Understanding the Clinical Microbiome
Biological Engineering Thesis Proposal

Claire Duvallet

October 11, 2016

Contents

Abstract

In spite of the recent increase in research on the human microbiome, there is not a clear consensus on the relationship between human microbial communities and disease. Microbes colonize our entire bodies, supplementing our body's functions, priming and training our immune systems, providing resistance to colonization by pathogens, and contributing to maintenance of health or progression of disease. However, major knowledge gaps exist in this field. The microbiota of certain body sites have been much less studied than others, and even the most extensively studied body sites lack a synthesized understanding of the clinical relevance of human-microbe associations. Additionally, translating microbial associations into biological hypotheses remains challenging due to the lack of centralized tools and databases for assigning biological meaning to groups of microbes.

This thesis will expand our understanding of the clinical microbiome in three ways. First, I will characterize the relationships between the microbial communities of the aerodigestive tract and their associations with clinical factors, increasing our basic understanding of this under-studied system. Next, I will perform a meta-analysis of published case-control gut microbiome studies across many disease states, synthesizing many existing studies by identifying consistent microbial markers of health and disease. Finally, I will curate groups of biologically related microbes to enable enrichment analyses and generalizable interpretations of results from microbiome studies.

1 Overall objectives and specific aims

1.1 Overall objectives

The research presented in this thesis is united by a common purpose: advancing our understanding of the clinical human microbiome. First, I will enrich our basic understanding of an under-studied microbial system by characterizing the relationships between the microbial communities of the aerodigestive tract and their associations with clinical factors. Second, I will synthesize results across many studies of a well-studied system, moving the field toward a better understanding of and consensus on the associations between gut microbial communities and human disease. To do this, I will perform a meta-analysis of published case-control gut microbiome studies across many disease states to identify consistent microbial markers of health and disease. Finally, I will curate existing knowledge on microbial communities and develop a tool for inferring generalizable biological hypotheses from existing and future microbiome studies. Together, this work will contribute new knowledge to the exciting field of human microbiome research and will empower researchers to draw clinically meaningful insights from their exisiting and future analyses.

1.2 Specific Aims

Aim 1 Apply standard methods to identify microbial community characteristics associated with gastro-esophogeal reflux disease and aspiration.

1. Quantify relationships between lung, gastric, and throat microbial communities.
2. Identify clinical modulators of lung, gastric, and throat microbial communities.

Aim 2 Perform a meta-analysis of gut microbiome studies to identify consistent microbial signatures within and across multiple diseases.

1. Compile and process publicly available case-control gut microbiome studies with a standardized method.
2. Determine whether certain microbes are consistently associated with disease in general or with specific diseases.
3. Identify relationships between physiologically-related diseases by comparing their microbial characteristics.

Aim 3 Enable generalizable interpretations of microbiome analyses by assigning bacteria to groups with similar characteristics and known associations with disease.

1. Combine existing databases with targeted literature searches to define *microbe sets* based on known biological relationships.
2. Use machine-learning techniques to extract disease-associated *microbe sets* from datasets collected in Aim 2.
3. Develop these *microbe sets* into a collaborative tool for use in interpreting new microbiome studies.

2 Background and significance

Three to five pages

The topics addressed in this thesis are broad, and are all connected by the motivation to better understand clinically-relevant associations between microbes and their human hosts. My work will focus on the microbial communities of two major body systems: the aerodigestive and gastrointestinal tracts. To study these, I will use a combination of traditional analytical techniques, supplemented by novel methods as required. This section will provide background on the aerodigestive tract, the gut microbiome, and current analytical techniques used in microbiome studies.

2.1 Aerodigestive tract

2.1.1 Physiology and disease

From an engineering perspective, the aerodigestive tract, consisting of the upper gastrointestinal and respiratory tracts, can be thought of as different compartments connected by the esophagus and trachea (Figure 1). The mass transport between the throat, stomach, and lung compartments is regulated by complex physiological mechanisms. Swallowing guides material from the mouth to the stomach, but may dysfunction and allow material to enter the lungs. The esophageal sphincter usually prevents material from leaving the stomach, but sometimes allows gastric contents into the lungs [2]. Finally, complex homeostatic mechanisms clear the lungs of foreign bodies and create a selective environment for microbes in the lungs [3, 4]. Understanding these complex physiological relationships is further complicated by the experimental and ethical considerations involved in sampling and studying the human aerodigestive tract [2].

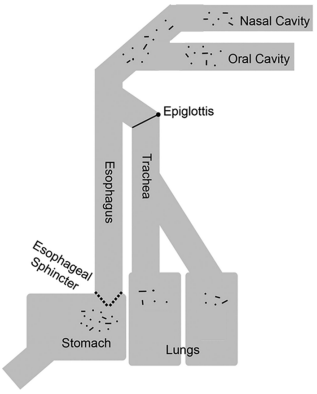

Figure 1: Schematic of the flow relationship between sites of the aerodigestive tract. Adapted from [1].

Gastro-esophogeal reflux disease (GERD) is a set of syndromes in which the reflux of stomach contents leads to troublesome symptoms or complications [5, 6]. The most common symptoms of GERD are regurgitation and heartburn, but the disease may also present asymptomatically [5, 6]. GERD can be diagnosed by demonstrating reflux of gastric contents, injury to the esophogus, or based on symptoms alone [5]. GERD affects 10-20% of people in Western Europe and North America, and can lead to severe complications such as Barrett's esophagus [5]. Proton-pump inhibitors (PPIs) are often prescribed for GERD, though long-term adverse effects of these drugs is becoming more widely understood [7, 8, 9]. In cases of severe reflux, fundoplication surgery may be recommended, in which part of the stomach is wrapped around the esophagus to prevent refluxate from leaving the stomach and going up into the esophagus [7].

Aspiration is another complex aerodigestive condition in which foreign material is inhaled, either through macro-aspiration resulting from dysfunctional swallowing or micro-aspiration which is common in healthy people [7, 4]. Clinically, aspiration is defined as the inhalation of foreign material such as food or gastric contents into the lungs [10]. Most aspiration

events are unwitnessed without obvious outward signs or symptoms, and may involved large quantities of aspirated material or may be micro-aspiration events [10]. Aspiration resulting from swallowing dysfunction can be diagnosed with a Modified Barium Swallow test (MBS) [3]. In the MBS procedure, patients are observed videoradiographically as they swallow varying quantities and viscosities of food or liquid impregnated with barium, a contrast agent. The swallowing process is observed and abnormalities like aspiration of contents past the vocal chords can be diagnosed [3, 11]. However, the MBS test cannot be used to determine aspiration of gastric contents or episodes of micro-aspiration, which may also have clinical relevance [10, 12]. No validated clinical biomarkers exist to diagnose and define gastric and micro-aspiration events[12], but they are thought to play a role in causing or exacerbating many respiratory diseases [9, 13, 14]. Currently, microaspiration of gastric contents is studied by measuring the concentration of bile or pepsin in the lungs, but such assays are rarely validated against gold-standard methods and have not undergone clinical validation [9, 12]. Further complicating the issue is that many healthy patients have a baseline level of micro-aspiration [4, 7].

GERD and aspiration are thought to be associated with many respiratory diseases, but the precise link and mechanisms underlying these associations remains very unclear [9]. The prevalence of GERD in respiratory diseases has been estimated to be up to 90% for some diseases, and often presents without the common symptoms of heartburn or regurgitation [9]. Studies have shown that GERD is related to adverse outcomes after lung transplantation [7] and reduced lung function in patients with cystic fibrosis [14]. Aspiration of stomach contents is thought to contribute to these adverse outcomes, either through the aspiration of bile triggering a change in the lung's environment and making it more favorable to colonization, or through direct aspiration of gastric bacteria leading to infection [13, 14]. However, because of the difficulties in diagnosing and studying microaspiration, aspiration of gastric contents, and GERD, precise causal links between reflux, aspiration, and lung disease have yet to be established [9, 14].

2.1.2 Microbiome of the aerodigestive tract

The microbial communities of the human aerodigestive tract are among the least studied human-associated microbiota. Neither gastric nor lung sites were included in the Human Microbiome Project, leading to a dearth of studies and data on these important body sites [1]. Previous studies have examined the relationships between microbial communities in the upper aerodigestive tract, with conflicting results [1, 14, 15, 16]. Some have shown striking similarities in the microbial communities across the respiratory tract [1, 14] while others argue that different sites have distinct microbial communities [15], and that these vary even within individual sites like the lungs [17, 18]. The lungs have classically thought to be sterile and free of bacteria in healthy people [1, 2, 16]. However, both culture-based and culture-independent studies of the lung microbiome have recovered bacteria, mostly of the *Prevotella*, *Veillonella*, and *Streptococcus* genera [1]. Although the balance between factors that shape the lung microbiome remains to be fully elucidated, lung microbial communities are shaped by the balance of immigration, elimination, and active colonization of microbes throughout the respiratory tract [1, 4]. Existing studies have examined the microbiome of patients with cystic fibrosis, chronic obstructive pulmonary disorder, and asthma, as well as

smokers and patients on PPI therapy [14, 15, 17]. However, many of these studies are limited by small sample sizes and sampling constraints. The stomach has also traditionally thought to be relatively sterile due to its low pH [19]. Many studies of the gastric microbiota have focused on *Helicobacter pylori*, which colonizes many healthy patients and is also implicated in the development of gastric cancer and other stomach diseases [20]. These consistently find that *H. pylori* dominates the mucosal community when it is present [19, 20]. Other culture-independent studies have found diverse gastric communities in both the mucosa and lumen healthy patients' stomachs [1, 15, 19]. While the majority of the gastric flora is likely seeded by the oral microbiota (via swallowing), the stomach also contains its own unique community [1, 19].

2.2 Gut microbiome

2.2.1 Gut microbiome in health and disease

The human gastrointestinal tract is integral to health and disease and its microbiota is, in turn, integral to its functioning. The intestine digests food and absorbs nutrients, and also plays important roles in maintaining metabolic homeostasis, regulating hormone levels, and communicating sensory signals with the brain. The microbes in our guts provide essential functions to our well being. They help us harvest energy from the food we eat and digest otherwise undigestible fibers, train our immune system, break down xenobiotics and other foreign products, and release metabolites and hormones that provide chemical signals to our body's regulatory mechanisms [21, 22, 23]. These signals can act locally in the gut and can also have larger systemic effects, for example by sending signals through the vagus nerve to the brain via the 'gut-brain axis' [22, 24].

Because of this complex interplay between host and microbes, many diseases have been hypothesized to be associated with the gut microbiome. These include metabolic disorders, inflammatory and auto-immune diseases, pathogenic diarrhea, and others [21, 23, 25, 26]. The relationship between the gut microbiome and obesity has been extensively studied in mouse models and human patients. These studies have found encouraging results and causal associations in mice but relatively little consensus in humans [23, 27, 28, 29]. Related disorders like metabolic syndrome and diabetes have also been examined, with similarly little consensus on specific microbial markers of these conditions [30, 31]. Inflammatory bowel disease (IBD) is a chronic disease characterized by mucosal inflammation of the GI tract. Animal studies relating gut microbes, immune function, and IBD have pointed to an important role for bacteria in affecting the progression of IBD [32]. Recent 16S studies have focused on classifying IBD patients based on their fecal microbiota and on identifying discriminatory taxa in stool [33, 34]. IBD patients have distinct microbial profiles that distinguish them from controls, but the specific microbes which drive this distinction have not been identified, and it is likely that larger community structure patterns play an important role in shaping the IBD microbiota. [29, 33, 34]. Colorectal cancer (CRC) has long been associated with the intestinal microbiome, with bacterial metabolites thought to contribute to the development of CRC and specific bacteria having been isolated from CRC tumors [35, 36]. 16S studies have analyzed disease associations with tumor- and lumen-associated microbiota, and have consistently found enrichment of the *Fusobacterium* genus in CRC patients [35, 37, 38, 39].

Diarrheal diseases caused by intestinal pathogens have also been extensively surveyed using 16S methods, especially in the context of *Clostridium difficile* infection (CDI) and related fecal microbiota transplants [40, 41]. Finally, diseases like rheumatoid arthritis, autism, Parkinson's, HIV, and others have also been examined for microbial associations, though these fields remain relatively unexplored [22, 26, 42, 43].

2.2.2 Existing understanding of the gut microbiome

Although specific microbiome-disease associations remain unclear, general characteristics of the gut microbiome are relatively well-understood. People have unique gut microbial communities, very few microbes can be consistently found across the majority of people, and many gut communities are dominated by one or two phyla (Bacteroidetes and Firmicutes) [44]. Our gut microbiome is stable over time and can also change rapidly in response to disease, antibiotics, travel, and diet [45]. These perturbations can be fully reversible and can also have long-term effects [45]. Dysbiosis is often discussed as an "imbalance" of gut microbes, though is generally applied to mean any community disruption related to disease [46]. Generally, less diverse communities are thought to be associated with disease, though many recent studies find no significant association with diversity and disease [29]. Early mouse studies associated the ratio of Bacteroidetes to Firmicutes with altered phenotypes, but few subsequent studies have found similar associations in human patients [23, 47].

Combining existing studies to increase our ability to find consistent disease associations is a promising approach, but the most recent of these meta-analyses have had mixed results [29, 47]. In some cases like IBD, strong and consistent signals can be found across studies but no specific microbes have been found to be consistently associated with IBD [29]. Meta-analyses of obesity studies also tend to find no clear taxonomic associations with obesity [47, 29], even though the microbiome has been causally linked to obesity in mouse models [23, 27]. Other meta-analysis studies are not relevant to extracting clinically-relevant microbial associations: many of these test the ability of various statistical and machine learning methods to extract biomarkers or classify disease states, without much interpretion of what the statistical results mean clinically [48, 49, 50, 51].

2.3 Analytical background and significance

2.3.1 Data generation, analysis and associated challenges

A common way that researchers study the human microbiome is to do amplicon-based next generation sequencing of the complex microbial communities. This culture-independent method begins with extracting DNA from a sample of interest, amplifying the universally conserved bacterial 16S rRNA gene, and sequencing by one of the available technologies such as 454 Pyrosequencing or, more recently, Illumina HiSeq or MiSeq [48, 47]. The resulting reads are quality-controlled and often processed into Operational Taxonomic Units (OTUs), clusters of similar sequences which serve as proxies for bacterial species [48]. To interpret results, researchers assign taxonomies to OTUs using a variety of methods, for example by mapping them to annotated reference genomes or using Bayesian inference trained on a reference set of annotated bacteria [52, 53]. Each of these processing steps affects the eventual output data, and there are no accepted standardized methods followed by all studies.

The data that results from these surveys can be very challenging to analyze. The datasets are often very high-dimensional, with hundreds of OTUs present in a given cohort which may only have tens of samples [47]. The data is also incredibly sparse: only very few OTUs tend to be present in many of the samples, and most entries in the data matrix are zeros [50]. Furthermore, strong batch effects between studies result from differences in experimental and computational processing steps. For example, different taxonomy databases contain different microbes or conflicting names for the same bacteria, making it difficult to compare even published, annotated results across studies. These issues may be major contributors to the lack of consensus on the role of the microbiome in disease, in spite of the broad availability of data and studies.

While there exist no established standards for processing or analyzing 16S data, most studies take similar approaches to gleaning insight from case-control cohorts [48]. Alpha diversity, the diversity of species within each sample, is usually compared across groups of interest. Beta diversity, the diversity between samples, is also frequently compared to understand whether samples within groups are more similar to each other than they are to the other group(s) [48]. Finally, most studies perform univariate non-parametric tests on the abundance of OTUs to find bacteria significantly associated with the condition of interest [54]. However, because of the high-dimensionality of the data and the often very low sample sizes, many studies yield no significant results [54].

2.3.2 Interpreting taxonomy-based microbiome analyses

Few analytical tools exist to interpret the lists of significant OTUs which often result from 16S analyses into biological hypotheses. Similar groups of bacteria are frequently associated with health or disease states, but identifying the patterns which group these bateria remains a manual task for researchers. Typically, once significant OTUs are found for a certain condition, researchers perform a literature search and hope to find previous mechanistic studies on these bacteria. Other more seasoned researchers can often look at a list and infer over-representation of certain phenotypes, such as spore-formers or short-chain fatty acid-producing bacteria. However, few systematic approaches to extract meaning from significant OTU associations is currently used in the field.

Enrichment analysis is a powerful way to directly identify biologically meaningful patterns in high-dimensional data. Enrichment analysis is widely used in RNA expression studies and has been proposed for use in metabolomics studies [55, 56]. Gene Set Enrichment Analysis (GSEA) introduced this statistical method to biomedical applications. In GSEA, genes are ranked by their differential expression between two conditions. Then, *a priori*-defined groups of related genes are analyzed for over- or under-representation at either end of the ranked list. Rather than asking whether individual genes are correlated with a phenotype, GSEA allows for the identification of groups of genes which change together. This allows for identification of significant phenotypes where individual genes do not exhibit large enough changes to reach significance on their own. It also enables more direct biological interpretation, since the gene sets are defined *a priori* based on biological knowledge. Enrichment analyses like GSEA could be incredibly useful in microbiome studies, where many phenotype associations are likely to result from groups of bacteria working together, and high-dimensional datasets frequently produce few significant OTU-level phenotype associations.

Enrichment analyses rely upon the existence of curated sets of the features of interest (i.e. genes, metabolites, or microbes). No such grouping of microbes, i.e. *microbe sets*, currently exist. In GSEA, genes were grouped into gene sets based on their common pathways, functions, locations in the chromosome, and associations with disease [55]. Similar annotations exist in some microbial databases, but none of these databases or tools have been used to define groups of related microbes. ImG contains approximately 10,000 annotated bacterial genomes, but the annotations are not fully complete and do not span all categories of possible interest [57]. SourceTracker can be used to label microbial communities according to their environmental source, but requires input training sets with each use in order to learn and make the classifications [58]. Finally, bioinformatic tools have been developed that can infer functional content (PICRUST) [59] or metabolic pathways (HUMANN) [60] from 16S data. Again, these tools are dataset-specific and have not been generalized to define biologically related groups of organisms in a study-independent way.

3 Research design and methods

Six to eight pages

The research presented in this thesis is united by a common purpose: advancing our understanding of the clinical human microbiome. First, I will enrich our basic understanding of an under-studied microbial system. Second, I will collect and synthesize results from many studies of a well-studied system, to move the field toward a better understanding of the associations between microbial communities and human disease. Finally, I will curate existing knowledge on microbial communities to develop a tool for inferring generalizable biological hypotheses from existing and future microbiome studies. Together, these aims will contribute new knowledge to an exciting field of research and will empower researchers to draw more meaningful insights on the most clinically-relevant factors of the human microbiome.

3.1 Aim 1: Aerodigestive microbiota associated with GERD and aspiration

GERD, aspiration, and respiratory infections are three related conditions with complex and unclear interactions. We know that aspirating patients are at a higher risk for respiratory infections, and that many patients who present with idiopathic respiratory problems have a high prevalence of GERD. Furthermore, the microbial communities of the aerodigestive tract are connected and likely exchange bacterial members, which may contribute to respiratory infections. We hypothesize that the microbial communities of the aerodigestive tract are extensively exchanging microbes, and that certain clinical conditions like aspiration or GERD may modulate the amount of exchange happening across various sites.

To address this hypothesis, we will first identify which microbes are exchanged across sites and define a metric to quantify the "extent" of this exchange. To define this metric, we will incorporate both the co-occurence and the abundance of microbes in the two sites and calculate it for each site-combination. Next, we will identify clinical factors that have an effect on microbial exchange in the aerodigestive tract. Specifically, we will investigate how aspiration, reflux, and PPI use affects the similarity of communities and the exchange of

microbes between sites in the aerodigest tract. We hypothesize that aspiration will increase the lung-throat connection and that reflux and PPI use will strengthen the stomach-lung connection. Quantitatively describing the amount of microbial exchange happening in the aerodigestive tract and determining clinical modulators of this exchange will contribute new knowledge to our current understanding of the aerodigestive microbiome, and could inform future aerodigestive investigations and treatments.

3.1.1 Aerodigestive patient cohort

The cohort presented in this work represents the largest collection of human aerodigestive tract samples of its kind. It consists of 261 patients recruited by Rachel Rosen (M.D., GI/Nutrition) and her team at Boston Children's Hospital over the course of the past 6 years. Multiple samples were taken from patients: throat swabs, gastric fluid, and broncho-alveolar lavages (BAL) (Table 1). To acquire a BAL sample, a bronchoscope is inserted into the lungs of an anaesthetized patient, saline is flushed through the bronchoscope, and then suctioned back up [16]. Gastric fluid is suctioned during an endoscopy, and throat swabs are acquired by brushing the posterior tongue [15]. Many patients in this cohort were monitored for GERD with 24-hour impedance monitoring [5], which identifies the total number of reflux episodes, the percent of time each patient was refluxing, and the acidic or non-acidic nature of the reflux event. A subset were also tested for aspiration with an MBS test, and most patients have metadata for being on or off PPI medication.

Sites	**N**
gastric, throat, & BAL	87
gastric & throat	45
gastric & BAL	34
BAL & throat	9

Table 1: Number of patients with data for each combination of sites.

3.1.2 Quantify exchange of microbes between lung, gastric, and throat communities

To understand the microbial exchange between sites in the aerodigestive tract, we must first identify which microbes are being exchanged and then quantify the extent of this exchange across the sites. We define a microbe as exchanged between two sites if the Spearman correlation of its abundance in both sites is greater than 0.5. In other words, if a microbe is being exchanged between sites, we expect that if we see more of it in one site, then we will also see more of it in the other. To calculate this correlation, we consider only patients who have the microbe present in both sites (blue dots in Figure 2). We quantify the extent of exchange, p_s by asking how many of the total patients have the microbe present in both sites. In other words, p_s is the percentage of patients who are exchanging that microbe across their two sites.

One factor to consider when drawing conclusions from the p_s metric is that because of the low bacterial biomass in the gastric and lung sites, it is possible that some microbes which are "exchanged" across these sites are simply both being seeded by the environment. However, if these microbes are phylogenetically related or if they are known members of the gastric or lung communities, this would indicate that the OTUs are being selected for by the environment and are relevant community members. Small phylogenetic distance

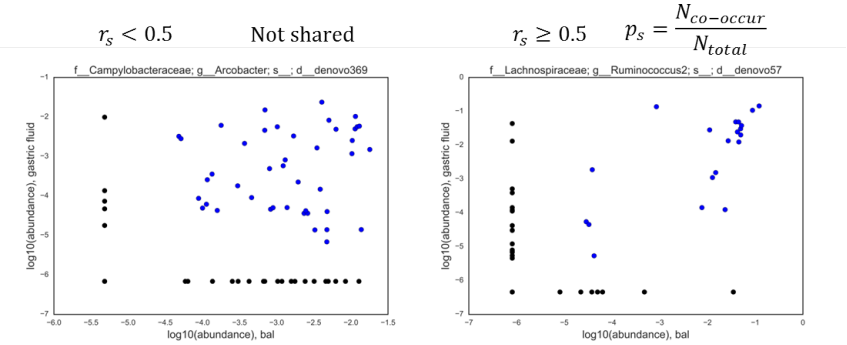

Figure 2: If the abundance of a microbe when it is present in both sites is correlated, then we consider it exchanged across those sites (blue points, right panel). p_s is then calculated as the percentage of patients who have the microbe present in both sites (blue points divided by total points, where each point represents one patient). Example of microbes which are (A) not exchanged and (B) exchanged between stomach and lungs.

between exchanged microbes would indicate either non-random seeding of the aerodigestive communities or non-random selection of randomly-seeded microbes. In either case, the identified microbes would be a relevant part of the microbial communities. Also, we can compare the exchanged OTUs we identify with previous work to determine whether these microbes could be considered commensals, or if they have functions that would allow them to survive in their specific niches.

3.1.3 Identify clinical modulators of lung, gastric, and throat microbial communities

Once we quantify microbial exchange within the aerodigestive tract, we can begin to ask which clinical factors modulate the exchange that occurs between sites. We hypothesize that aspiration will modulate the amount of throat-lung and gastric-lung exchange, that reflux will modulate the amount of gastric-lung exchange, and that PPI use may also affect the gastric-lung connection, or may simply change the whole communities. For each of these clinical factors, we will calculate a new p_s for each previously-defined exchanged microbe, within each patient subgroup. In other words, p'_s is the percent of the patients in one subgroup that have each exchanged microbe present in both of the sites. We will also investigate the overall community similarity (i.e. beta diversity) between the two sites for each patient subgroup. Finally, we may pursue more exploratory analyses like univariate comparisons of OTU abundances in patients with and without each condition.

Our first hypothesis is that aspirators will have more exchange between their throat and lungs and between their stomachs and their lungs. Aspirating patients are at higher risk for respiratory infections, which may be a result of stomach or oral microbes successfully seeding the lungs because the airway is no longer adequately protected. Reflux surgery is often prescribed to aspirators with chronic infections, as it is thought that the refluxate of these patients carries microbes which seed the lungs. The dataset has 48 patients with abnormal MBS test results (aspirators) and 63 patients with normal results.

Our next hypothesis is that patients with more severe GERD will have more exchange

between the stomach and lung communities. Because we are interested in GERD that may modulate the stomach-lung connection by actually reaching the top of the esophogaus, we will focus our analyses on proximal, i.e. full-column, reflux. Our GERD data quantifies the percent of reflux which is full-column, and does not provide a hard cutoff above which reflux is considered "severe". Thus, another goal of this work is to identify if such a threshold exists. We will examine how exchange between and similarity of the lung ang gastric communities changes as a function of this "severity" cutoff. If we find a cutoff in which the amount of exchange in patients who are above the cutoff is significantly higher than in patients who are below the cutoff, then perhaps this threshold could define the clinically-relevant amount of reflux that is problematic. The dataset has 125 patients with reflux testing.

Finally, we will investigate how PPIs modulate the exchange between all three aerodigestive sites. PPIs reduce the acidity of the gastric fluid, which may increase the bacterial load of the refluxate and lead to more frequent colonization of the lungs [61]. We hypothesize that patients on PPIs will have altered gastric fluid microbial communities, and that the exchange between the lungs and stomach may be increased. The dataset has 114 patients on PPIs and 85 patients not on PPIs.

An important consideration when interpreting these results is that this work does not address the direction of microbial exchange between aerodigestive sites, nor does it directly link increased microbial exchange with adverse outcomes like respiratory infections. We assume that most of the community is being seeded from the throat, but do not explicitly know the balance between immigration, elimination, and growth of microbes in each site [1]. Follow up studies focusing on patients who develop respiratory infections or who frequently have GERD- or aspiration-associated respiratory infections should be undertaken to directly link the exchange between communities with adverse clinical outcomes.

3.2 Aim 2: Meta-analysis of gut microbiome studies

By combining results from existing gut microbiome case-control studies, we can move the field toward a consolidated understanding of consistent microbial markers of gut-related diseases. We hypothesize that certain bacteria will often be associated with disease, and that some of these bacteria will be associated with many different types of diseases while others will be unique to one or two conditions. Additionally, we hypothesize that microbial signatures of health and disease will be more similar in similar diseases (i.e. in diabetes and obesity vs. in diabetes and autism).

To address our hypotheses, we will first acquire a comprehensive collection of case-control gut microbiome datasets and process them with standardized methods. We will analyze each dataset individually and synthesize the results from all datasets with basic meta-analysis techniques to identify microbes consistently associated with health or general disease. We will also perform a similar intra-disease meta-analysis for studies analyzing the same disease to identify microbes that may be specific markers of certain diseases and not others. Finally, we will identify relationships between physiologically-related diseases by comparing their microbial characteristics across multiple datasets. This comprehensive pan-disease meta-analysis will consolidate the findings from many existing gut 16S microbiome studies, synthesizing our existing knowledge and generating new hypotheses to inform future mechanistic experiments and case-control analyses.

3.2.1 Compile and process gut microbiome datasets

To perform a meta-analysis, we need to collect a comprehensive selection of 16S gut microbiome case-control studies. We will identify these studies through a targeted literature search. Inclusion criteria for datasets is those that sequenced any region of the 16S rRNA gene of human stool samples, with at least 15 patients in the case (i.e. disease) group. Studies which focus exclusively on children under 5 will be excluded from these analyses, as the infant gut microbiome does not resemble that of adults [49]. We will also consider only datasets with publicly-available data, either from data repositories like SRA or from personal email communication with authors. We will not apply for special permissions to use IRB-protected data.

We will process these datasets using a standardized in-house pipeline developed by Thomas Gurry, a post-doc in the Alm lab. We will start with the rawest available data - in most cases, these will be fastq files but for some studies we will begin from quality-filtered fasta files. Sequences will be quality and length trimmed, clustered at 100% similarity, and assigned Latin taxonomic names using the RDP classifier. Samples with fewer than 100 reads will be removed from consideration. OTUs with fewer than 10 reads or which are present in less than 1% of samples will be removed. More stringent quality filtering may be considered in order to reduce noise in the dataset.

Different studies sequence different regions of the 16S gene, preventing us from using open-reference OTU sequences to compare microbes across studies. Therefore, we will collapse OTUs based on their taxonomic assignment and compare these across studies. Analyzing OTUs at the species or strain level would be ideal, but 16S data is limited in that most OTUs cannot be classified down to such taxonomic resolution. On the other hand, the majority of gut microbiome OTUs tend to be classifiable to genus level. Thus, collapsing to the genus level provides an optimal balance between the amount of taxonomic information and the amount of data we have to discard (i.e. unannotated OTUs). Furthermore, previous work has found that the maximal predictive power of the microbiota to distinguish between different phenotypes occurs at various taxonomic thresholds [50], so we will also consider higher-level taxonomic classifications in our in-depth meta-analyses.

3.2.2 Identify microbial markers of disease

Once we have processed all datasets in a standardized way, our first goal is to identify microbial markers of general health and disease. For each study, we will compare macro-summaries of the microbial communities (i.e. alpha diversity, Bacteroides/Firmicutes ratio) between cases and controls. Then, we will look for specific microbes associated with health or disease using univariate non-parametric statistics to compare abundances of microbes in cases and controls. To identify microbes which are *consistently* associated with health or disease, we will combine the results from all studies using the weighted Z-test, a weighted method used in meta-analyses for combining p-values [62]. This will yield an overall significance for each microbe, incorporating the results from all case-control studies.

Next, we aim to identify microbes which are consistently associated with *specific* diseases. For diseases which have more than 3 studies, we will perform the same meta-analysis as above with only the studies of that disease. There are many interesting possible outcomes from

this analysis. First, we could find bacteria which are not associated with disease in general but which *are* associated with that specific disease. We could also find bacteria which are significant markers of disease in general and that specific disease, but whose direction of change may differ in the two cases. For example, a bacteria could be significantly higher in Parkinson's patients but significantly lower in all other disease states. The microbes we identify with this analysis may be very interesting candidates for biomarkers or mechanistic follow-up studies.

It is possible that we struggle to find bacteria consistently associated with diseases because of technical batch effects, even where we expect to find a clear signal (i.e. in diseases which have had clear results from experimental or mechanistic studies [23, 27, 35]). Developing robust methods to overcome technical batch effects in 16S studies is not within the scope of this work, but there are many simpler options available to help mitigate severe batch effects. We will first try applying simple linear correction methods like subtracting the components which correlate closely with technical artifacts like read depth to our datasets. If standard meta-analysis methods still fail after this correction, we will also change the way we compare OTUs across studies: rather than depending on taxonomically-assigned closed-reference OTUs, we could identify associations between open-referenced OTUs and diseases within individual studies, and then compare the phylogenetic relationship of significant OTUs across all studies. Finally, if taxonomic approaches still do not yield results, we will consider approaching our meta-analysis from a functional point of view by using tools like PiCRUST or HUMANN to assign functionality to our observed taxonomies [59, 60].

3.2.3 Compare results between studies for related diseases

Our next hypothesis is that similar diseases will have similar signatures of dysbiosis. For example, we expect that metabolic diseases like obesity and diabetes will have more similar microbiota changes than they will to diarrheal diseases like *Clostridium difficile* infection or enteric diarrhea. We will summarize each dataset with one vector indicating its "microbial signature". This signature will be based on number and identities of microbes significantly associated with the cases and the direction of change of these microbes relative to the controls. Depending on the results from Section 3.2.2, we may also include factors like differences in alpha diversity or Bacteroides/Firmicutes ratios. Then, we will investigate which datasets cluster together in this "signature space" (Figure 3).

If a disease has a strong impact on or association with the gut microbiome, then we would expect the signatures from multiple studies of that disease to cluster very tightly together. If this is the case, we can extract the bacterial features which contribute the most to this tight clustering - these will then be most likely to be associated with that specific disease, and would be good candidates for further mechanistic explorations. On the other hand, if datasets of the same disease or similar conditions do not have similar microbial signatures, this may indicate that the microbiome is not inherently implicated in or affected by the disease. In this case, any signal that we see in the gut microbiome is likely driven by other non-disease effects, which are not necessarily the same across studies. Finally, if we find different diseases with similar underlying causes (i.e. inflammation) clustering near each other, then perhaps this would indicate that the microbiome is affected or involved with the underlying cause rather than the specific diseases. Such insights could help us design better

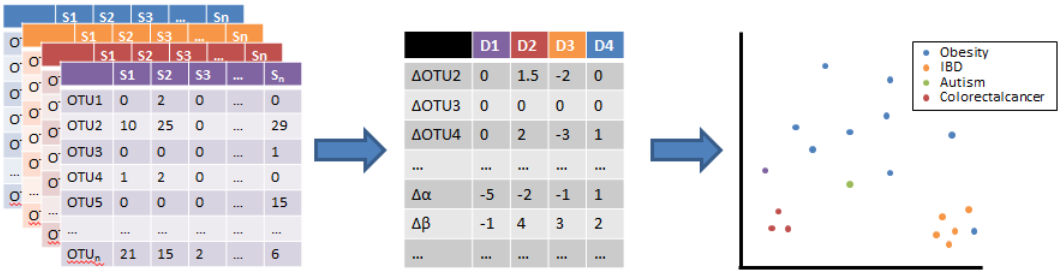

Figure 3: We will define a microbial signature for each dataset by summarizing the association of each of its OTUs with the disease of interest. We will also include other relevant features like alpha and beta diversities. These signature vectors will be used to cluster datasets and analyze resulting patterns.

experiments to follow up on mechanism or causal relationships.

3.3 Aim 3: Assigning bacteria to groups with similar functions and disease associations

In this aim, we will curate biologically-motivated *microbe sets* to enable easier interpretation of results from 16S microbiome analyses. By defining groups of related microbes, we will enable enrichment analyses like GSEA, but for microbiome data (Section 2.3.2, [55]). Enrichment analyses will allow for better biological understanding of results from individual studies as well as more consistent comparisons of results across studies reported in the literature.

To define *microbe sets* that facilitate enrichment analyses, we will begin by searching the literature for existing microbial annotation databases and papers which characterize broad groups of microbes. In parallel, we will also mine the datasets and results from Aim 2 for meaningful microbial associations with human phenotypes such as disease or inflammation. Finally, we will combine and package this information in a format that is easy to use and update as future studies contribute to the field. This tool will enable future microbiome scientists to extract more meaningful information from their microbiome studies, thus contributing significantly to increasing our understanding of the clinical and scientific relevance of the human microbiome.

3.3.1 Define microbe sets based on known biological relationships

Our first task is to curate and define *microbe sets* for use in enrichment analyses of microbiome datasets. We will perform an extensive literature search to identify existing databases and review papers with validated microbial phenotype annotations, with the goal of combining these databases and filling them out where they are missing annotations. We will begin with ImG which has approximately 10,000 annotated microbial genomes. About half of these genomes are human-associated bacteria, and half of those have annotations for categories like disease association, sporulation, and body site habitat. We will extract the 16S sequences for all the annotated human-associated microbes in this dataset, build a phylogenetic tree, and determine whether they adequately span the phylogenetic diversity we expect to find

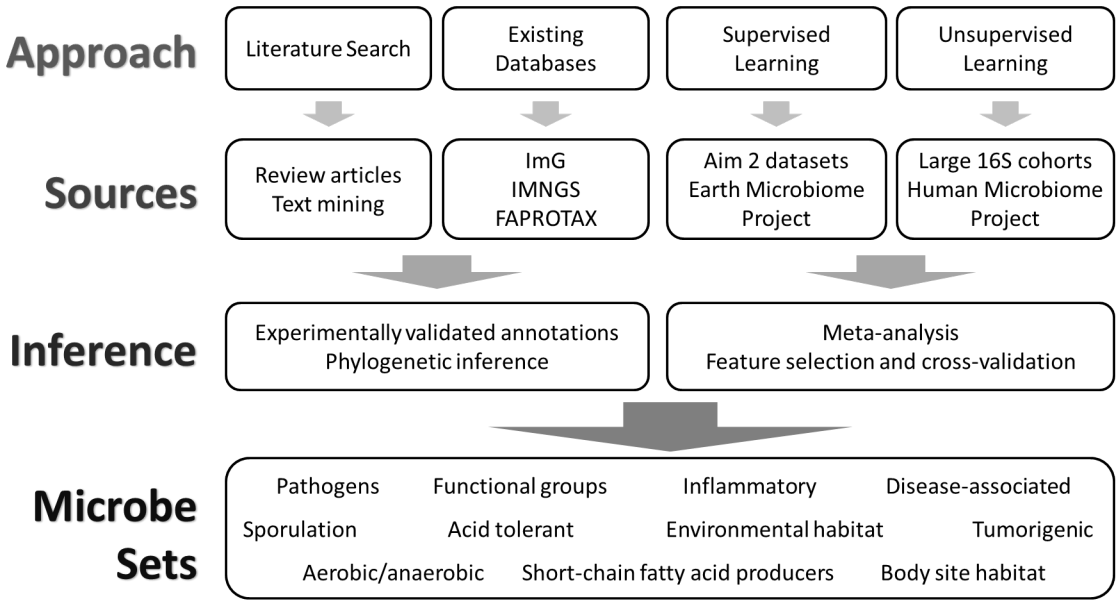

Figure 4: A variety of approaches will be used to define microbe sets, including manual curation from literature and database searches and data-driven methods ("Approach"). Many different kinds of resources will be drawn upon ("Sources") to infer groups of related microbes ("Inference") based on different categories ("Microbe Sets"). Databases described in [57, 64, 65]

in the human gut. If certain key clades are missing from these bacteria, we will manually include them from literature searches and NCBI queries.

In collaboration with Ilana Brito at Cornell, we will apply a combination of literature mining and bioinformatics approaches to fill out the missing annotations in the databases and to add our own fields of interest (Figure 4). We may also pursue unsupervised text mining of literature results to infer trait associations for well-studied bacteria [63]. As we populate our microbe annotations, we will also collect the relevant 16S sequences to build a tree. When possible, we will apply appropriate inference models to infer missing traits of leaf organisms.

3.3.2 Extract disease-associated microbe sets from datasets in Aim 2

In parallel, we will leverage the datasets collected in Aim 2 to extract additional disease-associated groups of microbes. We will define groups of microbes based on those that were markers for specific diseases and which distinguished broader phenotypes of health and disease in Section 3.2.3. We will also pursue machine learning-driven approaches to identify novel disease- or phenotype-associated *microbe sets*. We will restrict our feature space to genera which are present in the majority of studies in order to reduce the dimensionality of the problem and also to ensure that extracted groups are likely to be generalizeable to future studies. Random forest (RF) and support vector machine (SVM) classifiers are the most commonly used methods in microbiome datasets and have been shown to perform well in discriminating phenotypes based on 16S data [33, 48, 51]. We will apply these methods to various classification tasks (Table 2) and identify the most important features (RF) or those

Microbe set association	Classification task
General health/disease	All healthy vs. all disease
Diarrhea	CDI, EDD, IBS-D vs. controls
Neurological	Autism, Parkinson's vs. controls
Liver	NASH, MHE vs. controls
Metabolic syndrome	T1D, T2D, obesity, metabolic syndrome vs. controls
Autoimmune/inflammatory	T1D, rheumatoid arthritis, psoriatic arthritis, Crohn's disease vs. controls or non-autoimmune patients

Table 2: Classification tasks to identify groups of phenotype-associated microbes

with the highest support (SVM) in successful classifications. Cross-validating the extracted feature sets across different datasets with the same classification task will ensure the general discriminatory power of those microbes and prevent over-fitting.

These approaches may yield no *microbe sets* with phenotype associations that we are confident enough in to include in future enrichment analyses. Considering the diversity of the gut microbiome across people and the sparsity and high-dimensionality of 16S data, this result would not be surprising and underscores the importance of the manual curation work in Section 3.3.1 [50, 54]. If this happens, we will investigate higher-order taxa as possible *microbe set*-defining features, as this will reduce the dimensionality, sparsity, and inter-personal variability in the datasets. Another approach could be to convert each 16S community to functional profiles [59] and perform feature selection on these functional community profiles, and then convert these selected functions into taxonomically-defined groups of microbes. We could convert discriminatory functions back to taxa by either identifying the bacteria which most frequently have the discriminating function(s) in the datasets of interest, or by identifying all bacteria which have that function across all datasets.

3.3.3 Develop collaborative tool for interpreting microbiome studies

We will make the microbe set annotations available to researchers for further research and development. Through our literature searches, we will identify to the format of databases researchers have found most useful and strive to package our *microbe set* annotations in an easy-to-use format. We will likely begin with one very large text file containing all of the microbes, 16S sequences, and metadata that we have gathered. In addition to distributing the annotations and *microbe set* definitions themselves, we will also package them into a tool that researchers can use to interpret their 16S studies. Our software will take as input an OTU table and labels for different categories of samples (i.e. cases and controls). It will perform enrichment analysis on the OTU table and return the results to researchers, similar to the Broad's GSEA tool [55]. All of this work will be done using public open-source tools like GitHub to encourage collaboration and dissemination of our findings.

Developing a database of microbial annotations is a daunting task due to the vast diversity and complexity of microbes. We recognize the inherent difficulty of this task, and do not

expect to produce a fully comprehensive database. However, because our annotations are intended to serve as a tool for biological interpretations and hypothesis generation, even a partially-complete database will be extremely valuable in reducing the number of false-negative results in case- control studies and also providing coherent biological interpretations of existing results. We also recognize that our work will contribute to the beginning of a systematic grouping of phenotypically-associated microbes, and so we will ensure that the architecture of our database and annotation tool is easily accessible and modifiable by other researchers.

4 Preliminary studies

Three to four pages

4.1 Aim 1

4.1.1 Microbiome community exchange

Using our definition of p_s (3.1.2), we identified over 100 OTUs exchanged across sites of the aerodigestive tract. As expected, the majority of these were exchanged between the throat and stomach. Interestingly, the stomach and lungs also had a significant amount of microbial exchange, and these communities were almost as similar to each other as the throat and stomach communities were. These findings support the hypothesis that frequent microaspiration of gastric contents into the lungs is occurring. While certain bacteria may be preferentially exchanged between the lungs and stomach, the lower number of exchanged microbes as compared to the throat and lungs, combined with the similar similarity of communities, indicates that much of the exchange may be more stochastic, and not selecting for specific community members across many patients. Finally, we observed a decreasing trend in number of exchange microbes across throat-stomach, lung-stomach, and throat-lung sites, respectively, for all phyla except Proteobacteria. More Proteobacteria were exchanged between lungs and stomach than throat and stomach. Proteobacteria are known aerobes, and so may be preferentially selected for colonization in the lungs after microaspiration from the stomach.

4.1.2 Modulators of community exchange

We observed a distinct difference in the amount of exchange between the throat and lungs of aspirators versus patients with normal MBS results. We looked at the previously-defined microbes that were exchanged between the throat and lungs and re-caculated p_s, the percent of patients who have that microbe present in both sites, for the aspirator and non-aspirator groups separately. 25% of aspirators shared these microbes across their throats and lungs, while only 13% of non-aspirators did. Additionally, the throat and lung communities were significantly more similar in aspirating patients than non-aspirators. These results indicate that a consequence of abnormal swallowing dysfunction is likely a seeding of the lungs with oral bacteria. Interestingly, the stomach and lung communities of aspirating patients were slightly more similar to each other than in non-aspirators. By definition, aspirating patients

Dataset ID	Diseases	Median reads/sample	Year	Platform	Region

Table 3: Datasets

are unable to protect their airways from foreign material - while most apparent in swallowing, such dysfunction may also enable more material to move between the stomach and lungs as well.

The effect of reflux on the stomach-lung connection was less striking but still apparent. Specifically, the stomach and lung communities were more similar to each other in patients with very frequent full-column reflux. However, here was no significant difference in the extent of microbial exchange between the stomach and lungs. In other words, the previously-defined stomach-lung exchanged microbes were as likely to be found in both sites in patients with frequent full-column reflux than in patients with less severe reflux. This may indicate a more stochastic seeding of the lungs from the stomach rather than reproducible exchange of the same microbes across patients. In this case, stomach microbes are not necessarily growing in the lungs but nonetheless do have measurable impact on the lung's microbial community.

4.2 Aim 2

4.2.1 Collecting and reprocessing 16S case-control datasets

Our extensive literature review has so far identified 54 suitable case-control 16S datasets. 29 datasets and their associated metadata have been downloaded and processed through our in-house pipeline. Characteristics of these datasets are shown in Table 3

4.2.2 Identify general patterns of health and diseases

We first summarized overall community structure using Shannon's alpha diversity index (SDI). We observed significant batch effects across studies, likely due to the differences in sequencing depth. After standardizing SDI within studies and combining similar disease states, we observed little difference in overall community structure between cases and controls. An exception to this was seen in diarrheal diseases (*Clostridium difficile* infection and enteric diarrheal disease), in which alpha diversity was significantly lower in the cases.

We next performed univariate comparisons at the genus level for abundance in cases vs. controls for each study. This analysis revealed a similar result: diarrheal diseases showed striking shifts in many microbes, while other diseases had less obvious patterns of dysbiosis. We observed that many bacteria seemed to have relatively consistent shifts between cases and controls across many different diseases **??**. This supports the hypothesis that there is a general signature for disease, in other words, that sick people have altered microbiomes. In light of this finding, identifying microbial shifts that are unique to individual diseases will be crucial to finding specific biomarkers for diagnostic purposes and to motivate mechanistic investigations.

5 Gud werds

We hypothesize that there is a clinically-relevant exchange of bacteria within the aerodigestive tract that may be altered in certain disease states.

Another important consideration in this work may be if study-associated effects are larger than biological effects. For example, when we compare 'microbial signatures' across datasets, it's possible that the largest signal driving dataset clustering is sequencer or 16S region sequenced, rather than disease state. There are many approaches we could take to correct for such batch effects:

1. Subtracting the principal components corresponding to the technical artifacts.
2. Build a model that accounts for these technical artifacts by including them as factors in the model.
3. Non-parametric correction, like sample- or OTU-wise quantile normalization, using controls in each study as the reference distribution.

This work will improve our understanding of the clinical relevance of the human microbiome and will also provide new approaches and tools for analyzing future studies.

We will summarize each dataset's microbial communities methods commonly used in the literature: univariate non-parameteric statistical tests on relative abundances, alpha and beta diversity in different types of patients, and ratios of Firmicutes to Bacteroides in healthy vs. disease patients. It is generally thought that low alpha diversity is a marker of dysbiosis (REF), and that while most people have a Firmicutes/ Bacteroides ratio of (XXX), in certain diseases this ratio may be different (REF). (MAYBE BACKGROUND?) By analyzing each study in the same way from raw data, we can reduce the study-wise batch effects and increase our ability to identify general trends in the gut microbiome in health and disease. We will identify consistent markers of disease by using standard meta-analysis methods, comparing the effect sizes and directionality of bacteria across studies, and using Fisher's method to determine overall significant of a microbe (REFS).

"Thus, humans are super organisms: http://science.sciencemag.org/content/312/5778/1355"

The microbiota of the aerodigestive tract is poorly studied, and we have little understanding of how the microbial communities in different aeordigest sites are related or affected by disease. In contrast, the gut microbiome has been extensively studied through many case-control studies. However, these studies have frequently yielded inconsistent or incomparable results. Existing meta-analyses have not extended to more than one or two diseases, and thus can not determine whether significant microbes are associated with specific diseases or with disease in general. Finally, there are no existing tools that can be used to extract general biological insights from groups of disease-associated microbes.

Fecal microbiota transplants have demonstrated the incredible causal ability of the microbiome to affect health in human patients. Germ-free mouse models have shown that microbes are necessary for healthy functioning, and specific animal models have allowed for probing mechanistic understanding of host-microbial interactions (autism 440 ref, gordon mouse experiment ref).

Additionally, many of the most successful existing meta-analyses combine vastly different types of microbial communities and non-case- control experimental designs. The positive results from these studies are not particularly biologically insightful: it is a much easier task to differentiate vastly different communities (like the skin vs. the gut) than it is to

Category	Approach
Pathogens	Targeted literature search, literature mining, & databases
Body sites	Literature mining, machine learning on HMP data
Environmental associations	Literature mining, machine learning on EMP data
Growth rate	Inference from 16S sequences in datasets from Aim 2 and HMP
Obesity-associated	Targeted literature search & machine learning on Aim 2 datasets
Inflammation-associated	Targeted literature search & machine learning on Aim 2 datasets
Miscellaneous functions (acid-tolerant, mucus-degrading, etc)	Targeted literature search, unsupervised PiCRUST clustering, genome mining

Table 4: Possible approaches to define microbe sets of interest

differentiate subtleties contributing to health and disease (like the inflamed gut vs. the healthy gut) [48].

Every meta-analysis performed on 16S data has observed strong batch effects between studies and noted the need for large sample sizes to extract any meaningful signal [47], [29],[48], [49].

This definition depends on correlated abundances and not simply co-occurence, and is superior to simple co-occurence because of the significant overlap in the members of the three aeordigestive communities[1], [16].

To determine how lung, gastric, and throat microbial communities are related, we will calculate p_s for each of the site combinations. We will then see if there are apparent phylogenetic relationships between the exchanged microbes. We expect to find significantly more exchange between the throat and stomach, and very little exchange between the throat and lungs.

This work will be the beginning of what will hopefully become a new approach to interpreting 16S datasets - moving the field from asking simply "what's different?" toward a more critical interpretation of "why are things different?"

References

[1] C.M. Bassis, J.R. Erb-Downward, R.P. Dickson, C.M. Freeman, T.M. Schmidt, V.B. Young, J.M. Beck, J.L. Curtis, and G.B. Huffnagle. Analysis of the upper respiratory tract microbiotas as the source of the lung and gastric microbiotas in healthy individuals. *MBio*, 6(2):e00037–15, 2015. doi: 10.1128/mBio.00037-15. URL `http://dx.doi.org/10.1128/mBio.00037-15`.

[2] J.M. Beck, V.B. Young, and G.B. Huffnagle. The microbiome of the lung. *Translational Research*, 160(4):258–266, 2012. doi: 10.1016/j.trsl.2012.02.005. URL `http://dx.doi.org/10.1016/j.trsl.2012.02.005`.

[3] B. Martin-Harris and B. Jones. The videofluorographic swallowing study. *Physical medicine and rehabilitation clinics of North America*, 19(4):769–778, 2008. doi: 10.1016/j.pmr.2008.06.004. URL http://dx.doi.org/10.1016/j.pmr.2008.06.004.

[4] R.P. Dickson, F.J. Martinez, and G.B. Huffnagle. The role of the microbiome in exacerbations of chronic lung diseases. *The Lancet*, 384 (9944):691–702, 2014. doi: 10.1016/S0140-6736(14)61136-3. URL http://dx.doi.org/10.1016/S0140-6736(14)61136-3.

[5] N. Vakil, S.V. Van Zanten, P. Kahrilas, J. Dent, and R. Jones. The montreal definition and classification of gastroesophageal reflux disease: a global evidence-based consensus. *The American journal of gastroenterology*, 101(8):1900–1920, 2006. doi: doi:10.1111/j.1572-0241.2006.00630.x. URL http://dx.doi.org/10.1111/j.1572-0241.2006.00630.x.

[6] J. Dent, H.B. El-Serag, M. Wallander, and S. Johansson. Epidemiology of gastro-oesophageal reflux disease: a systematic review. *Gut*, 54(5):710–717, 2005. doi: 10.1136/gut.2004.051821. URL http://dx.doi.org/10.1136/gut.2004.051821.

[7] M.P. Sweet, M.G. Patti, C. Hoopes, S.R. Hays, and J.A. Golden. Gastro-oesophageal reflux and aspiration in patients with advanced lung disease. *Thorax*, 64(2):167–173, 2009. doi: 10.1136/thx.2007.082719. URL http://dx.doi.org/10.1136/thx.2007.082719.

[8] F. Imhann, M.J. Bonder, A.V. Vila, J. Fu, Z. Mujagic, L. Vork, E.F. Tigchelaar, S.A. Jankipersadsing, M.C. Cenit, H.J. Harmsen, and G. Dijkstra. Proton pump inhibitors affect the gut microbiome. *Gut*, 65(5):740–748, 2016. doi: 10.1053/j.gastro.2015.06.043. URL http://dx.doi.org/10.1053/j.gastro.2015.06.043.

[9] L. A. Houghton, A. S. Lee, H. Badri, K. R. DeVault, and J. A. Smith. Respiratory disease and the oesophagus: reflux, reflexes and microaspiration. *Nature Reviews Gastroenterology & Hepatology*, 13(8):445–460, 2016. doi: 10.1038/nrgastro.2016.91. URL http://dx.doi.org/10.1038/nrgastro.2016.91.

[10] K. Raghavendran, J. Nemzek, L.M. Napolitano, and P.R. Knight. Aspiration-induced lung injury. 39(4):818–826, 2011. doi: 10.1097/CCM.0b013e31820a856b. URL http://dx.doi.org/10.1097/CCM.0b013e31820a856b.

[11] B. Martin-Harris, J.A. Logemann, S. McMahon, M. Schleicher, and J. Sandidge. Clinical utility of the modified barium swallow. *Dysphagia*, 15(3):136–141, 2000. doi: 10.1007/s004550010015. URL http://dx.doi.org/10.1007/s004550010015.

[12] A. Lee, E. Festic, P.K. Park, K. Raghavendran, O. Dabbagh, A. Adesanya, O. Gajic, and R.R. Bartz. Characteristics and outcomes of patients hospitalized following pulmonary aspiration. *Chest*, 146(4):899–907, 2014. doi: 10.1378/chest.13-3028. URL http://dx.doi.org/10.1378/chest.13-3028.

[13] F.J. Reen, D.F. Woods, Mooij, M.J., M.N. Chrinn, D. Mullane, L. Zhou, J. Quille, D. Fitzpatrick, J.D. Glennon, G.P. McGlacken, and C. Adams. Aspirated bile: a major host trigger modulating respiratory pathogen colonisation in cystic fibrosis patients. *European Journal of Clinical Microbiology & Infectious Diseases*, 33(10):1763–1771, 2014. doi: doi:10.1007/s10096-014-2133-8. URL http://dx.doi.org/10.1007/s10096-014-2133-8.

[14] H. Al-Momani, A. Perry, C.J. Stewart, R. Jones, A. Krishnan, Robertson, A.G., S. Bourke, S. Doe, S.P. Cummings, A. Anderson, and T. Forrest. Microbiological profiles of sputum and gastric juice aspirates in cystic fibrosis patients. *Scientific Reports*, 6, 2016. doi: doi:10.1038/srep26985. URL http://dx.doi.org/10.1038/srep26985.

[15] R. Rosen, L. Hu, J. Amirault, U. Khatwa, D.V. Ward, and A. Onderdonk. 16S community profiling identifies proton pump inhibitor related differences in gastric, lung, and oropharyngeal microflora. *The Journal of pediatrics*, 166(4):917–923, 2015. doi: 10.1016/j.jpeds.2014.12.067. URL http://dx.doi.org/10.1016/j.jpeds.2014.12.067.

[16] E.S. Charlson, K. Bittinger, A.R. Haas, A.S. Fitzgerald, I. Frank, A. Yadav, F.D. Bushman, and R.G. Collman. Topographical continuity of bacterial populations in the healthy human respiratory tract. *American journal of respiratory and critical care medicine*, 184(8):957–963, 2011. doi: 10.1164/rccm.201104-0655OC. URL http://dx.doi.org/10.1164/rccm.201104-0655OC.

[17] J.R. Erb-Downward, D.L. Thompson, M.K. Han, C.M. Freeman, L. McCloskey, L.A. Schmidt, V.B. Young, G.B. Toews, J.L. Curtis, B. Sundaram, and F.J. Martinez. Analysis of the lung microbiome in the healthy smoker and in COPD. *PloS one*, 6(2):e16384, 2011. doi: 10.1371/journal.pone.0016384. URL http://dx.doi.org/10.1371/journal.pone.0016384.

[18] R.P. Dickson, J.R. Erb-Downward, C.M. Freeman, L. McCloskey, J.M. Beck, G.B. Huffnagle, and J.L. Curtis. Spatial variation in the healthy human lung microbiome and the adapted island model of lung biogeography. *Annals of the American Thoracic Society*, 12(6):821–830, 2015. doi: 10.1513/AnnalsATS.201501-029OC. URL http://dx.doi.org/10.1513/AnnalsATS.201501-029OC.

[19] The noncolonic microbiome: does it really matter? *Current gastroenterology reports*, 12(4):259–262, 2010. doi: 10.1007/s11894-010-0111-6. URL http://dx.doi.org/10.1007/s11894-010-0111-6.

[20] E.M. Bik, P.B. Eckburg, S.R. Gill, K.E. Nelson, E.A. Purdom, F. Francois, G. Perez-Perez, M.J. Blaser, and D.A. Relman. Molecular analysis of the bacterial microbiota in the human stomach. *Proceedings of the National Academy of Sciences of the United States of America*, 103(3):732–737, 2006. doi: 10.1073/pnas.0506655103. URL http://dx.doi.org/10.1073/pnas.0506655103.

[21] L. Zhu, S.S. Baker, C. Gill, W. Liu, R. Alkhouri, R.D. Baker, and S.R. Gill. Characterization of gut microbiomes in nonalcoholic steatohepatitis (NASH) patients: a

connection between endogenous alcohol and NASH. *Hepatology*, 57(2):601–609, 2013. doi: 10.1002/hep.26093. URL http://dx.doi.org/10.1002/hep.26093.

[22] D.W. Kang, J.G. Park, Z.E. Ilhan, G. Wallstrom, J. LaBaer, J.B. Adams, and R. Krajmalnik-Brown. Reduced incidence of Prevotella and other fermenters in intestinal microflora of autistic children. *PloS one*, 8(7):e68322, 2013. doi: 10.1371/journal.pone.0068322. URL http://dx.doi.org/10.1371/journal.pone.0068322.

[23] P.J. Turnbaugh, R.E. Ley, M.A. Mahowald, V. Magrini, E.R. Mardis, and J.I. Gordon. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122):1027–1131, 2006. doi: 10.1038/nature05414. URL http://dx.doi.org/10.1038/nature05414.

[24] J. Son, L.J. Zheng, L.M. Rowehl, X. Tian, Y. Zhang, W. Zhu, L. Litcher-Kelly, K.D. Gadow, G. Gathungu, C.E. Robertson, D. Ir, D.N. Frank, and E. Li. Comparison of fecal microbiota in children with autism spectrum disorders and neurotypical siblings in the simons simplex collection. *PLOS ONE*, 10(10):e0137725, 2015. doi: 10.1371/journal.pone.0137725. URL http://dx.doi.org/10.1371/journal.pone.0137725.

[25] P. Singh, T.K. Teal, T.L. Marsh, J.M. Tiedje, R. Mosci, K. Jernigan, A. Zell, D.W. Newton, H. Salimnia, P. Lephart, D. Sundin, W. Khalife, R.A. Britton, J.T. Rudrik, and S.D. Manning. Intestinal microbial communities associated with acute enteric infections and disease recovery. *Microbiome*, 3(1), sep 2015. doi: 10.1186/s40168-015-0109-2. URL http://dx.doi.org/10.1186/s40168-015-0109-2.

[26] F. Scheperjans, V. Aho, P.A.B. Pereira, K. Koskinen, L. Paulin, E. Pekkonen, E. Haapaniemi, S. Kaakkola, J. Eerola-Rautio, P. Pohja, E. Kinnunen, K. Murros, and P. Auvinen. Gut microbiota are related to parkinson's disease and clinical phenotype. *Movement Disorders*, 30(3):350–358, dec 2014. doi: 10.1002/mds.26069. URL http://dx.doi.org/10.1002/mds.26069.

[27] V. K. Ridaura, J. J. Faith, F. E. Rey, J. Cheng, A. E. Duncan, A. L. Kau, N. W. Griffin, V. Lombard, B. Henrissat, J. R. Bain, M. J. Muehlbauer, O. Ilkayeva, C. F. Semenkovich, K. Funai, D. K. Hayashi, B. J. Lyle, M. C. Martini, L. K. Ursell, J. C. Clemente, W. Van Treuren, W. A. Walters, R. Knight, C. B. Newgard, A. C. Heath, and J. I. Gordon. Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science*, 341(6150):1241214–1241214, sep 2013. doi: 10.1126/science.1241214. URL http://dx.doi.org/10.1126/science.1241214.

[28] J.S. Escobar, B. Klotz, B.E. Valdes, and G.M. Agudelo. The gut microbiota of colombians differs from that of americans, europeans and asians. *BMC microbiology*, 14(1):1, 2014. doi: 10.1186/s12866-014-0311-6. URL http://dx.doi.org/10.1186/s12866-014-0311-6.

[29] Walters W., Xu Z., and Knight R. Meta-analyses of human gut microbes associated with obesity and ibd. *FEBS Letters*, 588:4223–4233, 2014. doi: 10.1016/j.febslet.2014.09.039. URL http://dx.doi.org/10.1016/j.febslet.2014.09.039.

[30] V. Stadlbauer, B. Leber, S. Lemesch, S. Trajanoski, M. Bashir, A. Horvath, M. Tawdrous, T. Stojakovic, G. Fauler, P. Fickert, C. Högenauer, I. Klymiuk, P. Stiegler, M. Lamprecht, T.R. Pieber, N.J. Tripolt, and H Sourij. Lactobacillus casei shirota supplementation does not restore gut microbiota composition and gut barrier in metabolic syndrome: A randomized pilot study. *PLOS ONE*, 10(10):e0141399, 2015. doi: 10.1371/journal.pone.0141399. URL http://dx.doi.org/10.1371/journal.pone.0141399.

[31] M.C. Ross, D.M. Muzny, J.B. McCormick, R.A. Gibbs, S.P. Fisher-Hoch, and J.F. Petrosino. 16s gut community of the cameron county hispanic cohort. *Microbiome*, 3(1):7, 2015. doi: 10.1186/s40168-015-0072-y. URL http://dx.doi.org/10.1186/s40168-015-0072-y.

[32] C.P. Tamboli, C. Neut, P. Desreumaux, and J.F. Colombel. Dysbiosis in inflammatory bowel disease. *Gut*, (1):1–4, 2004. doi: 10.1136/gut.53.1.1. URL http://dx.doi.org/10.1136/gut.53.1.1.

[33] E. Papa, M. Docktor, C. Smillie, S. Weber, S.P. Preheim, D. Gevers, G. Giannoukos, D. Ciulla, D. Tabbaa, J. Ingram, D.B. Schauer, D.V. Ward, J.R. Korzenik, R.J. Xavier, A. Bousvaros, and E.J. Alm. Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease. *PLoS ONE*, 7(6):e39242, 2012. doi: 10.1371/journal.pone.0039242. URL http://dx.doi.org/10.1371/journal.pone.0039242.

[34] D. Gevers, S. Kugathasan, L.A. Denson, Y. Vázquez-Baeza, W. Van Treuren, B. Ren, E. Schwager, D. Knights, S. Song, M. Yassour, X.C. Morgan, A.D. Kostic, C. Luo, A. González, D. McDonald, Y. Haberman, T. Walters, S. Baker, J. Rosh, M. Stephens, M. Heyman, J. Markowitz, R. Baldassano, A. Griffiths, F. Sylvester, D. Mack, S. Kim, W. Crandall, J. Hyams, C. Huttenhower, R. Knight, and R. Xavier. The treatment-naive microbiome in new-onset crohn's disease. *Cell Host & Microbe*, 15(3):382–392, mar 2014. doi: 10.1016/j.chom.2014.02.005. URL http://dx.doi.org/10.1016/j.chom.2014.02.005.

[35] G. Zeller, J. Tap, A.Y. Voigt, S. Sunagawa, J.R. Kultima, P.I. Costea, A. Amiot, J. Bohm, F. Brunetti, N. Habermann, R. Hercog, M. Koch, A. Luciani, D.R. Mende, M.A. Schneider, P. Schrotz-King, C. Tournigand, J.T. Nhieu, T. Yamada, J. Zimmermann, V. Benes, M. Kloor, C.M. Ulrich, M. von Knebel Doeberitz, I. Sobhani, and P. Bork. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology*, 10(11):766–766, 2014. doi: 10.15252/msb.20145645. URL http://dx.doi.org/10.15252/msb.20145645.

[36] W. Chen, F. Liu, Z. Ling, X. Tong, and C. Xiang. Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. *PLoS ONE*, 7(6):e39743, 2012. doi: 10.1371/journal.pone.0039743. URL http://dx.doi.org/10.1371/journal.pone.0039743.

[37] T. Wang, G. Cai, Y. Qiu, N. Fei, M. Zhang, X. Pang, W. Jia, S. Cai, and L. Zhao. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *The ISME Journal*, 6(2):320–329, 2011. doi: 10.1038/ismej.2011.109. URL http://dx.doi.org/10.1038/ismej.2011.109.

[38] N. Wu, X. Yang, R. Zhang, J. Li, X. Xiao, Y. Hu, Y. Chen, F. Yang, N. Lu, Z. Wang, C. Luan, Y. Liu, B. Wang, C. Xiang, Y. Wang, F. Zhao, G.F. Gao, S. Wang, L. Li, H. Zhang, and B. Zhu. Dysbiosis signature of fecal microbiota in colorectal cancer patients. *Microb Ecol*, 66(2):462–470, 2013. doi: 10.1007/s00248-013-0245-9. URL http://dx.doi.org/10.1007/s00248-013-0245-9.

[39] J.P. Zackular, M.A.M. Rogers, M.T. Ruffin, and P.D. Schloss. The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prevention Research*, 7(11):1112–1121, 2014. doi: 10.1158/1940-6207.capr-14-0129. URL http://dx.doi.org/10.1158/1940-6207.CAPR-14-0129.

[40] A.M. Schubert, M.A. Rogers, C. Ring, J. Mogle, J.P. Petrosino, V.B. Young, D.M. Aronoff, and P.D. Schloss. Microbiome data distinguish patients with clostridium difficile infection and non-c. difficile-associated diarrhea from healthy controls. *mBio*, 5(3):e01021–14–e01021–14, 2014. doi: 10.1128/mbio.01021-14. URL http://dx.doi.org/10.1128/mBio.01021-14.

[41] C. Vincent, D.A. Stephens, V.G. Loo, T.J. Edens, M.A. Behr, K. Dewar, and A.R. Manges. Reductions in intestinal clostridiales precede the development of nosocomial clostridium difficile infection. *Microbiome*, 1(1):18, 2013. doi: 10.1186/2049-2618-1-18. URL http://dx.doi.org/10.1186/2049-2618-1-18.

[42] J.U. Scher, A. Sczesnak, R.S. Longman, N. Segata, C. Ubeda, C. Bielski, T. Rostron, V. Cerundolo, E.G. Pamer, S.B. Abramson, C. Huttenhower, and D.R. Littman. Expansion of intestinal prevotella copri correlates with enhanced susceptibility to arthritis. *eLife*, 2, 2013. doi: 10.7554/elife.01202. URL http://dx.doi.org/10.7554/eLife.01202.

[43] D.M. Dinh, G.E. Volpe, C. Duffalo, S. Bhalchandra, A.K. Tai, A.V. Kane, C.A. Wanke, and H.D. Ward. Intestinal microbiota, microbial translocation, and systemic inflammation in chronic HIV infection. *Journal of Infectious Diseases*, 211(1):19–27, 2014. doi: 10.1093/infdis/jiu409. URL http://dx.doi.org/10.1093/infdis/jiu409.

[44] Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, 2012. doi: 10.1038/nature11234. URL http://dx.doi.org/10.1038/nature11234.

[45] L.A. David, A.C. Materna, J. Friedman, M.I. Campos-Baptista, M.C. Blackburn, A. Perrotta, S.E. Erdman, and E.J. Alm. Host lifestyle affects human microbiota on daily timescales. *Genome Biol*, 15(7):R89, 2014. doi: 10.1186/gb-2014-15-7-r89. URL http://dx.doi.org/10.1186/gb-2014-15-7-r89.

[46] E.K. Costello, K. Stagaman, L. Dethlefsen, B.J.M. Bohannan, and D.A. Relman. The application of ecological theory toward an understanding of the human microbiome. *Science*, 336(6086):1255–1262, 2012. doi: 10.1126/science.1224203. URL http://dx.doi.org/10.1126/science.1224203.

[47] Sze M.A. and Schloss P.D. Looking for a signal in the noise: revisiting obesity and the microbiome. *mBio*, 7(4):e01018–16, 2016. doi: 10.1128/mBio.01018-16. URL http://dx.doi.org/10.1128/mBio.01018-16.

[48] D. Knights, E. Costello, and R. Knight. Supervised classification of the human microbiota. *FEMS Microbiology Reviews*, 35:343–359, 2010. doi: 10.1111/j.1574-6976.2010.00251.x. URL http://dx.doi.org/10.1111/j.1574-6976.2010.00251.x.

[49] C.A. Lozupone, J. Stombaugh, A. Gonzalez, G. Ackermann, D. Wendel, Y. Vzquez-Baeza, J.K. Jansson, J.I. Gordon, and R. Knight. Meta-analyses of studies of the human microbiota. *Genome research*, 23(10):1704–1714, 2013. doi: 10.1101/gr.151803.112. URL http://dx.doi.org/10.1101/gr.151803.112.

[50] D. Knights, L. Parfrey, J. Zaneveld, C. Lozupone, and R. Knight. Human-associated microbial signatures: Examining their predictive value. *Cell Host & Microbe*, 10(4):292–296, 2011. doi: 10.1016/j.chom.2011.09.003. URL http://dx.doi.org/10.1016/j.chom.2011.09.003.

[51] E. Pasolli, D.T. Truong, F. Malik, L. Waldron, and N. Segata. Machine learning meta-analysis of large metagenomic datasets: Tools and biological insights. *PLOS Computational Biology*, 12(7):e1004977, 2016. doi: 10.1371/journal.pcbi.1004977. URL http://dx.doi.org/10.1371/journal.pcbi.1004977.

[52] Q. Wang, G.M. Garrity, J.M. Tiedje, and J.R. Cole. Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73(16):5261–5267, jun 2007. doi: 10.1128/aem.00062-07. URL http://dx.doi.org/10.1128/AEM.00062-07.

[53] D. McDonald, M.N. Price, J. Goodrich, E.P. Nawrocki, T.Z. DeSantis, A. Probst, G.L. Andersen, R. Knight, and P. Hugenholtz. An improved greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal*, 6(3):610–618, dec 2011. doi: 10.1038/ismej.2011.139. URL http://dx.doi.org/10.1038/ismej.2011.139.

[54] F. Wang, J. Kaplan, B. Gold, M. Bhasin, N. Ward, R. Kellermayer, B. Kirschner, M. Heyman, S. Dowd, S. Cox, H. Dogan, B. Steven, G. Ferry, S. Cohen, R. Baldassano, C. Moran, E. Garnett, L. Drake, H. Otu, L. Mirny, T. Libermann, H. Winter, and K. Korolev. Detecting microbial dysbiosis associated with pediatric crohn disease despite the high variability of the gut microbiota. *Cell Reports*, 14(4):945–955, 2016. doi: 10.1016/j.celrep.2015.12.088. URL http://dx.doi.org/10.1016/j.celrep.2015.12.088.

[55] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, and J. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*, 102(43):15545–15550, 2005. doi: 10.1073/pnas.0506580102. URL http://dx.doi.org/10.1073/pnas.0506580102.

[56] J. Xia and D. Wishart. MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res*, 38:W71–W77, 2010. doi: 10.1093/nar/gkq329. URL http://dx.doi.org/10.1093/nar/gkq329.

[57] V.M. Markowitz, I.A. Chen, K. Palaniappan, K. Chu, E. Szeto, M. Pillay, A. Ratner, J. Huang, T. Woyke, M. Huntemann, I. Anderson, K. Billis, N. Varghese, K. Mavromatis, A. Pati, N.N. Ivanova, and N.C. Kyrpides. IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Research*, 42(D1):D560–D567, 2013. doi: 10.1093/nar/gkt963. URL http://dx.doi.org/10.1093/nar/gkt963.

[58] D. Knights, J. Kuczynski, E.S. Charlson, J. Zaneveld, M.C. Mozer, R.G. Collman, F.D. Bushman, R. Knight, and S.T. Kelley. Bayesian community-wide culture-independent microbial source tracking. *Nature methods*, 8(9):761–763, 2011. doi: 10.1038/nmeth.1650. URL http://dx.doi.org/10.1038/nmeth.1650.

[59] M.G. Langille, J. Zaneveld, J.G. Caporaso, D. McDonald, D. Knights, J.A. Reyes, J.C. Clemente, D.E. Burkepile, Thurber, R.L.V., R. Knight, and R.G. Beiko. Predictive functional profiling of microbial communities using 16s rRNA marker gene sequences. *Nature biotechnology*, 31(9):814–821, 2013. doi: 10.1038/nbt.2676. URL http://dx.doi.org/10.1038/nbt.2676.

[60] S. Abubucker, N. Segata, J. Goll, A.M. Schubert, J. Izard, B.L. Cantarel, B. Rodriguez-Mueller, J. Zucker, M. Thiagarajan, B. Henrissat, and O. White. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol*, 8(6):e1002358, 2012. doi: 10.1371/journal.pcbi.1002358. URL http://dx.doi.org/10.1371/journal.pcbi.1002358.

[61] R. Rosen, N. Johnston, K. Hart, U. Khatwa, E. Katz, and S. Nurko. Higher rate of bronchoalveolar lavage culture positivity in children with nonacid reflux and respiratory disorders. *The Journal of pediatrics*, 159(3):504–506, 2011. doi: 10.1016/j.jpeds.2011.05.021. URL http://dx.doi.org/10.1016/j.jpeds.2011.05.021.

[62] D.V. Zaykin. Optimally weighted Ztest is a powerful method for combining probabilities in metaanalysis. *Journal of evolutionary biology*, 24(8):1836–1841, 2011. doi: 10.1111/j.1420-9101.2011.02297.x. URL http://dx.doi.org/10.1111/j.1420-9101.2011.02297.x.

[63] J.O. Korbel, T. Doerks, L.J. Jensen, C. Perez-Iratxeta, S. Kaczanowski, S.D. Hooper, M.A. Andrade, and P. Bork. Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol*, 3(5):e134, 2005. doi: 10.1371/journal.pbio.0030134. URL http://dx.doi.org/10.1371/journal.pbio.0030134.

[64] S. Louca, L.W. Parfrey, and M. Doebeli. Decoupling function and taxonomy in the global ocean microbiome. *Science*, 353(6305):1272–1277, 2016. doi: 10.1126/science.aaf4507. URL http://dx.doi.org/10.1126/science.aaf4507.

[65] I. Lagkouvardos, D. Joseph, M. Kapfhammer, S. Giritli, M. Horn, D. Haller, and T. Clavel. IMNGS: A comprehensive open resource of processed 16s rRNA microbial profiles for ecology and diversity studies. *Scientific Reports*, 6:33721, 2016. doi: 10.1038/srep33721. URL http://dx.doi.org/10.1038/srep33721.