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 some body sites have been much less studied than others. Even extensively studied body sites lack a synthesized understanding of human-microbe-disease associations. Finally, interpreting exploratory microbiome analyses into biological hypotheses remains challenging due to the lack of centralized tools and databases for assigning biological traits to groups of microbes.

This thesis will expand our understanding of the clinical microbiome in three ways. First, I will quantify the exchange between microbial communities in the aerodigestive tract and investigate their associations with aspiration and gastro-esophogeal reflux disease. This work will increase our understanding of the under-studied microbial communities of the aerodigestive tract. Next, I will re-analyze published gut microbiome studies across many disease states to identify disease-specific microbes and shared microbial responses to disease. No comparative studies currently exist which examine microbial communities across a wide variety of diseases. Finally, I will curate a database which links microbes to their biological traits, thus enabling generalizable interpretations of results from microbiome analyses. 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 existing and future analyses.

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 quantify the exchange across microbial communities in the aerodigestive tract and examine their associations with clinical factors such as aspiration and gastro-esophogeal reflux disease. Second, I will re-analyze published gut microbiome studies in a broad comparative analysis. I will look for microbes which are frequently enriched or depleted across many disease states as well as ones which are specifically altered in only one or two diseases. Finally, I will curate a database that annotates bacterial sequences with their associated biological traits. The information in this database will be compiled by mining and combining existing literature, genomes, and datasets related to the human microbiome.

1.2 Specific aims

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

1. Quantify microbial exchange across sites in the aerodigestive tract.
2. Evaluate the effect of aspiration and gastro-esophogeal reflux disease on aerodigestive microbial communities.

Aim 2 Re-analyze published gut microbiome studies to identify disease-associated microbes and shared microbial responses to disease.

1. Compile and re-process publicly available case-control gut microbiome studies with a standardized method.
2. Identify generic microbial responses to disease shared across all studied disease states.
3. Identify microbes consistently associated with specific diseases.

Aim 3 Curate a database of 16S sequences annotated with associated general biological traits.

1. Combine existing databases and perform targeted literature searches to annotate microbes based on known biological traits.
2. Use machine-learning techniques to extract disease-associated groups of bacteria from datasets collected in Aim 2.
3. Develop these microbial annotations into a collaborative tool for use in interpreting new microbiome studies.

2 Background and significance

The topics addressed in this thesis are 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 related to the invasive sampling necessary to study the human aerodigestive tract [2].

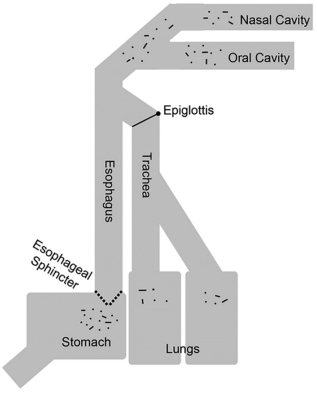

Figure 1: Schematic of the flow relationship between sites of the aerodigestive tract. Adapted from [1].

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, 13], but they are thought to play a role in causing or exacerbating many respiratory diseases [9, 14, 15]. 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, 13].

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 [15]. 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 [14, 15]. 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, 13, 15].

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]. 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 precise 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 [15, 17, 18]. However, many of these studies are limited by small sample sizes due to the invasive nature of sampling human lungs. 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 *Helicobacter 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 of healthy patients' stomachs [1, 17, 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]. Previous studies have examined the relationships between microbial communities in the upper aerodigestive tract, with conflicting results [1, 15, 17, 16]. Some have shown striking similarities in the microbial communities across the respiratory tract [1, 15] while others argue that different sites have distinct microbial communities [17], and that these vary even within individual sites like the lungs [18, 21].

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 [22, 23, 24]. 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' [23, 25].

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 [22, 24, 26, 27]. 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 [24, 28, 29, 30]. Related disorders like metabolic syndrome and diabetes have also been examined, with similarly little consensus on specific microbial markers of these conditions [31, 32]. Inflammatory bowel disease (IBD) is a chronic disease characterized by mucosal inflammation of the gastrointestinal tract. Animal studies relating gut microbes, immune function, and IBD have pointed to an important role for bacteria in affecting the progression of IBD [33]. Recent studies have focused on classifying IBD patients based on their fecal microbiota and on identifying discriminatory taxa in stool [34, 35]. 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 [30, 34, 35]. 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 [36, 37]. Studies have analyzed disease associations with tumor- and lumen-associated microbiota, and have consistently found enrichment of the *Fusobacterium* genus in CRC patients [36, 38, 39, 40]. 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 [41, 42]. Finally, diseases like rheumatoid arthritis, autism, Parkinson's, HIV, and others have also been examined for microbial associations, though these fields remain relatively unexplored [23, 27, 43, 44].

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 their own 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) [45]. Our gut microbiome is stable over time and can also change rapidly in response to disease, antibiotics, travel, and diet [46]. These perturbations can be fully reversible or can also have long-term effects [46]. Dysbiosis is often discussed as an "imbalance" of gut microbes, though is generally applied to mean any community disruption related to disease [47]. Generally, less diverse communities are thought to be associated with disease, though studies often find no significant association with diversity and disease [30]. Early mouse studies associated the ratio of Bacteroidetes to Firmicutes with altered phenotypes, but few subsequent studies have found similar associations in human patients [24, 48].

Combining existing studies to increase our ability to find consistent disease associations is a promising approach, but existing meta-analyses have had mixed results [30, 48]. 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 [30]. Meta-analyses of obesity studies also tend to find no clear taxonomic associations with obesity [48, 30], even though the microbiome has been causally linked to obesity in mouse models [24, 28]. 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 biologically [49, 50, 51, 52].

2.3 Analytical background and significance

2.3.1 Data generation, standard analysis methods and associated challenges

A common way that researchers study the human microbiome is to do amplicon-based next generation sequencing of 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 [49, 48]. 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 [49]. To interpret results, researchers assign taxonomies to OTUs using a variety of methods, for example by mapping them to annotated reference genomes or using a Bayesian classifier trained on a reference set of annotated bacteria [53, 54]. 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 [48]. The data is also incredibly sparse: only few OTUs tend to be present in many of the samples, and most entries in the data matrix are zeros [51]. 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 results across studies. These issues may be large 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 researchers take similar approaches to gleaning insight from case-control cohorts [49]. 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) [49]. Finally, most studies perform univariate non-parametric tests on the abundance of OTUs to find bacteria significantly associated with the condition of interest [55]. However, because of the high-dimensionality of the data and the often very low sample sizes, many studies yield no significant results [55].

2.3.2 Interpreting taxonomy-based microbiome analyses

Few analytical tools exist to interpret the lists of significant OTUs resulting from 16S analyses into biological hypotheses. When bacteria are found to be associated with the study condition of interest, identifying the patterns which group these bacteria together is 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 reported associations or 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 upper gastrointestinal tract bacteria. However, no tools providing a systematic approach to extract meaning from OTUs currently exist. Additionally, in many cases few or no OTUs are found to be significantly associated with a condition of interest at all.

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 [56, 57]. 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.

Interpreting the biological significance of a list of significant OTUs through enrichment analysis relies upon the existence of trait annotations for the OTUs in that list. There currently does not exist a database which maps a microbe's taxonomy (via its 16S sequence) to its biological traits. In GSEA, genes were grouped into gene sets based on their common pathways, functions, locations in the chromosome, and associations with disease [56]. Similar information exists for bacteria in some microbial databases, but none of these databases or tools have been combined to define groups of microbes based on their general traits. The Integrated Microbial Genomes & Microbiomes (IMG) database contains approximately 10,000 annotated bacterial genomes, but the annotations are not fully complete and do not span all categories of interest [58]. 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 the environmental associations and classify the input dataset [59]. Finally, bioinformatic tools like PiCRUSt have been developed that can infer functional content from 16S data [60], but these annotate microbial communities with genes and KEGG pathways present in the communities and do not necessarily translate these to generalizable biological traits. Existing tools and databases for microbial annotations are often based either on inferences from genetic content or on experimentally validated phenotypes, but rarely combine these two methods to create a comprehensive resource that maps microbes to their varied biological traits.

3 Research design and methods

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

The microbial communities of the lungs, stomach, and throat are connected and likely exchange members, but the effects of GERD and aspiration on these microbial communities and their exchange are unclear [1]. **We hypothesize that there is extensive microbial exchange occurring across all sites of the aerodigestive tract, and that certain clinical conditions like aspiration or GERD 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 this exchange. Next, we will investigate how aspiration and GERD affect aerodigestive microbial communities. Aspirating patients are at a higher risk for respiratory infections, and many patients who present with idiopathic respiratory problems have a high prevalence of GERD [9, 13]. Therefore, we hypothesize that aspiration will increase the microbial exchange between the lungs and the throat, and that reflux will increase the exchange between the lungs and the stomach. Quantitatively describing the amount of microbial exchange happening in the aerodigestive tract and determining clinical modulators of this exchange could inform treatments targeted toward reducing the exchange across specific sites.

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 [17]. 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 a Modified Barium Swallow (MBS) test.

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 microbial exchange across sites in the aerodigestive tract

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 it has a significant, positive Spearman correlation of its abundance in both sites.** In other words, if a microbe is consistently exchanged between sites, we expect that if we see little of it in one site in one patient, then we will also see little of it in the other site. If we see more of it in the first site in a different patient, then we also expect to see more of it in the other site. To calculate this correlation, we consider only patients who have the microbe present in both sites (blue points in Figure 2). **We quantify the extent of exchange, p_s by determining what percentage of patients are exchanging that microbe across their two sites.** In other words, p_s is the number of patients who have the microbe present in both sites divided by the total number of patients. We will calculate this metric for each OTU across all site-combinations, i.e. throat and lung, stomach and lung, and stomach and throat.

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. If exchanged microbes are closely related, this 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

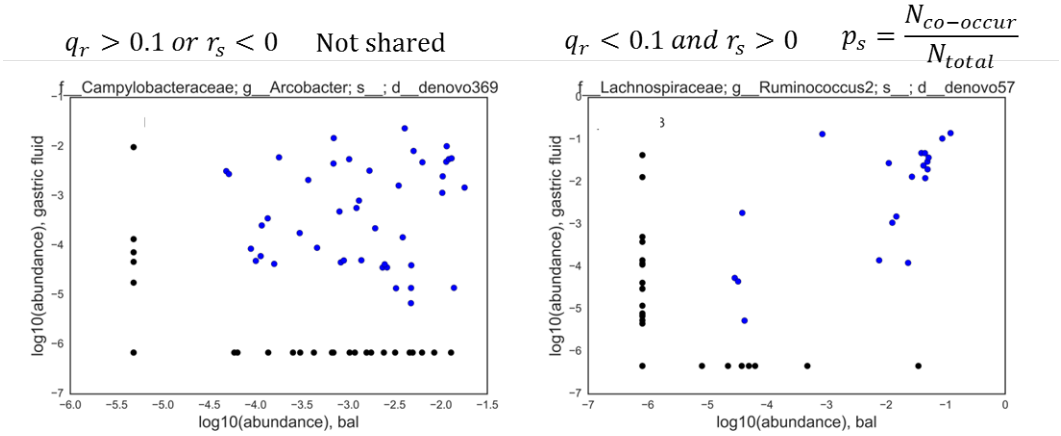

Figure 2: If the abundance of a microbe when it is present in both sites is significantly positively correlated, then we consider it exchanged across those sites (blue points, right panel). p_s is then calculated as the percentage of patients who have the microbe present in both sites (blue points divided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

niches.

3.1.3 Evaluate the effect of aspiration and gastro-esophogeal reflux disease on aerodigestive microbial communities

Once we quantify microbial exchange within the aerodigestive tract, we can begin to ask how conditions like aspiration and GERD modulate the exchange that occurs between them. **Our first hypothesis is that aspirators will have more exchange between their throat and lungs than non-aspirators.** Because aspirators cannot reliably protect their airways from foreign materials in general, they may also have more exchange between their stomachs and 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. Identifying how the throat-lung and stomach-lung connections are affected by aspiration would shed light on the involvement of the microbiota in these two clinical hypotheses. 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 their stomach and lung communities. Because we are interested in GERD that may modulate the stomach-lung connection, we will focus our analyses on full-column reflux, i.e. reflux that reaches the proximal (top) part of the esophagus, since it is more likely to enter the lungs. Our GERD data quantifies the percent of reflux which is full-column, but does not provide a hard cutoff above which reflux is considered "severe". Thus, a secondary goal of this work is to identify if such a threshold exists. We will examine how exchange between and similarity of the lung and gastric communities changes as a

function of the "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.

For each of these clinical factors, **we will compare the amount of microbial exchange occurring between the sites of interest in the two groups of patients**, i.e. non-aspirators vs. aspirators or less severe reflux vs. severe reflux. We will calculate a new p_s within each patient subgroup for each of the previously-defined exchanged microbes. In other words, p'_s will be the percentage of the patients in one subgroup that have the exchanged microbe present in both of their sites. **We will also investigate how community similarities across aerodigestive sites are affected by the conditions.** We will calculate the beta diversity between the two sites for each patient who has sequencing data for both of the sites. If a clinical factor increase the amount of exchange between two sites, we expect to see either more exchange of specific microbes and/or more similar overall communities across the two sites.

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 downstream communities are 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 subsequent adverse clinical outcomes.

3.2 Aim 2: Re-analyze published gut microbiome studies to identify disease-associated microbes

Many researchers have found associations between gut microbial communities and individual diseases, but combining results from different studies about the same disease often yields conflicting results. To date, no one has performed a comprehensive comparison of the results from gut microbiome studies across all disease states with standardized processing and analysis methods. **We hypothesize that re-analyzing existing gut microbiome datasets across all diseases will identify bacteria associated with a general response to disease and others associated with specific disease states.** To address our hypotheses, we will first acquire and re-process a comprehensive collection of case-control gut microbiome datasets. We will perform univariate tests for microbial associations with disease in each individual dataset. We will then combine the results from these univariate analyses across all datasets and, separately, across all datasets of the same disease [61]. Microbes which are significantly enriched or depleted across all datasets will be considered part of a shared microbial response to disease, whereas those which are only associated with one or two individual diseases will be considered putative microbial markers of those diseases. These findings will increase our understanding of the role that microbial communities play in maintaining health or promoting disease, point to potential biomarkers for specific diseases, and motivate future mechanistic studies on the relationship between microbes and their human hosts.

3.2.1 Compile and re-process publicly available case-control gut microbiome studies

To perform a meta-analysis, we will first 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 ones which contain 16S rRNA sequences from 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 [50]. 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 least processed data available - in most cases, these will be raw 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 [53]. 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 a good balance between the taxonomic resolution we can get with the amount of data we have to discard (i.e. unannotated OTUs). Additionally, previous work has found that the maximal predictive power of the microbiota to distinguish between different phenotypes occurs at various taxonomic thresholds [51], so we will also consider higher-level taxonomic classifications in our meta-analyses.

3.2.2 Identify generic microbial responses to disease shared across all studied disease states

Once we re-process existing gut microbiome datasets in a standardized way, **we hypothesize that we will identify broad community shifts related to generic disease.** We will begin by comparing summaries of the microbial communities in each disease state. For example, we will compare the differences in alpha diversity for all cases and controls of the same disease across multiple datasets. We can combine alpha diversities from different studies in two ways. First, we can standardize the alpha diversities within each study (i.e. by subtracting the mean and dividing by the standard deviation), combine samples from studies focused on the same disease, and perform a statistical comparison of the alpha diversity across all samples in the case or control groups. Alternatively, we can do individual statistical comparisons within each study and combine these p-values together with the weighted Z-

test, a method used in meta-analyses for combining p-values [61]. We will begin with alpha diversity because it is the most frequently reported metric in the microbiome literature. We will also similarly consider other community summaries like the Bacteroides/Firmicutes ratio, beta diversity between and within cases and controls, and phyla abundances.

We also hypothesize that certain microbes are part of a shared response to disease. We will use univariate non-parametric statistics to compare the abundance of individual microbes between cases and controls. To compare microbes across disparate studies, we will analyze OTUs that have been collapsed to the genus level. We will combine the univariate results across all studies using the weighted Z-test [61] to determine the overall significance of each microbe across all diseases. Genera which are significant after multiple hypothesis corrections and meta-analysis combination are those which can be considered to contribute to general phenotypes of health (if they are consistently more abundant in controls) or disease (if they are consistently more abundant in cases). We will also look for similar consistent associations with higher-order taxa, in order to identify whole clades of related organisms which are associated with health or disease.

3.2.3 Identify microbes consistently associated with specific diseases

Next, we aim to identify microbes which are consistently associated with *specific* diseases. We will combine univariate results for individual diseases by using the same meta-analysis method described above, but now only combining datasets which analyze the same individual disease state [61]. We expect to find two types of disease-specific microbes from this analysis. The first is bacteria which are not part of the generic response to disease but which are significantly enriched or depleted in an individual disease state. The second is bacteria which *are* enriched or depleted in the generic response to disease, but which have a different directionality in a specific disease. For example, a microbe could be significantly depleted overall across all studies, but significantly enriched when looking just within the obesity studies. The microbes we identify with this analysis may be very interesting candidates for biomarkers or mechanistic follow-up studies.

In this aim, we may 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 [24, 28, 36]). 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 batch effects. First, we can apply simple linear correction methods like subtracting the principal components of variation which correlate closely with technical artifacts like read depth or sample size. Another option is to use open-reference OTUs rather than collapsing to genus and compare the phylogenetic relationships of the significant OTUs across different studies. Finally, we could also approach our meta-analysis from a functional point of view by using tools like PiCRUSt to assign functionality to our observed taxonomies [60].

3.3 Aim 3: Curate a database of 16S sequences annotated with general biological traits

Microbiome studies yield associations between 16S sequences and disease, but many of these sequences do not have functional annotations in existing databases. Tools that connect unannotated sequences to functions do not include general biological traits, making it difficult to interpret results from microbiome studies into biological hypotheses. **In this aim, we will develop a centralized database of 16S sequences annotated with their associated biological traits.** We will begin by searching the literature for existing databases, combining their annotations and supplementing them with literature search and genome-mining as needed. We will begin with traits like sporulation and body site habitat. We will also include disease associations that we identified in Aim 2 and search for other putative phenotype-associations in these datasets. Finally, we will package our database into a tool that researchers can use to interpret their case-control microbiome studies into biological hypotheses and possible mechanistic associations.

3.3.1 Combine existing databases and perform targeted literature searches to annotate microbes based on known biological traits

In order to begin curating our database of microbes and their traits, **we will perform an extensive literature search to identify existing databases and comprehensive review papers with experimentally validated microbial phenotypes**. We will combine these databases and identify where they are missing annotations. We will begin with IMG, a database with approximately 10,000 annotated microbial genomes. About half of these genomes are human-associated bacteria, and half of those have annotations for traits like disease association, sporulation, and body site habitat. We will extract the 16S sequences for the microbes in this database and ensure that all of the organisms we identified in Aim 2 are represented in IMG's annotated microbes. If specific microbes are missing from IMG, we will manually include them and their associated metadata from NCBI queries and targeted literature searches.

In collaboration with Ilana Brito, Assistant Professor of Biomedical Engineering at Cornell University, **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 traits of interest** (Figure 3). For example, in order to annotate spore-forming microbes, we will first perform a literature search to identify known sporulation genes. We will then search for these genes in the whole genomes of representative organisms or in the 16S-based functional profiles inferred by PiCRUSt [60]. Organisms which have these genes will be annotated as spore-formers. As another example, annotating traits like body site habitat will depend mostly on literature search since there are likely no clear genes conferring this trait. We will also extract habitat associations from large datasets like the Human Microbiome Project which sequenced multiple body sites [45]. Throughout our annotation process, we will need to balance the confidence we have in the annotations with the breadth of microbes we are able to annotate. For example, some bacteria which are known spore-formers do not contain any known sporulation genes. We will fill these out to the best of the field's existing knowledge, but our database will certainly still contain a vast number of

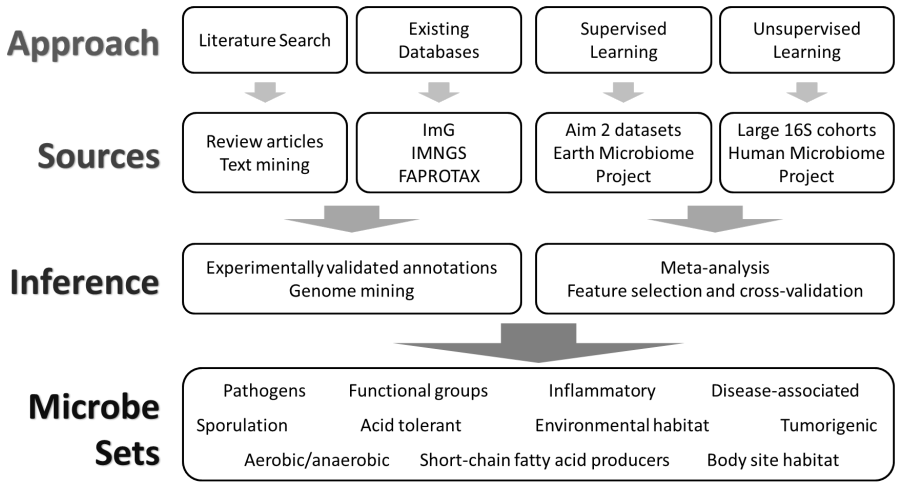

Figure 3: A variety of approaches will be used to define microbe sets, including manual curation from literature and database searches and data-driven methods ("Approach"). Many different kinds of resources will be drawn upon ("Sources") to infer groups of related microbes ("Inference") based on different categories ("Microbe Sets"). Databases described in [58, 62, 63]

false negatives. On the other hand, microbes which have certain sporulation genes may not express them in a biologically meaningful way in the human-associated environment. Therefore, it will be important for our database to include information on how the annotations were derived, i.e. from experimental studies, genetic content inference, or taxonomy-based functional inferences.

3.3.2 Extract disease-associated groups of bacteria from datasets in Aim 2

In parallel with our literature-based annotations, we will use the datasets collected in Aim 2 to identify additional disease-associated groups of microbes. We will annotate microbes based on whether they were enriched or depleted in the shared microbial response to disease or if they were associated with specific diseases. We will also extract additional disease- and phenotype-associated microbes by applying machine learning methods to the datasets in Aim 2 for various classification tasks (Table 2). We will consider only genera which were present in the majority of studies in order to both reduce the dimensionality of the classification task and also ensure that extracted groups of microbes 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 studies and have been shown to perform well in discriminating phenotypes based on 16S data [34, 49, 52]. We will apply these methods to various classification tasks (Table 2) and identify the most important features (RF) or those with the highest support (SVM) for each successful classification. 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 few or no groups of microbes with phenotype associations that we are confident enough in to include in our database. Considering the diversity of

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.

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 manually supervised curation work described in Section 3.3.1 [51, 55]. If this happens, we will investigate higher-order taxa as possible features, as this will reduce the dimensionality, sparsity, and inter-personal variability in the datasets. Another approach is to directly convert each 16S community to functional profiles [60], 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 our microbial trait annotations available to researchers as both a flat database and as a packaged tool to interpret 16S studies. Through our literature searches, we will identify the format of databases that researchers have found most useful and strive to package our annotations in a user friendly, useful format. We will likely begin with one very large text file containing all of the microbes, 16S sequences, and trait annotations that we have gathered. In addition to distributing the annotations 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 Gene Set Enrichment Analysis tool [56]. 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 microbiome studies. It will also be immensely useful to researchers by

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

4.1 Aim 1: Aerodigestive microbiota associated with GERD and aspiration

4.1.1 Aerodigestive communities exchange microbes across all sites

We identified over 150 OTUs exchanged across sites in the aerodigestive tract (Appendix 5.2, Figure 8) using our definition of p_s (Section 3.1.2). As expected, the majority of exchange occurred between the throat and stomach [1]. Interestingly, the stomach and lung communities were almost as similar to each other as the throat and stomach communities were, but had less frequent exchange of specific microbes (Figure 4). These findings support the hypothesis that frequent non-specific microaspiration of gastric contents into the lungs is occurring, in which the exchange of microbes is more stochastic and not necessarily consistently selecting for specific community members across many patients. Finally, we observed a decreasing trend in number of exchanged microbes across throat and stomach, lung and stomach, and throat and lung sites, respectively, for all phyla except Proteobacteria (Appendix 5.2, Figure 8). More Proteobacteria OTUs were exchanged between lungs and stomach than between the throat and stomach. Proteobacteria are known aerobes, and so may be preferentially selected for colonization in the lungs after microaspiration from the stomach. These results support the existing hypothesis that the majority of exchange across aerodigestive sites occurs between the throat and stomach, and also shows that stomach and lung communities may be more related than previously thought.

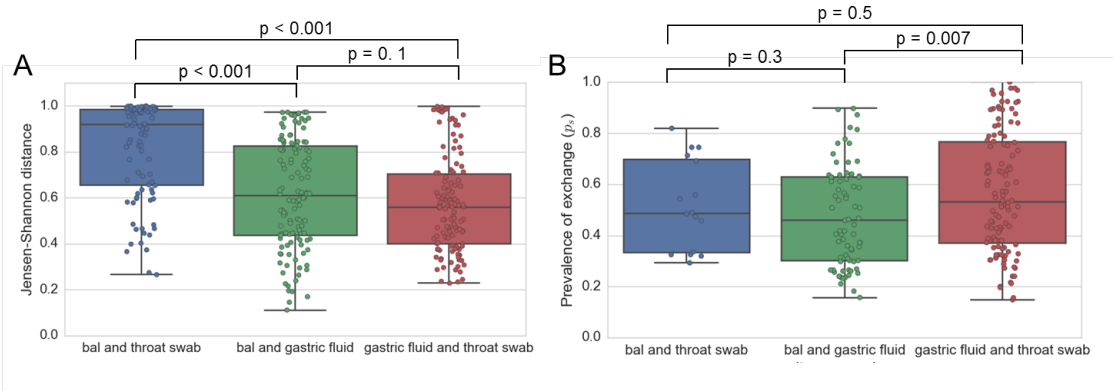

Figure 4: Community similarities and exchange across sites. (A) Jensen-Shannon distance (JSD) between sites within each patient. Identical communities have a JSD = 0; completely different communities have JSD = 1. (B) p_s for each exchanged OTU across all site combinations. P-values calculated with a non-parametric Kruskal-Wallis test.

4.1.2 Aspiration increases throat-lung exchange and gastro-esophogeal reflux affects stomach-lung relationship

We observed a distinct increase in the amount of exchange between the throat and lungs of aspirators relative to non-aspirators (Figure 5B). 40% of aspirators shared the previously-defined throat-lung microbes between their throats and lungs, while only 15% of non-aspirators did. Additionally, the throat and lung communities were significantly more similar in aspirating patients than non-aspirators (Figure 5A). 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, but not significantly so. This indicates that the stomach is not likely a major source of bacterial seeding of the lungs, even in aspirators. One of the current treatments for aspirating patients who have frequent respiratory infections is fundoplication, an invasive surgery that prevents refluxate from exiting the stomach, because it is thought that the bacteria in the refluxate is seeding the lungs and resulting in infection. These findings show that fundoplication surgery may not be the best course of action in aspirators, since the bacterial exchange between stomach and lungs is not significantly different from the exchange in normal patients [13, 64].

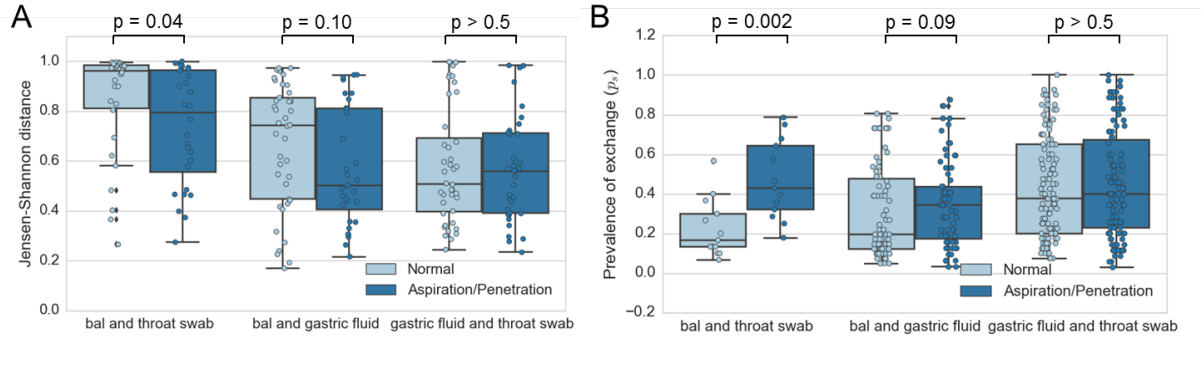

Figure 5: Community similarities and exchange across sites, stratitified by aspiration status. (A) and (B) as in Figure 4.

As expected, stomach and lung communities of patients with severe reflux were more similar to each other than in patients without severe reflux (Figure 6A). Interestingly, patients with severe reflux were not more likely to exchange the previously-defined stomach-lung microbes between their stomachs and lungs (Figure 6B). In this work, 'severe reflux' was defined as reflux in which more than half of events were full-column events. Future work will determine the effect of this 'severity' threshold on the amount of exchange, with the aim of identifying whether a severity threshold exists above which bacterial exchange between the stomach and lungs becomes significantly higher than in non-severe reflux patients.

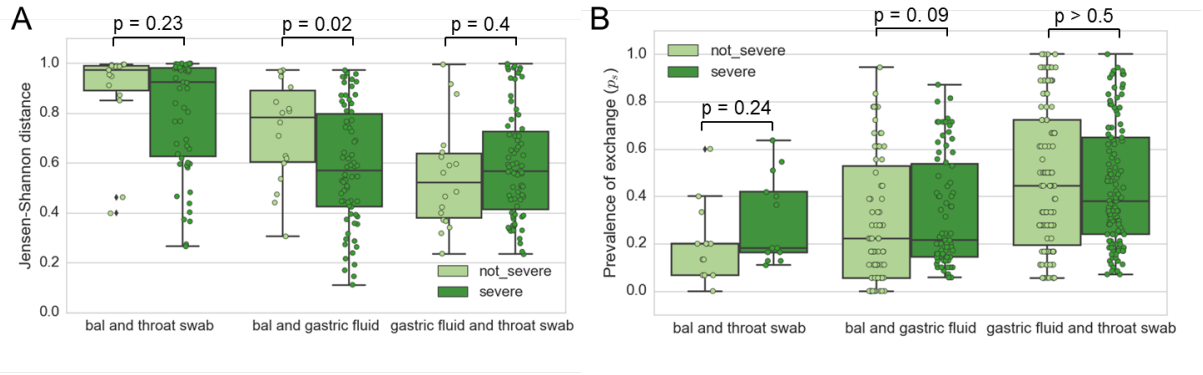

Figure 6: Community similarities and exchange across sites, stratitified by reflux severity. (A) and (B) as in Figure 4.

4.2 Aim 2: Meta-analysis of gut microbiome studies

4.2.1 Collect and process 16S case-control datasets

Our literature review has identified over 50 suitable case-control 16S datasets, 28 of which have been downloaded with associated metadata and processed through our in-house pipeline. Characteristics of these datasets, including sample size, disease states, and median sequencing depth, are shown in Appendix 5.1, Table 3.

4.2.2 General microbial signature of diseases is especially apparent in diarrheal diseases

We first summarized overall community structure using Shannon's alpha diversity index (SDI). As expected, we observed significant batch effects across studies, likely due to the differences in sequencing depth (Appendix 5.2, Figure 9). 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 (Figure 7). Interestingly, disease states which had multiple studies that defined their cases or controls differently also showed significant differences in alpha diversity. For example, one IBD study [34] recruited controls with non-inflammatory conditions of the gastrointestinal tract while others [35, 65, 66] used healthy patients as controls. Additionally, some obesity studies [29, 67, 68] labeled patients as "overweight" in addition to "obese" while others [32, 69] included only "healthy" and "obese" patient category labels. These cases were the only non-diarrheal significant differences in alpha diversity and were less striking than the differences in diarrheal disease, and are hypothesized to be driven by batch effects rather than biology. Further work to investigate this hypothesis will involve applying different standardization techniques and including the individual studies as factors in the statistical comparisons across cohorts.

We next performed within-study univariate comparisons of genus level-abundances in cases vs. controls. Diarrheal diseases had striking shifts in many microbes, while other

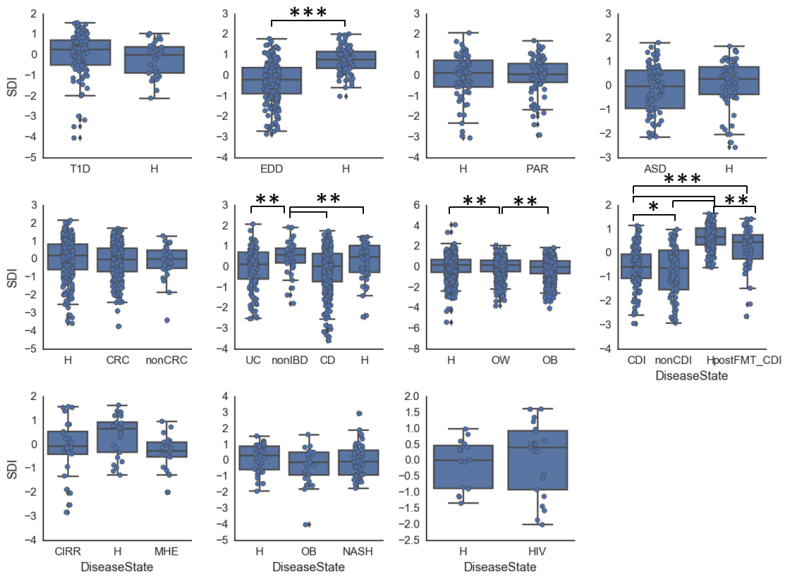

Figure 7: Alpha diversity by study type. Q-values calculated by independent t-tests and corrected for Benjamini-Hochberg false discovery rate. $q \approx 0$ (***), $q<0.0001$ (**), $q<0.01$ (*)

diseases had less obviously apparent microbial indicators of disease (Appendix 5.2, Figure 10). We also observed that some bacteria seemed to have relatively consistent shifts across many different diseases. We confirmed the significance of these associations by combining the p-values of each genus across all studies using the weighted Z-test method [61]. We found many genera in the family *Enterobacteriaceae* to be associated with disease in general, while many genera in the families *Lachnospiraceae*, and *Ruminococcaceae* were associated with healthy controls. These results support the hypothesis that there is a general signature for disease, in other words, that sick people have altered microbiomes. In light of this finding, focusing on microbial shifts that are unique to individual diseases will be crucial to identify disease-specific biomarkers that could be used for diagnostic purposes and to motivate future mechanistic investigations of microbial interactions with disease.

5 Appendix

5.1 Supplementary Tables

Dataset ID	Year	Disease(s)	N control	N case	Median reads per sample	Sequencer	16S Region	Ref.
asd_kb	2013	ASD	20	19	1345	454	V2-V3	[23]
asd_son	2015	ASD	44	59	4777	Miseq	V1-V2	[25]
cdi_schu	2014	CDI	243	93	3557	454	V3-V5	[41]
cdi_vincent	2013	CDI	18	17	5518	454	V1-V3	[42]
cdi_young	2014	CDI	18	27	16516	Miseq	V4	[70]
crc_baxter	2016	CRC	122	120	9476	Miseq	V4	[71]
crc_xiang	2012	CRC	22	21	1152	454	V1-V3	[37]
crc_zackular	2014	CRC	58	30	54269	MiSeq	V4	[40]
crc_zeller	2014	CRC	75	41	120612	MiSeq	V4	[36]
crc_zhao	2012	CRC	54	44	161	454	V3	[38]
crc_zhu	2013	CRC	18	12	1835	454	V3	[39]
edd_singh	2015	EDD	82	222	2573	454	V3-V5	[26]
hiv_dinh	2015	HIV	15	21	3248	454	V3-V5	[44]
ibd_alm	2012	UC, CD	24	66	1303	454	V3-V5	[34]
ibd_eng	2009	UC, CD	32	32	2658	454	V5-V6	[65]
ibd_gevers	2014	CD	16	146	9773	Miseq	V4	[35]
ibd_hut	2012	UC, CD	27	186	995	454	V3-V5	[66]
mhe_zhang	2013	CIRR, MHE	25	46	487	454	V1-V2	[72]
nash_baker	2013	NASH, OB	16	47	9904	454		[22]
nash_chan	2013	NASH	22	32	1743	454	V1-V2	[73]
ob_escobar	2014	OW, OB	10	20	1126	454	V1-V3	[29]
ob_goodrich	2014	OW, OB	451	528	27364	Miseq	V4	[67]
ob_gord	2009	OW, OB	61	219	1569	454	V2	[68]
ob_ross	2015	OB	26	37	1583	454	V1-V3	[32]
ob_zup	2012	OB	167	117	1392	454	V1-V3	[69]
par_schep	2015	PAR	74	74	2351	454	V1-V3	[27]
t1d_alkanani	2015	T1D	23	89	9117	MiSeq	V4	[74]
t1d_mejia	2014	T1D	8	21	4702	454	V3-V5	[75]

Table 3: Datasets currently collected and processed through standardized pipeline. Disease labels: ASD = Austism spectrum disorder, CDI = *Clostridium difficile* infection, CRC = colorectal cancer, EDD = enteric diarrheal disease, UC = Ulcerative colitis, CD = Crohn's disease, CIRR = Liver cirrhosis, MHE = minimal hepatic encephalopathy, NASH = non-alcoholic steatohepatitis, OW = overweight, OB = obese, PAR = Parkinson's disease, T1D = Type I Diabetes.

5.2 Supplementary Figures

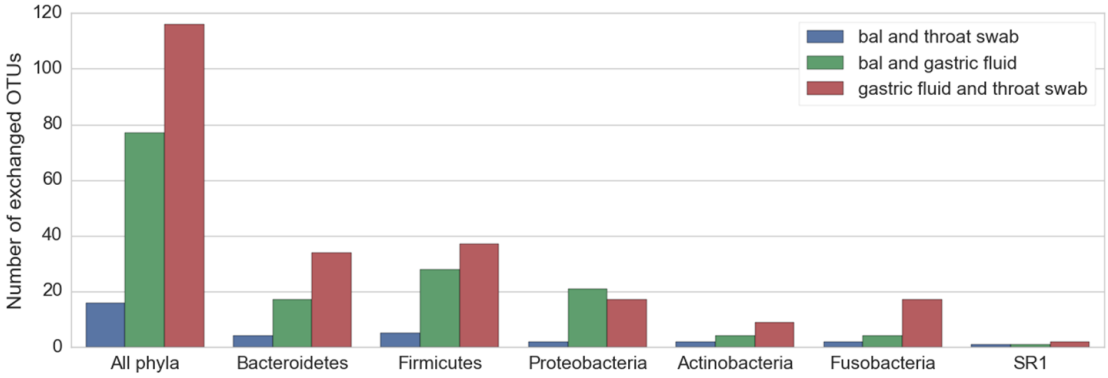

Figure 8: Number of exchanged OTUs across aerodigestive tract sites, separated by phyla.

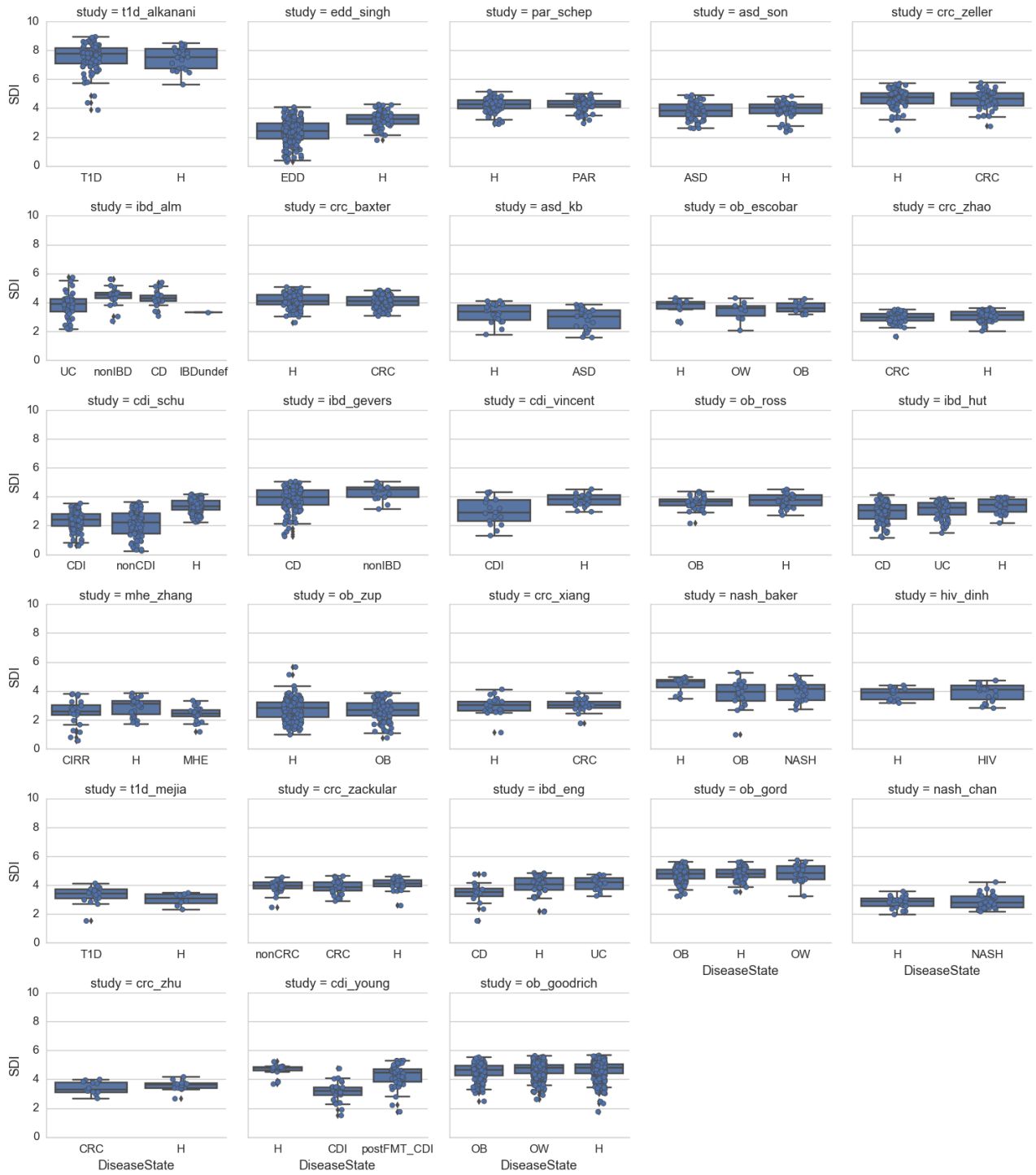

Figure 9: Shannon alpha diversity index (SDI), stratified by individual studies. Note how the range of SDI varies across studies.

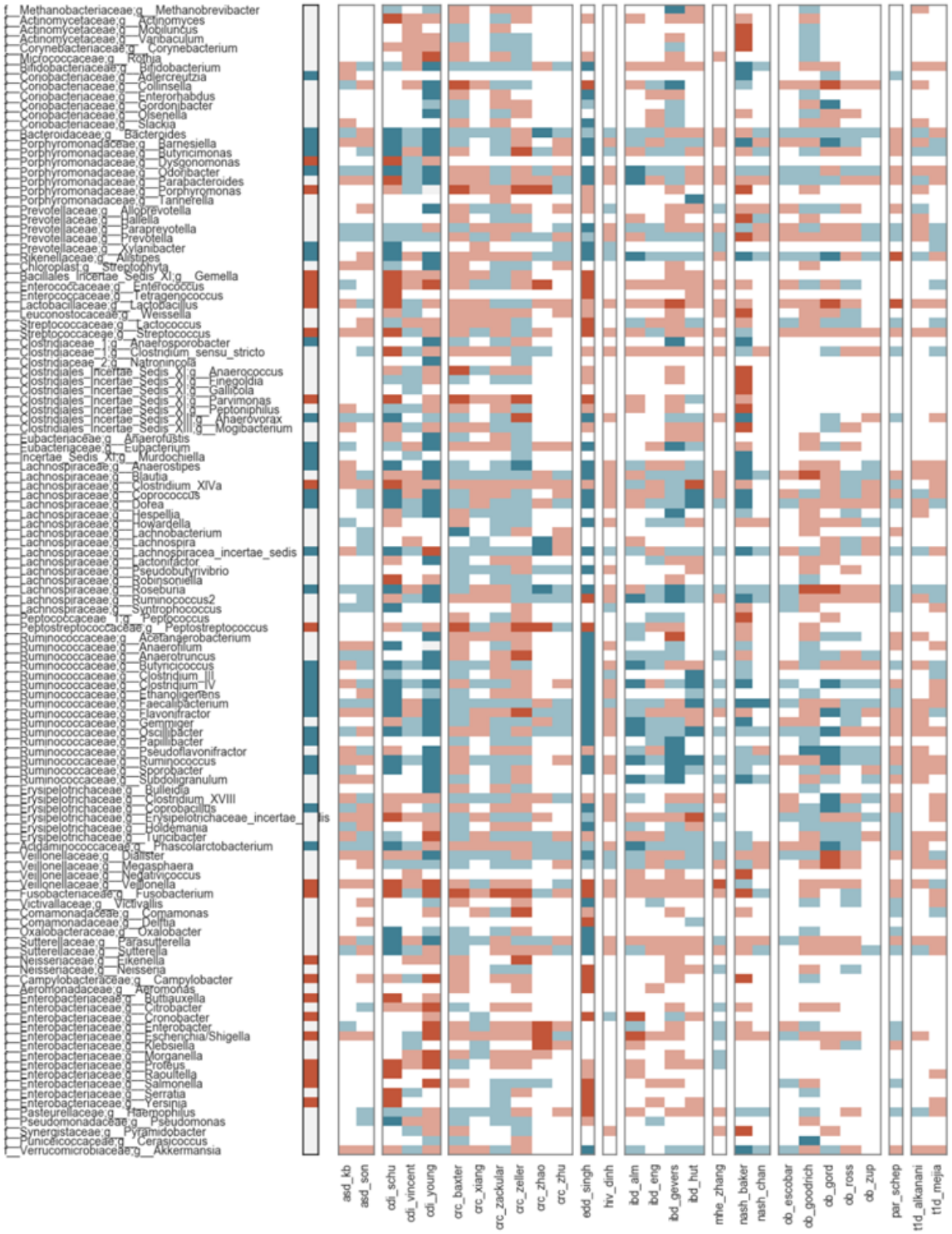

Figure 10: Heatmap of genus-level associations with disease across all compiled studies. Color indicates whether the mean abundance of the organism was higher in the control patients (blue) or disease (red). Shade indicates whether the association was significant after FDR correction (dark color) or not significant (light color), $\alpha = 0.05$. Genus associations which were significant after combining q-values across all studies are shown in the bar to the left of the heatmap. Studies are labeled along the x-axis and described in Appendix 5.1, Table 3; genera are labeled along the y-axis.

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] Fernando M. de Benedictis, Virgilio P. Carnielli, and Diletta de Benedictis. Aspiration lung disease. *Pediatric Clinics of North America*, 56(1):173–190, 2009. doi: 10.1016/j.pcl.2008.10.013. URL http://dx.doi.org/10.1016/j.pcl.2008.10.013.

[14] 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.

[15] 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.

[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] 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.

[18] 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.

[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] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] Hisayoshi Kawahara, Hiroomi Okuyama, Akio Kubota, Takaharu Oue, Yuko Tazuke, Makoto Yagi, and Akira Okada. Can laparoscopic antireflux surgery improve the quality of life in children with neurologic and neuromuscular handicaps? *Journal of Pediatric Surgery*, 39(12):1761–1764, 2004. doi: 10.1016/j.jpedsurg.2004.08.034. URL http://dx.doi.org/10.1016/j.jpedsurg.2004.08.034.

[65] B.P. Willing, J. Dicksved, J. Halfvarson, A.F. Andersson, M. Lucio, Z. Zheng, G. J"arnerot, C. Tysk, J.K. Jansson, and L. Engstrand. A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. *Gastroenterology*, 139(6):1844–1854.e1, dec 2010. doi: 10.1053/j.gastro.2010.08.049. URL http://dx.doi.org/10.1053/j.gastro.2010.08.049.

[66] X.C. Morgan, T.L. Tickle, H. Sokol, D. Gevers, K.L. Devaney, D.V. Ward, J.A. Reyes, S.A. Shah, N. LeLeiko, S.B. Snapper, A. Bousvaros, J. Korzenik, B.E. Sands, R.J. Xavier, and C. Huttenhower. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol*, 13(9):R79, 2012. doi: 10.1186/gb-2012-13-9-r79. URL http://dx.doi.org/10.1186/gb-2012-13-9-r79.

[67] J.K. Goodrich, J.L. Waters, A.C. Poole, J.L. Sutter, O. Koren, R. Blekhman, M. Beaumont, W. Van Treuren, R. Knight, J.T. Bell, T.D. Spector, A.G. Clark, and R.E. Ley. Human genetics shape the gut microbiome. *Cell*, 159(4):789–799, nov 2014. doi: 10.1016/j.cell.2014.09.053. URL http://dx.doi.org/10.1016/j.cell.2014.09.053.

[68] P.J. Turnbaugh, M. Hamady, T. Yatsunenko, B.L. Cantarel, A. Duncan, R.E. Ley, M.L. Sogin, W.J. Jones, B.A. Roe, J.P. Affourtit, M. Egholm, B. Henrissat, A.C. Heath, R. Knight, and J.I. Gordon. A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–484, 2008. doi: 10.1038/nature07540. URL http://dx.doi.org/10.1038/nature07540.

[69] M.L. Zupancic, B.L. Cantarel, Z. Liu, E.F. Drabek, K.A. Ryan, S. Cirimotich, C. Jones, R. Knight, W.A. Walters, D. Knights, E.F. Mongodin, R.B. Horenstein, B.D. Mitchell, N. Steinle, S. Snitker, A.R. Shuldiner, and C.M. Fraser. Analysis of the gut microbiota in the old order Amish and its relation to the metabolic syndrome. *PloS one*, 7(8):e43052, 2012. doi: 10.1371/journal.pone.0043052. URL http://dx.doi.org/10.1371/journal.pone.0043052.

[70] I. Youngster, J. Sauk, C. Pindar, R.G. Wilson, J.L. Kaplan, M.B. Smith, E.J. Alm, D. Gevers, G.H. Russell, and E.L. Hohmann. Fecal microbiota transplant for relapsing

clostridium difficile infection using a frozen inoculum from unrelated donors: A randomized, open-label, controlled pilot study. *Clinical Infectious Diseases*, 58(11):1515–1522, 2014. doi: 10.1093/cid/ciu135. URL http://dx.doi.org/10.1093/cid/ciu135.

[71] N.T. Baxter, M.T. Ruffin, M.A. Rogers, and P.D. Schloss. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Medicine*, 8(1), 2016. doi: 10.1186/s13073-016-0290-3. URL http://dx.doi.org/10.1186/s13073-016-0290-3.

[72] Z. Zhang, H. Zhai, J. Geng, R. Yu, H. Ren, H. Fan, and P. Shi. Large-scale survey of gut microbiota associated with MHE via 16s rRNA-based pyrosequencing. *Am J Gastroenterol*, 108(10):1601–1611, jul 2013. doi: 10.1038/ajg.2013.221. URL http://dx.doi.org/10.1038/ajg.2013.221.

[73] V.W. Wong, C. Tse, T.T. Lam, G.L. Wong, A.M. Chim, W.C. Chu, D.K. Yeung, P.T. Law, H. Kwan, J. Yu, J.J. Sung, and H.L. Chan. Molecular characterization of the fecal microbiota in patients with nonalcoholic steatohepatitis – a longitudinal study. *PLoS ONE*, 8(4):e62885, apr 2013. doi: 10.1371/journal.pone.0062885. URL http://dx.doi.org/10.1371/journal.pone.0062885.

[74] A.K. Alkanani, N. Hara, P.A. Gottlieb, D. Ir, C.E. Robertson, B.D. Wagner, D.N. Frank, and D. Zipris. Alterations in intestinal microbiota correlate with susceptibility to type 1 diabetes. *Diabetes*, 64(10):3510–3520, 2015. doi: 10.2337/db14-1847. URL http://dx.doi.org/10.2337/db14-1847.

[75] M.E. Mejía-León, J.F. Petrosino, N.J. Ajami, M.G. Domínguez-Bello, and A.M.C. de la Barca. Fecal microbiota imbalance in mexican children with type 1 diabetes. *Sci. Rep.*, 4, 2014. doi: 10.1038/srep03814. URL http://dx.doi.org/10.1038/srep03814.