrated bacteria onto each of 3 BHI plates.

8.8 Smear drops over plate using fresh, sterile inoculum loop.

8.9 Incubate at 37°C for 24-48 hours, periodically check for growth by visual inspection. Store remaining rehydrated bacteria at 4°C until growth is confirmed.

8.10 After 48 hours of incubation, assuming good growth, photograph plates, inspect cells under microscope. Check for coral glow under black light.

8.11 Count cells on hemocytometer to determine concentration and practice hemocytometer technique (see section 4.17)

9. Freezing and Storing – *R dentocariosa*

9.1 Seal 1 plate with parafilm and store in 4°C.

9.2 Make more BHI broth if needed (see section 5.3).

9.3 From remaining plates, scrape each colony (being sure to get the edge of the colony where new growth is happening) into labeled cryotubes containing .5 mL BHI and pipette up and down to resuspend.

9.4 Add .5 mL 50% glycerol to each cyrotube and shake to mix. Freeze at -80°C.

10. Grow Microbes Under Uniform Conditions

This section is written to be done at 1 time for all 3 microbes, but can be broken down into 2-3 groups, depending on confidence and ability of experimenter(s).

10.1 Mix 55.5 g BHI media, and 1.5 L distilled water. Bring to a boil on a hot plate using a stir bar.

10.2 Autoclave media at 115°C for 10 minutes, let cool until it can be handled without burning hands.

10.3 Clean and disinfect BSC.

10.4 Pipette 10 mL BHI media into each of 12 Nest culture flasks.

10.5 Remove a cryotube of each microbe stored in BHI from the -80 freezer, store on dry ice.

10.6 Using a fresh, sterile inoculum loop for each microbe, scrape the top of the frozen media and place in labeled culture flask, stirring to mix.

10.7 Repeat 12.6 twice for each microbe, resulting in 3 flasks of each microbe.

10.8 For *E nidulans*: using a fresh, sterile inoculum loop, scrape the top of the frozen media and streak fungus onto a malt extract agar plate created in 4.11. This will serve as control that the frozen stock survived and that the fungus can grow at the higher temperature.

10.9 For *P acnes*: using a fresh, sterile inoculum loop, scrape the top of the frozen media and streak onto a TSA plate. This will serve as a control that the frozen stock survived.

10.10 Incubate at 37°C with mild shaking overnight to 24 hours.

10.11 Mix 37 g BHI media, 15 g agar, and 1 L distilled water. Bring to a boil on a hot plate using a stir bar.

10.12 Autoclave media at 115°C for 10 minutes, let cool until it can be handled without burning hands.

10.13 Pour media into 16-20 culture plates. Wait for plates to solidify.

10.14 After incubation, check growth every 3-6 hours (subject to sufficient change in growth in that time). To check growth:

10.14.1 Photograph the flasks.

10.14.2 Pipette up and down several times to ensure a representative sample. Remove 150-200 uL from the flask to an Eppendorf tube.

10.14.3 Take 100 uL from the Eppendorf tube to count on a hemocytometer following step 4.17.

10.14.4 At every other time point (every 6-12 hours), drop remaining ~100 uL onto a BHI plate, and smear using a fresh, sterile inoculum loop. Make sure this plate is labeled with the time.

10.14.5 Return flasks and new plate(s) to incubator.

10.15 After incubation and through the end of all experiments, maintain a serial culture for each microbe using serial splitting just prior to log phase growth (presumably splitting every ~24 hours). To do this, take 1 mL of the current serial culture and add to a flask containing 9 mL new BHI media. Periodically store a serial split at 4°C after the 1 mL has been removed and/or create a frozen stock by mixing .5 mL of culture with .5 mL of 50% glycerol and store at -80°C.

10.16 Check timed plates (created in 12.14.4) after 24 hours (subject to sufficient growth), count colonies as colony forming units.

10.17 Stop taking measurements when no change or negative growth has been observed for 3 consecutive time points.

11. Dual Cultures

This section is written to be done at 1 time for all 3 pairs of microbes but can be broken down into 2-3 groups, depending on the confidence and ability of the experimenter(s).

11.1 Mix 74 g BHI media and 2 L distilled water. Bring to a boil on a hot plate using a stir bar.

11.2 Autoclave media at 115°C for 10 minutes, let cool until it can be handled without burning hands.

11.3 Pipette 10 mL BHI media into each of 18 Nest culture flasks.

11.4 Based on concentration, pipette XXX cells from latest serial culture (see step 12.15) of microbe A into newly labeled flask.

11.5 Based on concentration, pipette XXX cells from latest serial culture of microbe B into same labeled flask.

11.6 Repeat steps 13.5 and 13.6 twice more resulting in 3 flasks for the pair microbe A and microbe B.

11.7 Repeat steps 13.5-7 for every pair of microbes (see sections 1.5.1-6) which should result in 18 flasks.

11.8 Incubate at 37°C with mild shaking overnight to 24 hours.

11.9 Mix 37 g BHI media, 15 g agar, and 1 L distilled water. Bring to boil on a hot plate using a stir bar.

11.10 Autoclave media at 115°C for 10 minutes, let cool until it can be handled without burning hands.

11.11 Pour media into 16-20 culture plates. Wait for plates to solidify.

11.12 After incubation, check growth every 3-6 hours (subject to sufficient change in growth in that time in monocultures). Check growth following steps 12.14.1-5.

Additionally, photograph and count the cells under the black light for dual cultures containing *C minutissium* and *P acnes*.

11.13 Check timed plates after 24 hours (subject to sufficient growth), count colonies as colony forming units.

11.14 Stop taking measurements when no change or negative growth has been observed for 3 consecutive time points.

12. Tri-cultures

12.1 Mix 37 g BHI media and 1 L distilled water. Bring to a boil on a hot plate using a stir bar.

12.2 Autoclave media at 115°C for 10 minutes, let cool until it can be handled without burning hands.

12.3 Clean and disinfect BSC.

12.4 Pipette 10 mL BHI media into each of 18 Nest culture flasks.

12.5 Based on concentration, pipette XXX cells from latest serial culture (see step 10.15) of microbe A into newly labeled flask.

12.6 Based on concentration, pipette XXX cells from latest serial culture of microbe B into same labeled flask.

12.7 Based on concentration, pipette XXX cells from latest serial culture of microbe C into same labeled flask.

12.8 Repeat steps 12.5 and 12.6 twice more resulting in 3 flasks for the triad microbes A, B, and C.

12.9 Incubate at 37°C with milk shaking overnight to 24 hours.

12.10 Mix 37 g BHI media, 15 g agar, and 1 L distilled water. Bring to boil on a hot plate using a stir bar.

12.11 Autoclave media at 115°C for 10 minutes, let cool until it can be handled without burning hands.

12.12 Pour media into 16-20 culture plates. Wait for plates to solidify.

12.13 After incubation, check growth ever 3-6 hours (subject to sufficient change in growth in that time in monocultures). Check growth following steps 10.14.1-5. Additionally, photograph and count the cells under the black light for tri-cultures containing *C minutissium* and *P acnes*.

12.14 Check timed plates after 24 hours (subject to sufficient growth), count colonies as colony forming units.

12.15 Stop taking measurements when no change or negative growth has been observed for 3 consecutive time points.

BIBLIOGRAPHY

1. **Escherich T**. 1886. Die Darmbakterien des Säuglings und ihre Beziehungen zur Physiologie der Verdauung [Enterobacteria of infants and their relation to digestion physiology].

2. **Hilty M**, **Burke C**, **Pedro H**, **Cardenas P**, **Bush A**, **Bossley C**, **Davies J**, **Ervine A**, **Poulter L**, **Pachter L**, **Moffatt MF**, **Cookson WOC**. 2010. Disordered Microbial Communities in Asthmatic Airways. PLoS One **5**:e8578.

3. **Erb-Downward JR**, **Thompson DL**, **Han MK**, **Freeman CM**, **McCloskey L**, **Schmidt LA**, **Young VB**, **Toews GB**, **Curtis JL**, **Sundaram B**, **Martinez FJ**, **Huffnagle GB**. 2011. Analysis of the Lung Microbiome in the "Healthy" Smoker and in COPD. PLoS One **6**:e16384.

4. **Morris A**, **Beck JM**, **Schloss PD**, **Campbell TB**, **Crothers K**, **Curtis JL**, **Flores SC**, **Fontenot AP**, **Ghedin E**, **Huang L**, **Jablonski K**, **Kleerup E**, **Lynch S V**, **Sodergren E**, **Twigg H**, **Young VB**, **Bassis CM**, **Venkataraman A**, **Schmidt TM**, **Weinstock GM**, **Project on behalf of the LHIVM**. 2013. Comparison of the Respiratory Microbiome in Healthy Nonsmokers and Smokers. Am J Respir Crit Care Med **187**:1067–1075.

5. **Gollwitzer ES**, **Saglani S**, **Trompette A**, **Yadava K**, **Sherburn R**, **McCoy KD**, **Nicod LP**, **Lloyd CM**, **Marsland BJ**. 2014. Lung microbiota promotes tolerance to allergens in neonates via PD-L1. Nat Med **20**:642–647.

6. **Iwai S**, **Huang D**, **Fong S**, **Jarlsberg LG**, **Worodria W**, **Yoo S**, **Cattamanchi A**, **Davis JL**, **Kaswabuli S**, **Segal M**, **Huang L**, **Lynch S V.** 2014. The Lung Microbiome of Ugandan HIV-Infected Pneumonia Patients Is Compositionally and Functionally Distinct from That of San Franciscan Patients. PLoS One **9**:e95726.

7. **Huang YJ**, **Nelson CE**, **Brodie EL**, **DeSantis TZ**, **Baek MS**, **Liu J**, **Woyke T**, **Allgaier M**, **Bristow J**, **Wiener-Kronish JP**, **Sutherland ER**, **King TS**, **Icitovic N**, **Martin RJ**, **Calhoun WJ**, **Castro M**, **Denlinger LC**, **DiMango E**, **Kraft M**, **Peters SP**, **Wasserman SI**, **Wechsler ME**, **Boushey HA**, **Lynch S V.** 2011. Airway microbiota and bronchial hyperresponsiveness in patients with suboptimally controlled asthma. J Allergy Clin Immunol **127**:372–381.e3.

8. **Arumugam M**, **Raes J**, **Pelletier E**, **Le Paslier D**, **Yamada T**, **Mende DR**, **Fernandes GR**, **Tap J**, **Bruls T**, **Batto J-M**, **Bertalan M**, **Borruel N**, **Casellas F**, **Fernandez L**,

Gautier L, **Hansen T**, **Hattori M**, **Hayashi T**, **Kleerebezem M**, **Kurokawa K**, **Leclerc M**, **Levenez F**, **Manichanh C**, **Nielsen HB**, **Nielsen T**, **Pons N**, **Poulain J**, **Qin J**, **Sicheritz-Ponten T**, **Tims S**, **Torrents D**, **Ugarte E**, **Zoetendal EG**, **Wang J**, **Guarner F**, **Pedersen O**, **de Vos WM**, **Brunak S**, **Doré J**, **Antolín M**, **Artiguenave F**, **Blottiere HM**, **Almeida M**, **Brechot C**, **Cara C**, **Chervaux C**, **Cultrone A**, **Delorme C**, **Denariaz G**, **Dervyn R**, **Foerstner KU**, **Friss C**, **van de Guchte M**, **Guedon E**, **Haimet F**, **Huber W**, **van Hylckama-Vlieg J**, **Jamet A**, **Juste C**, **Kaci G**, **Knol J**, **Lakhdari O**, **Layec S**, **Le Roux K**, **Maguin E**, **Mérieux A**, **Melo Minardi R**, **M'rini C**, **Muller J**, **Oozeer R**, **Parkhill J**, **Renault P**, **Rescigno M**, **Sanchez N**, **Sunagawa S**, **Torrejon A**, **Turner K**, **Vandemeulebrouck G**, **Varela E**, **Winogradsky Y**, **Zeller G**, **Weissenbach J**, **Ehrlich SD**, **Bork P**. 2011. Enterotypes of the human gut microbiome. Nature **473**:174–180.

9. **Alekseyenko A V**, **Perez-Perez GI**, **De Souza A**, **Strober B**, **Gao Z**, **Bihan M**, **Li K**, **Methé BA**, **Blaser MJ**. 2013. Community differentiation of the cutaneous microbiota in psoriasis. Microbiome **1**:31.

10. **Segal LN**, **Alekseyenko A V**, **Clemente JC**, **Kulkarni R**, **Wu B**, **Chen H**, **Berger KI**, **Goldring RM**, **Rom WN**, **Blaser MJ**, **Weiden MD**. 2013. Enrichment of lung microbiome with supraglottic taxa is associated with increased pulmonary inflammation. Microbiome **1**:19.

11. **Dickson RP**, **Erb-Downward JR**, **Freeman CM**, **Walker N**, **Scales BS**, **Beck JM**, **Martinez FJ**, **Curtis JL**, **Lama VN**, **Huffnagle GB**. 2014. Changes in the Lung Microbiome following Lung Transplantation Include the Emergence of Two Distinct Pseudomonas Species with Distinct Clinical Associations. PLoS One **9**:e97214.

12. **Dickson RP**, **Erb-Downward JR**, **Martinez FJ**, **Huffnagle GB**. 2016. The Microbiome and the Respiratory Tract. Annu Rev Physiol **78**:481–504.

13. **Morris A**. 2014. Heart-lung interaction via infection. Ann Am Thorac Soc **11 Suppl 1**:S52-6.

14. **Lozupone C**, **Cota-Gomez A**, **Palmer BE**, **Linderman DJ**, **Charlson ES**, **Sodergren E**, **Mitreva M**, **Abubucker S**, **Martin J**, **Yao G**, **Campbell TB**, **Flores SC**, **Ackerman G**, **Stombaugh J**, **Ursell L**, **Beck JM**, **Curtis JL**, **Young VB**, **Lynch S V**, **Huang L**, **Weinstock GM**, **Knox KS**, **Twigg H**, **Morris A**, **Ghedin E**, **Bushman FD**, **Collman RG**, **Knight R**, **Fontenot AP**, **Lung HIV Microbiome Project**. 2013. Widespread colonization of the lung by Tropheryma whipplei in HIV infection. Am J Respir Crit Care Med **187**:1110–7.

15. **Hubbell SP**. 2001. The Unified Neutral Theory of Biodiversity and Biogeography. Princeton, NJ.

16. **Sloan WT**, **Lunn M**, **Woodcock S**, **Head IM**, **Nee S**, **Curtis TP**. 2006. Quantifying the roles of immigration and chance in shaping prokaryote community structure. Environ Microbiol **8**:732–740.

17. **Venkataraman A**, **Bassis CM**, **Beck JM**, **Young VB**, **Curtis JL**, **Huffnagle GB**,

Schmidt TM. 2015. Application of a neutral community model to assess structuring of the human lung microbiome. MBio **6**:e02284-14.

18. **Bassis CM**, **Erb-Downward JR**, **Dickson RP**, **Freeman CM**, **Schmidt TM**, **Young VB**, **Beck JM**, **Curtis JL**, **Huffnagle GB**. 2015. Analysis of the upper respiratory tract microbiotas as the source of the lung and gastric microbiotas in healthy individuals. MBio **6**:e00037.

19. **Dickson RP**, **Erb-Downward JR**, **Freeman CM**, **McCloskey L**, **Beck JM**, **Huffnagle GB**, **Curtis JL**. 2015. Spatial Variation in the Healthy Human Lung Microbiome and the Adapted Island Model of Lung Biogeography. Ann Am Thorac Soc.

20. **Delhaes L**, **Monchy S**, **Fréalle E**, **Hubans C**, **Salleron J**, **Leroy S**, **Prevotat A**, **Wallet F**, **Wallaert B**, **Dei-Cas E**, **Sime-Ngando T**, **Chabé M**, **Viscogliosi E**. 2012. The Airway Microbiota in Cystic Fibrosis: A Complex Fungal and Bacterial Community—Implications for Therapeutic Management. PLoS One **7**:e36313.

21. **Carmody LA**, **Zhao J**, **Kalikin LM**, **LeBar W**, **Simon RH**, **Venkataraman A**, **Schmidt TM**, **Abdo Z**, **Schloss PD**, **LiPuma JJ**. 2015. The daily dynamics of cystic fibrosis airway microbiota during clinical stability and at exacerbation. Microbiome **3**:12.

22. **Sze MA**, **Hogg JC**, **Sin DD**. 2014. Bacterial microbiome of lungs in COPD. Int J Chron Obstruct Pulmon Dis **9**:229–38.

23. **Pragman AA**, **Kim HB**, **Reilly CS**, **Wendt C**, **Isaacson RE**. 2012. The lung microbiome in moderate and severe chronic obstructive pulmonary disease. PLoS One **7**:e47305.

24. **Sze MA**, **Dimitriu PA**, **Hayashi S**, **Elliott WM**, **McDonough JE**, **Gosselink J V**, **Cooper J**, **Sin DD**, **Mohn WW**, **Hogg JC**. 2012. The lung tissue microbiome in chronic obstructive pulmonary disease. Am J Respir Crit Care Med **185**:1073–80.

25. **Charlson ES**, **Diamond JM**, **Bittinger K**, **Fitzgerald AS**, **Yadav A**, **Haas AR**, **Bushman FD**, **Collman RG**. 2012. Lung-enriched organisms and aberrant bacterial and fungal respiratory microbiota after lung transplant. Am J Respir Crit Care Med **186**:536–45.

26. **Adami AJ**, **Cervantes JL**. 2015. The microbiome at the pulmonary alveolar niche and its role in Mycobacterium tuberculosis infection. Tuberculosis **95**:651–658.

27. **Beck JM**, **Schloss PD**, **Venkataraman A**, **Twigg H**, **Jablonski KA**, **Bushman FD**, **Campbell TB**, **Charlson ES**, **Collman RG**, **Crothers K**, **Curtis JL**, **Drews KL**, **Flores SC**, **Fontenot AP**, **Foulkes MA**, **Frank I**, **Ghedin E**, **Huang L**, **Lynch S V**, **Morris A**, **Palmer BE**, **Schmidt TM**, **Sodergren E**, **Weinstock GM**, **Young VB**, **Lung HIV Microbiome Project**. 2015. Multicenter Comparison of Lung and Oral Microbiomes of HIV-infected and HIV-uninfected Individuals. Am J Respir Crit Care Med **192**:1335–44.

28. **Cox MJ**, **Allgaier M**, **Taylor B**, **Baek MS**, **Huang YJ**, **Daly RA**, **Karaoz U**, **Andersen GL**, **Brown R**, **Fujimura KE**, **Wu B**, **Tran D**, **Koff J**, **Kleinhenz ME**, **Nielson D**,

Brodie EL, **Lynch S V.** 2010. Airway Microbiota and Pathogen Abundance in Age-Stratified Cystic Fibrosis Patients. PLoS One **5**:e11044.

29. **Zhao J**, **Schloss PD**, **Kalikin LM**, **Carmody LA**, **Foster BK**, **Petrosino JF**, **Cavalcoli JD**, **VanDevanter DR**, **Murray S**, **Li JZ**, **Young VB**, **LiPuma JJ**. 2012. Decade-long bacterial community dynamics in cystic fibrosis airways. Proc Natl Acad Sci U S A **109**:5809–14.

30. **Marri PR**, **Stern DA**, **Wright AL**, **Billheimer D**, **Martinez FD**. 2013. Asthma-associated differences in microbial composition of induced sputum. J Allergy Clin Immunol **131**:346-52–3.

31. **Sze MA**, **Dimitriu PA**, **Suzuki M**, **McDonough JE**, **Campbell JD**, **Brothers JF**, **Erb-Downward JR**, **Huffnagle GB**, **Hayashi S**, **Elliott WM**, **Cooper J**, **Sin DD**, **Lenburg ME**, **Spira A**, **Mohn WW**, **Hogg JC**. 2015. The Host Response to the Lung Microbiome in Chronic Obstructive Pulmonary Disease. Am J Respir Crit Care Med.

32. **Molyneaux PL**, **Cox MJ**, **Willis-Owen SAG**, **Mallia P**, **Russell KE**, **Russell A-M**, **Murphy E**, **Johnston SL**, **Schwartz DA**, **Wells AU**, **Cookson WOC**, **Maher TM**, **Moffatt MF**. 2014. The Role of Bacteria in the Pathogenesis and Progression of Idiopathic Pulmonary Fibrosis. Am J Respir Crit Care Med **190**:906–913.

33. **Garcia-Nunez M**, **Millares L**, **Pomares X**, **Ferrari R**, **Perez-Brocal V**, **Gallego M**, **Espasa M**, **Moya A**, **Monso E**. 2014. Severity-related changes of bronchial microbiome in chronic obstructive pulmonary disease. J Clin Microbiol **52**:4217–4223.

34. **Wu D**, **Hou C**, **Li Y**, **Zhao Z**, **Liu J**, **Lu X**, **Shang X**, **Xin Y**. 2014. Analysis of the bacterial community in chronic obstructive pulmonary disease sputum samples by denaturing gradient gel electrophoresis and real-time PCR. BMC Pulm Med **14**:179.

35. **Borewicz K**, **Pragman AA**, **Kim HB**, **Hertz M**, **Wendt C**, **Isaacson RE**. 2013. Longitudinal analysis of the lung microbiome in lung transplantation. FEMS Microbiol Lett **339**:57–65.

36. **Willner DL**, **Hugenholtz P**, **Yerkovich ST**, **Tan ME**, **Daly JN**, **Lachner N**, **Hopkins PM**, **Chambers DC**. 2013. Reestablishment of Recipient-associated Microbiota in the Lung Allograft Is Linked to Reduced Risk of Bronchiolitis Obliterans Syndrome. Am J Respir Crit Care Med **187**:640–647.

37. **Romani L**. 2011. Immunity to fungal infections. Nat Rev Immunol **11**:275–288.

38. **Iwasaki A**, **Medzhitov R**. 2015. Control of adaptive immunity by the innate immune system. Nat Immunol **16**:343–353.

39. **van Woerden HC**, **Gregory C**, **Brown R**, **Marchesi JR**, **Hoogendoorn B**, **Matthews IP**. 2013. Differences in fungi present in induced sputum samples from asthma patients and non-atopic controls: a community based case control study. BMC Infect Dis **13**:69.

40. **Cui L**, **Lucht L**, **Tipton L**, **Rogers MB**, **Fitch A**, **Kessinger C**, **Camp D**, **Kingsley L**, **Leo N**, **Greenblatt RM**, **Fong S**, **Stone S**, **Dermand JC**, **Kleerup EC**, **Huang L**, **Morris A**, **Ghedin E**. 2015. Topographic Diversity of the Respiratory Tract Mycobiome and Alteration in HIV and Lung Disease. Am J Respir Crit Care Med **191**:932–942.

41. **Cui L**, **Morris A**, **Ghedin E**. 2013. The human mycobiome in health and disease. Genome Med **5**:63.

42. **Deshmukh S**, **Rai M**. 2005. Biodiversity of Fungi: Their Role in Human Life. Science Pub Inc.

43. **Barnes PD**, **Marr KA**. 2006. Aspergillosis: spectrum of disease, diagnosis, and treatment. Infect Dis Clin North Am **20**:545–61, vi.

44. **Odds FC**. 1987. Candida Infections: An Overview. CRC Crit Rev Microbiol **15**:1–5.

45. **Shipley TW**, **Kling HM**, **Morris A**, **Patil S**, **Kristoff J**, **Guyach SE**, **Murphy JE**, **Shao X**, **Sciurba FC**, **Rogers RM**, **Richards T**, **Thompson P**, **Montelaro RC**, **Coxson HO**, **Hogg JC**, **Norris KA**. 2010. Persistent pneumocystis colonization leads to the development of chronic obstructive pulmonary disease in a nonhuman primate model of AIDS. J Infect Dis **202**:302–12.

46. **DeAngelis YM**, **Saunders CW**, **Johnstone KR**, **Reeder NL**, **Coleman CG**, **Kaczvinsky JR**, **Gale C**, **Walter R**, **Mekel M**, **Lacey MP**, **Keough TW**, **Fieno A**, **Grant RA**, **Begley B**, **Sun Y**, **Fuentes G**, **Scott Youngquist R**, **Xu J**, **Dawson TL**. 2007. Isolation and Expression of a Malassezia globosa Lipase Gene, LIP1. J Invest Dermatol **127**:2138–2146.

47. **Fleming A**. 1929. On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of B. influenzæ. Br J Exp Pathol **10**:226–236.

48. **Kim SH**, **Clark ST**, **Surendra A**, **Copeland JK**, **Wang PW**, **Ammar R**, **Collins C**, **Tullis DE**, **Nislow C**, **Hwang DM**, **Guttman DS**, **Cowen LE**. 2015. Global Analysis of the Fungal Microbiome in Cystic Fibrosis Patients Reveals Loss of Function of the Transcriptional Repressor Nrg1 as a Mechanism of Pathogen Adaptation. PLOS Pathog **11**:e1005308.

49. **Bittinger K**, **Charlson ES**, **Loy E**, **Shirley DJ**, **Haas AR**, **Laughlin A**, **Yi Y**, **Wu GD**, **Lewis JD**, **Frank I**, **Cantu E**, **Diamond JM**, **Christie JD**, **Collman RG**, **Bushman FD**. 2014. Improved characterization of medically relevant fungi in the human respiratory tract using next-generation sequencing. Genome Biol **15**:487.

50. **Harrison M**, **Twomey K**, **McCarthy Y**, **O'Connell O**, **Febrer M**, **Alston M**, **Ryan R**, **Plant B**. 2013. LSC 2013 abstract - The role of second-generation sequencing to characterize the fungal microbiota in the adult cystic fibrosis airway, and its correlation with standard culture-based methods and clinical phenotype. Eur Respir J **42**:OP02.

51. **Willger SD**, **Grim SL**, **Dolben EL**, **Shipunova A**, **Hampton TH**, **Morrison HG**, **Filkins**

LM, **O'Toole GA**, **Moulton LA**, **Ashare A**, **Sogin ML**, **Hogan DA**. 2014. Characterization and quantification of the fungal microbiome in serial samples from individuals with cystic fibrosis. Microbiome **2**:40.

52. **Krause R**, **Halwachs B**, **Thallinger GG**, **Klymiuk I**, **Gorkiewicz G**, **Hoenigl M**, **Prattes J**, **Valentin T**, **Heidrich K**, **Buzina W**, **Salzer HJF**, **Rabensteiner J**, **Prüller F**, **Raggam RB**, **Meinitzer A**, **Moissl-Eichinger C**, **Högenauer C**, **Quehenberger F**, **Kashofer K**, **Zollner-Schwetz I**. 2016. Characterisation of Candida within the Mycobiome/Microbiome of the Lower Respiratory Tract of ICU Patients. PLoS One **11**:e0155033.

53. **Willner D**, **Furlan M**, **Haynes M**, **Schmieder R**, **Angly FE**, **Silva J**, **Tammadoni S**, **Nosrat B**, **Conrad D**, **Rohwer F**. 2009. Metagenomic Analysis of Respiratory Tract DNA Viral Communities in Cystic Fibrosis and Non-Cystic Fibrosis Individuals. PLoS One **4**:e7370.

54. **Young JC**, **Chehoud C**, **Bittinger K**, **Bailey A**, **Diamond JM**, **Cantu E**, **Haas AR**, **Abbas A**, **Frye L**, **Christie JD**, **Bushman FD**, **Collman RG**. 2015. Viral Metagenomics Reveal Blooms of Anelloviruses in the Respiratory Tract of Lung Transplant Recipients. Am J Transplant **15**:200–209.

55. **Khemasuwan D**, **Farver CF**, **Mehta AC**. 2014. Parasites of the Air Passages. Chest **145**:883–895.

56. **Bowman SM**, **Free SJ**. 2006. The structure and synthesis of the fungal cell wall. BioEssays **28**:799–808.

57. **Klimek-Ochab M**, **Brzezińska-Rodak M**, **Zymańczyk-Duda E**, **Lejczak B**, **Kafarski P**. 2011. Comparative study of fungal cell disruption--scope and limitations of the methods. Folia Microbiol (Praha) **56**:469–75.

58. **Salter SJ**, **Cox MJ**, **Turek EM**, **Calus ST**, **Cookson WO**, **Moffatt MF**, **Turner P**, **Parkhill J**, **Loman NJ**, **Walker AW**. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol **12**:87.

59. **Weisburg WG**, **Barns SM**, **Pelletier DA**, **Lane DJ**. 1991. 16S ribosomal DNA amplification for phylogenetic study. J Bacteriol **173**:697–703.

60. **Dollive S**, **Peterfreund GL**, **Sherrill-Mix S**, **Bittinger K**, **Sinha R**, **Hoffmann C**, **Nabel CS**, **Hill DA**, **Artis D**, **Bachman MA**, **Custers-Allen R**, **Grunberg S**, **Wu GD**, **Lewis JD**, **Bushman FD**. 2012. A tool kit for quantifying eukaryotic rRNA gene sequences from human microbiome samples. Genome Biol **13**:R60.

61. **Bellemain E**, **Carlsen T**, **Brochmann C**, **Coissac E**, **Taberlet P**, **Kauserud H**. 2010. ITS as an environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases. BMC Microbiol **10**:189.

62. **DeSantis TZ**, **Hugenholtz P**, **Larsen N**, **Rojas M**, **Brodie EL**, **Keller K**, **Huber T**,

Dalevi D, **Hu P**, **Andersen GL**. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol **72**:5069–5072.

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

64. **Kõljalg U**, **Nilsson RH**, **Abarenkov K**, **Tedersoo L**, **Taylor AFS**, **Bahram M**, **Bates ST**, **Bruns TD**, **Bengtsson-Palme J**, **Callaghan TM**, **Douglas B**, **Drenkhan T**, **Eberhardt U**, **Dueñas M**, **Grebenc T**, **Griffith GW**, **Hartmann M**, **Kirk PM**, **Kohout P**, **Larsson E**, **Lindahl BD**, **Lücking R**, **Martín MP**, **Matheny PB**, **Nguyen NH**, **Niskanen T**, **Oja J**, **Peay KG**, **Peintner U**, **Peterson M**, **Põldmaa K**, **Saag L**, **Saar I**, **Schüßler A**, **Scott JA**, **Senés C**, **Smith ME**, **Suija A**, **Taylor DL**, **Telleria MT**, **Weiss M**, **Larsson K-H**. 2013. Towards a unified paradigm for sequence-based identification of fungi. Mol Ecol **22**:5271–5277.

65. **Kõljalg U**, **Larsson K-H**, **Abarenkov K**, **Nilsson RH**, **Alexander IJ**, **Eberhardt U**, **Erland S**, **Høiland K**, **Kjøller R**, **Larsson E**, **Pennanen T**, **Sen R**, **Taylor AFS**, **Tedersoo L**, **Vrålstad T**. 2005. UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. New Phytol **166**:1063–1068.

66. **de Hoog GS**, **Chaturvedi V**, **Denning DW**, **Dyer PS**, **Frisvad JC**, **Geiser D**, **Gräser Y**, **Guarro J**, **Haase G**, **Kwon-Chung K-J**, **Meis JF**, **Meyer W**, **Pitt JI**, **Samson RA**, **Taylor JW**, **Tintelnot K**, **Vitale RG**, **Walsh TJ**, **Lackner M**. 2015. Name Changes in Medically Important Fungi and Their Implications for Clinical Practice. J Clin Microbiol **53**:1056–1062.

67. **Taylor JW**. 2011. One Fungus = One Name: DNA and fungal nomenclature twenty years after PCR. IMA Fungus **2**:113–20.

68. **Hibbett DS**, **Binder M**, **Bischoff JF**, **Blackwell M**, **Cannon PF**, **Eriksson OE**, **Huhndorf S**, **James T**, **Kirk PM**, **Lücking R**, **Thorsten Lumbsch H**, **Lutzoni F**, **Matheny PB**, **McLaughlin DJ**, **Powell MJ**, **Redhead S**, **Schoch CL**, **Spatafora JW**, **Stalpers JA**, **Vilgalys R**, **Aime MC**, **Aptroot A**, **Bauer R**, **Begerow D**, **Benny GL**, **Castlebury LA**, **Crous PW**, **Dai Y-C**, **Gams W**, **Geiser DM**, **Griffith GW**, **Gueidan C**, **Hawksworth DL**, **Hestmark G**, **Hosaka K**, **Humber RA**, **Hyde KD**, **Ironside JE**, **Kõljalg U**, **Kurtzman CP**, **Larsson K-H**, **Lichtwardt R**, **Longcore J**, **Miądlikowska J**, **Miller A**, **Moncalvo J-M**, **Mozley-Standridge S**, **Oberwinkler F**, **Parmasto E**, **Reeb V**, **Rogers JD**, **Roux C**, **Ryvarden L**, **Sampaio JP**, **Schüßler A**, **Sugiyama J**, **Thorn RG**, **Tibell L**, **Untereiner WA**, **Walker C**, **Wang Z**, **Weir A**, **Weiss M**, **White MM**, **Winka K**, **Yao Y-J**, **Zhang N**. 2007. A higher-level phylogenetic classification of the Fungi. Mycol Res **111**:509–547.

69. **Robert V**, **Vu D**, **Amor ABH**, **van de Wiele N**, **Brouwer C**, **Jabas B**, **Szoke S**, **Dridi A**, **Triki M**, **Ben Daoud S**, **Chouchen O**, **Vaas L**, **de Cock A**, **Stalpers JA**, **Stalpers D**, **Verkley GJM**, **Groenewald M**, **Dos Santos FB**, **Stegehuis G**, **Li W**, **Wu L**, **Zhang R**,

Ma J, **Zhou M**, **Gorjón SP**, **Eurwilaichitr L**, **Ingsriswang S**, **Hansen K**, **Schoch C**, **Robbertse B**, **Irinyi L**, **Meyer W**, **Cardinali G**, **Hawksworth DL**, **Taylor JW**, **Crous PW**. 2013. MycoBank gearing up for new horizons. IMA Fungus **4**:371–9.

70. **White JR**, **Nagarajan N**, **Pop M**. 2009. Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. PLoS Comput Biol **5**:e1000352.

71. **Diaz PI**, **Strausbaugh LD**, **Dongari-Bagtzoglou A**. 2014. Fungal-bacterial interactions and their relevance to oral health: linking the clinic and the bench. Front Cell Infect Microbiol **4**:101.

72. **Tarkka M**, **Deveau AL**. 2016. An Emerging Interdisciplinary Field: Fungal–Bacterial Interactions, p. 162–178. *In* Irina S. Druzhinina, Chrisitan P Kubicek (eds.), The Mycota IV, 3rded. Springer.

73. **Sharma R**, **Ranjan R**, **Kapardar RK**, **Grover A**. 2005. SPECIAL SECTION: MICROBIAL DIVERSITY "Unculturable" bacterial diversity: An untapped resource. Curr Sci **89**.

74. **Staley JT**, **Konopka A**. 1985. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol **39**:321–346.

75. **Hawksworth DL**. 1991. The fungal dimension of biodiversity: magnitude, significance, and conservation. Mycol Res **95**:641–655.

76. **Pérez-Losada M**, **Castro-Nallar E**, **Bendall ML**, **Freishtat RJ**, **Crandall KA**. 2015. Dual Transcriptomic Profiling of Host and Microbiota during Health and Disease in Pediatric Asthma. PLoS One **10**:e0131819.

77. **Bolker BM**, **Brooks ME**, **Clark CJ**, **Geange SW**, **Poulsen JR**, **Stevens MHH**, **White JSS**. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol **24**:127–135.

78. **McCafferty J**, **Muhlbauer M**, **Gharaibeh RZ**, **Arthur JC**, **Perez-Chanona E**, **Sha W**, **Jobin C**, **Fodor AA**. 2013. Stochastic changes over time and not founder effects drive cage effects in microbial community assembly in a mouse model. ISME J **7**:2116–2125.

79. **Romero R**, **Hassan SS**, **Gajer P**, **Tarca AL**, **Fadrosh DW**, **Nikita L**, **Galuppi M**, **Lamont RF**, **Chaemsaithong P**, **Miranda J**, **Chaiworapongsa T**, **Ravel J**. 2014. The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. Microbiome **2**:4.

80. **Waldron L**, **Pintilie M**, **Tsao M-S**, **Shepherd FA**, **Huttenhower C**, **Jurisica I**. 2011. Optimized application of penalized regression methods to diverse genomic data. Bioinformatics **27**:3399–406.

81. **Zou H**, **Hastie T**. 2005. Regularization and variable selection via the elastic net. J R Stat Soc Ser B (Statistical Methodol **67**:301–320.

82. **Groll A**, **Tutz G**. 2014. Variable selection for generalized linear mixed models by L 1-penalized estimation. Stat Comput **24**:137–154.

83. **Schelldorfer J**, **Meier L**, **Bühlmann P**. 2014. GLMMLasso: An Algorithm for High-Dimensional Generalized Linear Mixed Models Using ℓ 1 -Penalization. J Comput Graph Stat **23**:460–477.

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

85. **Dannemiller KC**, **Reeves D**, **Bibby K**, **Yamamoto N**, **Peccia J**. 2014. Fungal High-throughput Taxonomic Identification tool for use with Next-Generation Sequencing (FHiTINGS). J Basic Microbiol **54**:315–321.

86. **Bohnen N**, **Degenaar CP**, **Jolles J**. 1992. Influence of age and sex on 19 blood variables in healthy subjects. Z Gerontol **25**:339–345.

87. **Holt PG**. 1987. Immune and inflammatory function in cigarette smokers. Thorax **42**:241–249.

88. **Appay V**, **Sauce D**. 2008. Immune activation and inflammation in HIV-1 infection: causes and consequences. J Pathol **214**:231–241.

89. **Yuan M**, **Lin Y**. 2006. Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B (Statistical Methodol **68**:49–67.

90. **Groll A**. 2014. glmmLasso: Variable selection for generalized linear mixed models by L1-penalized estimation.

91. **Bates D**, **Mächler M**, **Bolker BM**, **Walker SC**. 2015. Fitting linear mixed-effects models using lme4. ArXIV e-print; Press J Stat Softw.

92. **Nakagawa S**, **Schielzeth H**. 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods Ecol Evol **4**:133–142.

93. **Wilcox HE**, **Farrar MD**, **Cunliffe WJ**, **Holland KT**, **Ingham E**. 2007. Resolution of inflammatory acne vulgaris may involve regulation of CD4+ T-cell responses to Propionibacterium acnes. Br J Dermatol **156**:460–465.

94. **Benjamini Y**, **Hochberg Y**. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B.

95. **Tickle T**, **L W**, **Lu Y**, **Huttenhower C**. 2016. Multivariate association of microbial communities with rich metadata in high-dimensional studies. Prog.

96. **Chen EZ**, **Li H**. 2016. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. Bioinformatics btw308.

97. **Johnson PCD**, **Barry SJE**, **Ferguson HM**, **Müller P**. 2015. Power analysis for generalized linear mixed models in ecology and evolution. Methods Ecol Evol **6**:133–142.

98. **Sedgewick AJ**, **Shi I**, **Donovan RM**, **Benos P V.** 2016. Learning mixed graphical models with separate sparsity parameters and stability-based model selection. BMC Bioinformatics **17**:S175.

99. **Mukherjee PK**, **Chandra J**, **Retuerto M**, **Sikaroodi M**, **Brown RE**, **Jurevic R**, **Salata RA**, **Lederman MM**, **Gillevet PM**, **Ghannoum MA**. 2014. Oral Mycobiome Analysis of HIV-Infected Patients: Identification of Pichia as an Antagonist of Opportunistic Fungi. PLoS Pathog **10**:e1003996.

100. **Deng Y**, **Jiang Y-H**, **Yang Y**, **He Z**, **Luo F**, **Zhou J**. 2012. Molecular ecological network analyses. BMC Bioinformatics **13**:113.

101. **Marino S**, **Baxter NT**, **Huffnagle GB**, **Petrosino JF**, **Schloss PD**. 2014. Mathematical modeling of primary succession of murine intestinal microbiota. Proc Natl Acad Sci **111**:439–444.

102. **He X**, **McLean JS**, **Edlund A**, **Yooseph S**, **Hall AP**, **Liu S-Y**, **Dorrestein PC**, **Esquenazi E**, **Hunter RC**, **Cheng G**, **Nelson KE**, **Lux R**, **Shi W**. 2015. Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. Proc Natl Acad Sci **112**:244–249.

103. **Brown GD**, **Denning DW**, **Gow NAR**, **Levitz SM**, **Netea MG**, **White TC**. 2012. Hidden killers: human fungal infections. Sci Transl Med **4**:165rv13.

104. **Kurtz ZD**, **Müller CL**, **Miraldi ER**, **Littman DR**, **Blaser MJ**, **Bonneau RA**. 2015. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLoS Comput Biol **11**:e1004226.

105. **Paine RT**. 1995. A Conversation on Refining the Concept of Keystone Species. Conserv Biol **9**:962–964.

106. **Ze X**, **Le Mougen F**, **Duncan SH**, **Louis P**, **Flint HJ**. 2013. Some are more equal than others. Gut Microbes **4**:236–240.

107. **Morris A**, **Paulson JN**, **Talukder H**, **Tipton L**, **Kling H**, **Cui L**, **Fitch A**, **Pop M**, **Norris KA**, **Ghedin E**. 2016. Longitudinal analysis of the lung microbiota of cynomolgous macaques during long-term SHIV infection. Microbiome **4**:38.

108. **Foster JA**, **Krone SM**, **Forney LJ**, **Foster JA**, **Krone SM**, **Forney LJ**. 2008. Application of Ecological Network Theory to the Human Microbiome. Interdiscip Perspect Infect Dis **2008**:1–6.

109. **Navlakha S**, **Faloutsos C**, **Bar-Joseph Z**. 2015. MassExodus: modeling evolving networks in harsh environments. Data Min Knowl Discov **29**:1211–1232.

110. **Barberan A**, **Bates ST**, **Casamayor EO**, **Fierer N**. 2012. Using network analysis to explore co-occurrence patterns in soil microbial communities. ISME J **6**:343–351.

111. **Steele JA**, **Countway PD**, **Xia L**, **Vigil PD**, **Beman JM**, **Kim DY**, **Chow C-ET**, **Sachdeva R**, **Jones AC**, **Schwalbach MS**, **Rose JM**, **Hewson I**, **Patel A**, **Sun F**, **Caron DA**, **Fuhrman JA**. 2011. Marine bacterial, archaeal and protistan association networks reveal ecological linkages. ISME J **5**:1414–1425.

112. **Seed PC**. 2015. The human mycobiome. Cold Spring Harb Perspect Med **5**:a019810.

113. **Hoffmann C**, **Dollive S**, **Grunberg S**, **Chen J**, **Li H**, **Wu GD**, **Lewis JD**, **Bushman FD**. 2013. Archaea and Fungi of the Human Gut Microbiome: Correlations with Diet and Bacterial Residents. PLoS One **8**:e66019.

114. **Grice EA**, **Kong HH**, **Conlan S**, **Deming CB**, **Davis J**, **Young AC**, **Program NCS**, **Bouffard GG**, **Blakesley RW**, **Murray PR**, **Green ED**, **Turner ML**, **Segre JA**. 2009. Topographical and Temporal Diversity of the Human Skin Microbiome. Sci **324**:1190–1192.

115. **Findley K**, **Oh J**, **Yang J**, **Conlan S**, **Deming C**, **Meyer JA**, **Schoenfeld D**, **Nomicos E**, **Park M**, **Program NCS**, **Kong HH**, **Segre JA**. 2013. Human Skin Fungal Diversity. Nature **498**:367–370.

116. **Newman MEJ**. 2002. Assortative mixing in networks.

117. **Zijnge V**, **van Leeuwen MBM**, **Degener JE**, **Abbas F**, **Thurnheer T**, **Gmür R**, **M. Harmsen HJ**. 2010. Oral Biofilm Architecture on Natural Teeth. PLoS One **5**:e9321.

118. **Bor B**, **Cen L**, **Agnello M**, **Shi W**, **He X**. 2016. Morphological and physiological changes induced by contact-dependent interaction between Candida albicans and Fusobacterium nucleatum. Sci Rep **6**:27956.

119. **Schroeckh V**, **Scherlach K**, **Nutzmann H-W**, **Shelest E**, **Schmidt-Heck W**, **Schuemann J**, **Martin K**, **Hertweck C**, **Brakhage AA**. 2009. Intimate bacterial-fungal interaction triggers biosynthesis of archetypal polyketides in Aspergillus nidulans. Proc Natl Acad Sci **106**:14558–14563.

120. **Cao Y**, **Lin W**, **Li H**. 2016. Large Covariance Estimation for Compositional Data via Composition-Adjusted Thresholding.

121. **Aitchison J**. 1981. A new approach to null correlations of proportions. J Int Assoc Math Geol **13**:175–189.

122. **Liu H**, **Roeder K**, **Wasserman L**. 2010. Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models. Adv Neural Inf Process Syst **24**:1432–

1440.

123. **Csárdi G**, **Nepusz T**. 2006. The igraph software package for complex network research. InterJournal Complex Syst **1695**:1695.

124. **WELCH BL**. 1947. The generalisation of student's problems when several different population variances are involved. Biometrika **34**:28–35.

125. **Huang YJ**, **Kim E**, **Cox MJ**, **Brodie EL**, **Brown R**, **Wiener-Kronish JP**, **Lynch S V**. 2010. A persistent and diverse airway microbiota present during chronic obstructive pulmonary disease exacerbations. OMICS **14**:9–59.

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

127. **Huang YJ**, **Boushey HA**. 2015. The Sputum Microbiome in Chronic Obstructive Pulmonary Disease Exacerbations. Ann Am Thorac Soc **12**:S176–S180.

128. **Millares L**, **Pérez-Brocal V**, **Ferrari R**, **Gallego M**, **Pomares X**, **García-Núñez M**, **Montón C**, **Capilla S**, **Monsó E**, **Moya A**. 2015. Functional Metagenomics of the Bronchial Microbiome in COPD. PLoS One **10**:e0144448.

129. **Diaz PT**, **King ER**, **Wewers MD**, **Gadek JE**, **Neal D**, **Drake J**, **Clanton TL**. 2000. HIv infection increases susceptibility to smoking-induced emphysema*. Chest **117**:285S–285S.

130. **Crothers K**, **McGinnis K**, **Kleerup E**, **Wongtrakool C**, **Hoo GS**, **Kim J**, **Sharafkhaneh A**, **Huang L**, **Luo Z**, **Thompson B**, **Diaz P**, **Kirk GD**, **Rom W**, **Detels R**, **Kingsley L**, **Morris A**. 2013. HIV infection is associated with reduced pulmonary diffusing capacity. J Acquir Immune Defic Syndr **64**:271–8.

131. **Crothers K**, **Huang L**, **Goulet JL**, **Goetz MB**, **Brown ST**, **Rodriguez-Barradas MC**, **Oursler KK**, **Rimland D**, **Gibert CL**, **Butt AA**, **Justice AC**. 2011. HIV infection and risk for incident pulmonary diseases in the combination antiretroviral therapy era. Am J Respir Crit Care Med **183**:388–95.

132. **Kanehisa M**, **Sato Y**, **Kawashima M**, **Furumichi M**, **Tanabe M**. 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res **44**:D457-62.

133. **Kanehisa M**, **Goto S**. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res **28**:27–30.

134. **Li W**, **Zhang S**, **Liu C-C**, **Zhou XJ**. 2012. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. Bioinformatics **28**:2458–66.

135. **Ormerod KL**, **Wood DLA**, **Lachner N**, **Gellatly SL**, **Daly JN**, **Parsons JD**, **Dal'Molin**

CGO, **Palfreyman RW**, **Nielsen LK**, **Cooper MA**, **Morrison M**, **Hansbro PM**, **Hugenholtz P**. 2016. Genomic characterization of the uncultured Bacteroidales family S24-7 inhabiting the guts of homeothermic animals. Microbiome **4**:36.

136. **Li S**, **Park Y**, **Duraisingham S**, **Strobel FH**, **Khan N**, **Soltow QA**, **Jones DP**, **Pulendran B**. 2013. Predicting network activity from high throughput metabolomics. PLoS Comput Biol **9**:e1003123.

137. **Yi H**, **Yong D**, **Lee K**, **Cho Y-J**, **Chun J**. 2014. Profiling bacterial community in upper respiratory tracts. BMC Infect Dis **14**:583.

138. **Krishna P**, **Jain A**, **Bisen PS**. 2016. Microbiome diversity in the sputum of patients with pulmonary tuberculosis. Eur J Clin Microbiol Infect Dis **35**:1205–1210.

139. **Dewhirst FE**, **Chen T**, **Izard J**, **Paster BJ**, **Tanner ACR**, **Yu W-H**, **Lakshmanan A**, **Wade WG**. 2010. The Human Oral Microbiome. J Bacteriol **192**:5002–5017.

140. **Langille MGI**, **Zaneveld J**, **Caporaso JG**, **McDonald D**, **Knights D**, **Reyes JA**, **Clemente JC**, **Burkepile DE**, **Vega Thurber RL**, **Knight R**, **Beiko RG**, **Huttenhower C**. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol **31**:814–821.

141. **Abubucker S**, **Segata N**, **Goll J**, **Schubert AM**, **Izard J**, **Cantarel BL**, **Rodriguez-Mueller B**, **Zucker J**, **Thiagarajan M**, **Henrissat B**, **White O**, **Kelley ST**, **Methé B**, **Schloss PD**, **Gevers D**, **Mitreva M**, **Huttenhower C**. 2012. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol **8**:e1002358.

142. **Hyde ER**, **Andrade F**, **Vaksman Z**, **Parthasarathy K**, **Jiang H**, **Parthasarathy DK**, **Torregrossa AC**, **Tribble G**, **Kaplan HB**, **Petrosino JF**, **Bryan NS**. 2014. Metagenomic Analysis of Nitrate-Reducing Bacteria in the Oral Cavity: Implications for Nitric Oxide Homeostasis. PLoS One **9**:e88645.

143. **Li H**, **Duncan C**, **Townend J**, **Killham K**, **Smith LM**, **Johnston P**, **Dykhuizen R**, **Kelly D**, **Golden M**, **Benjamin N**, **Leifert C**. 1997. Nitrate-reducing bacteria on rat tongues. Appl Environ Microbiol **63**:924–30.

144. **Doel JJ**, **Benjamin N**, **Hector MP**, **Rogers M**, **Allaker RP**. 2005. Evaluation of bacterial nitrate reduction in the human oral cavity. Eur J Oral Sci **113**:14–9.

145. **Smith AJ**, **Benjamin N**, **Weetman DA**, **Mackenzie D**, **MacFarlane TW**. 1999. The Microbial Generation of Nitric Oxide in the Human Oral Cavity. Microb Ecol Heal Dis **11**.

146. **Palmerini CA**, **Palombari R**, **Perito S**, **Arienti G**. 2003. NO synthesis in human saliva. Free Radic Res **37**:29–31.

147. **Hyde ER**, **Luk B**, **Cron S**, **Kusic L**, **McCue T**, **Bauch T**, **Kaplan H**, **Tribble G**,

Petrosino JF, **Bryan NS**. 2014. Characterization of the rat oral microbiome and the effects of dietary nitrate. Free Radic Biol Med **77**:249–57.

148. **Stahmann K-P**, **Revuelta JL**, **Seulberger H**. 2000. Three biotechnical processes using Ashbya gossypii, Candida famata, or Bacillus subtilis compete with chemical riboflavin production. Appl Microbiol Biotechnol **53**:509–516.

149. **Sims GK**, **O'loughlin EJ**. 1992. Riboflavin Production during Growth of Micrococcus luteus on Pyridine. Appl Environ Microbiol **58**:3423–5.

150. **O 'kane DJ**. THE SYNTHESIS OF RIBOFLAVIN BY STAPHYLOCOCCI.

151. **Cribbs SK**, **Uppal K**, **Li S**, **Jones DP**, **Huang L**, **Tipton L**, **Fitch A**, **Greenblatt RM**, **Kingsley L**, **Guidot DM**, **Ghedin E**, **Morris A**. 2016. Correlation of the lung microbiota with metabolic profiles in bronchoalveolar lavage fluid in HIV infection. Microbiome **4**:3.

152. **Ringrose JH**, **Jeeninga RE**, **Berkhout B**, **Speijer D**. 2008. Proteomic studies reveal coordinated changes in T-cell expression patterns upon infection with human immunodeficiency virus type 1. J Virol **82**:4320–30.

153. **Brass AL**, **Dykxhoorn DM**, **Benita Y**, **Yan N**, **Engelman A**, **Xavier RJ**, **Lieberman J**, **Elledge SJ**. 2008. Identification of host proteins required for HIV infection through a functional genomic screen. Science **319**:921–6.

154. **Boelens MC**. 2008. Molecular genetic studies in epithelial cells of lung cancer and COPD patients.

155. **Al-Harbi NO**, **Imam F**, **Nadeem A**, **Al-Harbi MM**, **Korashy HM**, **Sayed-Ahmed MM**, **Hafez MM**, **Al-Shabanah OA**, **Nagi MN**, **Bahashwan S**. 2015. Riboflavin attenuates lipopolysaccharide-induced lung injury in rats. Toxicol Mech Methods **25**:417–23.

156. **Seekamp A**, **Hultquist DE**, **Till GO**. 1999. Protection by vitamin B2 against oxidant-mediated acute lung injury. Inflammation **23**:449–60.

157. **Anderson GP**, **Anderson**, **P. G**. 2016. Advances in understanding COPD. F1000Research **5**:2392.

158. **Wang M**, **Carver JJ**, **Phelan V V**, **Sanchez LM**, **Garg N**, **Peng Y**, **Nguyen DD**, **Watrous J**, **Kapono CA**, **Luzzatto-Knaan T**, **Porto C**, **Bouslimani A**, **Melnik A V**, **Meehan MJ**, **Liu W-T**, **Crüsemann M**, **Boudreau PD**, **Esquenazi E**, **Sandoval-Calderón M**, **Kersten RD**, **Pace LA**, **Quinn RA**, **Duncan KR**, **Hsu C-C**, **Floros DJ**, **Gavilan RG**, **Kleigrewe K**, **Northen T**, **Dutton RJ**, **Parrot D**, **Carlson EE**, **Aigle B**, **Michelsen CF**, **Jelsbak L**, **Sohlenkamp C**, **Pevzner P**, **Edlund A**, **McLean J**, **Piel J**, **Murphy BT**, **Gerwick L**, **Liaw C-C**, **Yang Y-L**, **Humpf H-U**, **Maansson M**, **Keyzers RA**, **Sims AC**, **Johnson AR**, **Sidebottom AM**, **Sedio BE**, **Klitgaard A**, **Larson CB**, **Boya P CA**, **Torres-Mendoza D**, **Gonzalez DJ**, **Silva DB**, **Marques LM**, **Demarque DP**, **Pociute E**, **O'Neill EC**, **Briand E**, **Helfrich EJN**, **Granatosky EA**, **Glukhov E**, **Ryffel F**, **Houson H**, **Mohimani H**, **Kharbush JJ**, **Zeng Y**, **Vorholt JA**, **Kurita KL**,

Charusanti P, **McPhail KL**, **Nielsen KF**, **Vuong L**, **Elfeki M**, **Traxler MF**, **Engene N**, **Koyama N**, **Vining OB**, **Baric R**, **Silva RR**, **Mascuch SJ**, **Tomasi S**, **Jenkins S**, **Macherla V**, **Hoffman T**, **Agarwal V**, **Williams PG**, **Dai J**, **Neupane R**, **Gurr J**, **Rodríguez AMC**, **Lamsa A**, **Zhang C**, **Dorrestein K**, **Duggan BM**, **Almaliti J**, **Allard P-M**, **Phapale P**, **Nothias L-F**, **Alexandrov T**, **Litaudon M**, **Wolfender J-L**, **Kyle JE**, **Metz TO**, **Peryea T**, **Nguyen D-T**, **VanLeer D**, **Shinn P**, **Jadhav A**, **Müller R**, **Waters KM**, **Shi W**, **Liu X**, **Zhang L**, **Knight R**, **Jensen PR**, **Palsson BØ**, **Pogliano K**, **Linington RG**, **Gutiérrez M**, **Lopes NP**, **Gerwick WH**, **Moore BS**, **Dorrestein PC**, **Bandeira N**. 2016. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol **34**:828–37.

159. **Caporaso JG**, **Lauber CL**, **Walters WA**, **Berg-Lyons D**, **Huntley J**, **Fierer N**, **Owens SM**, **Betley J**, **Fraser L**, **Bauer M**, **Gormley N**, **Gilbert JA**, **Smith G**, **Knight R**. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J **6**:1621–1624.

160. **Edgar RC**. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics **26**:2460–1.

161. **Ames SK**, **Hysom DA**, **Gardner SN**, **Lloyd GS**, **Gokhale MB**, **Allen JE**. 2013. Scalable metagenomic taxonomy classification using a reference genome database. Bioinformatics **29**:2253–60.

162. **Yu T**, **Park Y**, **Johnson JM**, **Jones DP**. 2009. apLCMS--adaptive processing of high-resolution LC/MS data. Bioinformatics **25**:1930–6.

163. **Zhou B**, **Wang J**, **Ressom HW**. 2012. MetaboSearch: Tool for Mass-Based Metabolite Identification Using Multiple Databases. PLoS One **7**:e40096.

164. **Cui Q**, **Lewis IA**, **Hegeman AD**, **Anderson ME**, **Li J**, **Schulte CF**, **Westler WM**, **Eghbalnia HR**, **Sussman MR**, **Markley JL**. 2008. Metabolite identification via the Madison Metabolomics Consortium Database. Nat Biotechnol **26**:162–164.

165. **Sud M**, **Fahy E**, **Cotter D**, **Brown A**, **Dennis EA**, **Glass CK**, **Merrill AH**, **Murphy RC**, **Raetz CRH**, **Russell DW**, **Subramaniam S**. 2007. LMSD: LIPID MAPS structure database. Nucleic Acids Res **35**:D527–D532.

166. **Mann HB**, **Whitney DR**. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other 50–60.

167. **Kim S**. 2015. ppcor: Partial and Semi-Partial (Part) Correlation.

168. **Kawaguchi A**. 2016. msma: Multiblock Sparse Multivariable Analysis.

169. **Tipton L**, **Ghedin E**, **Morris A**. 2016. The lung mycobiome in the next-generation sequencing era. Virulence 1–8.