Understanding the Clinical Microbiome
Biological Engineering Thesis Proposal

Claire Duvallet

October 11, 2016

Contents

Abstract

Analyzing the microbiome is hard. Getting clinical insights from analyses is even harder. I'm gonna do some analyses to give us insight into an under-studied clinical microbial system, do some meta-analyses to get direly-needed biological consensus on gut microbiome and disease, and propose a new tool for analyzing 16S datasets.

1 Overall objectives and specific aims

1.1 Overall objectives

In spite of the recent increase in research about the human microbiome, there is not a clear consensus on the relationship between human microbial communities and disease. Current 16S microbiome analyses typically study one patient cohort in one disease state, searching for individual disease-associated microbes. However, different published studies for the same disease often contain contradictory or inconsistent results. Existing meta-analyses rarely expand to more than one or two diseases, and thus do not distinguish between microbes which are associated with specific diseases from those which are associated with disease in general. Finally, there are no established tools to extract general biological insights from groups of disease-associated microbes.

This thesis will increase our understanding of the clinical microbiome by moving analyses from focusing on single microbes in individual diseases toward consolidation of groups of related microbes across many different kinds of diseases. In this thesis, I will first apply standard methods to characterize the under-studied microbiota of the aerodigestive tract of one patient cohort. Then, I will perform a comprehensive meta-analysis of gut microbiome studies across many disease states with multiple patient cohorts. Finally, I will develop a tool to enable generalizable interpretation of results from existing and future microbiome studies. 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.

1.2 Specific Aims

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

1. Determine how lung, gastric, and throat microbial communities are related.
2. Identify clinical modulators of lung, gastric, and throat microbial communities.

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

1. Compile and process publicly available case-control gut microbiome studies using a standardized method.
2. Identify microbes that are consistently associated with specific diseases and with disease in general.
3. Compare results betweens studies to identify similarities in microbial characteristics of physiologically-related diseases.

Aim 3 Enable generalizable interpretations of microbiome analyses by assigning bacteria to groups with similar functions 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 microbiome is a hot hot hot field and we know some stuff but we don't know a lot of stuff too.

My aims cover multiple body systems: aerodigestive and gut. My aims also span from one study to many, so need to talk about those approaches.

2.1 Biological background and significance

2.1.1 Aerodigestive tract

The aerodigestive tract consists of the upper gastrointestinal and respiratory tracts (REF). From an engineering perspective, the stomach, throat, and lungs can be thought of as different compartments connected by the esophagus and windpipe. (FIGURE) The mass transport between these 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 in some diseases is dysfunctional. Finally, complex homeostatic mechanisms clear the lungs of foreign bodies and create a very selective environment for microbes in the lungs (REFS REFS REFS). Thus, while the throat, stomach, and lung "compartments" are physically connected, the amount of material, including bacteria, flowing between them is not readily apparent.

2.1.2 Gastro-esophogeal reflux disease and aspiration

Gastro-esophogeal reflux disease (GERD) is set of syndromes in which the reflux of stomach contents leads to troublesome symptoms or complications [1],[2](FIGURE - Figure 2 of Vakil et al). The most common symptoms of GERD are regurgitation and heartburn, but the disease may also present asymptomatically[1],[2]. GERD can be diagnosed by demonstrating reflux of gastric contents (via pH and impedance monitoring), injury to the esophogus (via endoscopy), or based on symptoms alone[1].

Aspiration is the inhalation of foreign material such as food or gastric contents into the lungs [3]. 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 [3]. Aspiration resulting from swallowing dysfunction can be diagnosed with a Modified Barium Swallow test (MBS) [4]. 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 [5],[4]. However, the MBS test cannot be used to determine aspiration of gastric contents or episodes of micro-aspiration, which may also have clinical relevance [3],[6]. No validated clinical biomarkers exist to diagnose and define gastric and micro-aspiration events[6], but they are thought to play a role in causing or exacerbating many respiratory diseases [7],[8],[9]. 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],[6]. Further complicating the issue is that many healthy patients have a baseline level of micro-aspiration (REF).

2.1.3 GERD, aspiration, and respiratory disease

GERD is 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 (REF) and reduced lung function in patients with cystic fibrosis [8]. 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 [8],[7]. 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 [8],[9].

2.1.4 Microbiome of the aerodigestive tract

The microbiota of the human lungs and stomachs are among the least well-studied human-associated microbial communities. In fact, the lungs are classically thought to be sterile and free of bacteria in healthy people [10],[11]. 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 (HMP REF). However, subsequent studies have found suggestions that aerodigestive tract sites are microbially related, and that these microbiomes are altered in certain disease states (REFS).

Few studies have examined the relationship between sites of the upper aerodigestive tract [11]. Bassis et al. found that oral microbial communities were more similar to both stomach and lung communities than nasal communities. However, these authors did not examine the inter-relationships between all sites - their driving hypothesis was that the mouth is the source of the downstream communities. In this work, we will take a broader view: which of the upper aerodigestive body sites exchange microbes, are these the same microbes across multiple patients, and are there patterns for these shared microbes?

2.1.5 Lower gastrointestinal tract: physiology and microbiome

The human gastrointestinal tract is integral to health and disease. Its primary function is to digest food and absorb nutrients, and it also plays important roles in many other aspects of physiology. For example, it is the largest secretor of serotonin and something else (blood flow??) (REF?!). The lower gastrointestinal tract is populated by a complex microbial community, which serves important functions in digesting and salvaging additional nutrients, interacting with the immune system, and protecting the gut from invasion by pathogenic organisms (REF).

We know that the microbial communities in our guts are unique, modifiable, and integral to our overall health. Studies have consistently shown that gut microbial communities

are more similar within a person over time than across people (REF). Environmental disruptions like travel, change in diet, or antibiotics can have both reversible and irreversible changes on the composition of the gut microbiome (REF). Fecal microbiota transplants have demonstrated the causal ability of the microbiome to impact health in both animal models and human patients, and more targeted therapies like narrow-spectrum antibiotics or strain-specific probiotics are an active area of research and development (REF). Experiments with germ-free model organisms have also clearly demonstrated the microbiome's causal role in sustaining health and causing disease (REF).

2.1.6 Microbiome/disease associations of the lower gastrointestinal tract

Because of the microbiota's importance in the healthy function of the gastrointestinal tract, associations between microbial communities have been investigated for a wide variety of diseases (REFS).

Because of its important role in human physiology, many disorders are related to a dysfunctional intestinal tract. Metabolic syndromes like obesity and diabetes

2.2 Analytical background and significance

2.2.1 Data generation and challenges

The advent of high throughput next generation sequencing has enabled culture-independent analyses of complex microbial communities. Briefly, DNA is extracted from biological samples of interest and a portion of the universal bacterial 16S region is amplified. Data is sequenced using a variety of available technologies such as 454 Pyrosequencing or, more recently, Illumina HiSeq or MiSeq (REFS). The resulting reads are quality-controlled and processed into Operational Taxonomic Units (OTUs), clusters of similar sequences which serve as proxies for bacterial species. OTUs can be assigned bacterial taxonomies 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 (GG and RDP REFS).

The resulting datasets are often very high-dimensional, with hundreds of OTUs present in a given cohort which may only have tens of samples. 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. Furthermore, technical artifacts such as DNA extraction and PCR amplification biases are known issues that lead to strong batch effects across different studies. Different data processing, quality filtering, and name-assignment methods also lead to strong batch effects. 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 availability of many studies.

2.2.2 Current analytical approaches for 16S analyses

While there exist no established standards for processing or analyzing 16S data, most studies take similar approaches to gleaning insight from case-control cohorts. Alpha diversity, the diversity of species within each sample, is usually compared for case and control groups.

Many studies have found slight indications that alpha diversity is higher in healthy patients (REF). Beta diversity, the diversity between samples, is also frequently compared to understand whether samples within the control or case group are more similar to each other than they are to the other group. Finally, most studies perform univariate non-parametric tests on the abundance of OTUs to find bacteria significantly associated with disease or health. These analyses suffer from high-dimenstionality coupled with low power (few samples) - after multiple hypothesis testing corrections, many studies yield no significant results (REF).

Existing meta-analyses of microbiome datasets often focus on healthy patients or on just one or two diseases (REFS). These meta-analyses have had mixed results. In some cases like IBD, strong and consistent signals can be found across studies and in other cases. A meta-analysis of obesity studies found no clear taxonomic signal, and a difference in alpha diversity that was significant but likely too small to be biologically relevant[12]. Another study found no clear results in obesity datasets but did identify a strong signal associated with IBD across multiple datasets [13]. 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 meanginful signal [12],[13],[14],[15].

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 differentiate subtleties contributing to health and disease (like the inflamed gut vs. the healthy gut) [14].

2.2.3 Databases and tools for annotations

Few analytical tools exist to motivate biological hypotheses from the results of high-throughput case-control microbiome studies. Although studies often yield some significant OTU-disease associations, assigning biological interpretations to such results remains challenging. For example, the importance of short-chain fatty acid (SCFA) producers is well-accepted in IBD, and may play in a role in many other diseases as well (REFS). However, directly identifying enrichment of functionally-related bacteria (such as SCFA-producing bacteria) is not currently possible. Instead, researchers face the task of interpreting lists of individual OTUs into common biological themes or functions on their own (GSEA REF).

Enrichment analysis is a powerful way to 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 ([16], [17]. In Gene Set Enrichment Analysis (GSEA), genes are first ranked by differential expression between two conditions. Then, *a priori*-defined groups of related genes are analzed 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.

GSEA relies on the existence of curated gene sets, which has no existing analog (*i.e.*, microbe sets) in microbial databases. In GSEA, genes were grouped into gene sets based

on their common pathways, functions, locations in the chromosome, and associations with disease. 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 (IMG REF). 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 (REF SOURCETRACKER). Finally, bioinformatic tools have been developed that can functional content (PICRUST) or metabolic pathways (HUMANN) from 16S data. Again, these tools are dataset-specific and have not been generalized to define biologically related groups of organisms.

3 Research design and methods

Six to eight pages

3.1 Aim 1: Aerodigestive microbiota associated with GERD and aspiration

As discussed previously, patients with aerodigestive disorders like aspiration and GERD are at a higher risk for respiratory infections. We hypothesize that microbial communities in the aerodigestive tract share and exchange certain members which then contribute to infections, and that clinical factors like reflux or aspiration disease change the amount of bacterial exchange between aerodigestive sites.

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

Table 1: Aerodigestive site samples

Our patient cohort consists of 261 patients recruited by Rachel Rosen (M.D., GI/Nutrition) at Boston Children's Hospital for multiple studies over the course of the past 6 years. Multiple samples were taken from patients: throat swabs, gastric fluid, and bronchoalveolar lavages (BAL) (Table 1). 125 patients were monitored for full-column GERD and 112 patients were tested for aspiration. Overall, this cohort represents the largest existing human aerodigestive microbiome dataset.

3.1.1 Exchange of microbes between lung, gastric, and throat communities

To understand the microbial exchange between sites in the aerodigestive tract, we will first define a metric to quantify how "shared" microbes are between two sites. There are multiple ways to define this metric: as the percentage of patients who have the microbe present in both sites, as the correlation between the abundance of the microbe in one site with its abundance in the other site, or some combination of these two approaches. Because many microbes in the stomach and lungs are seeded by the oral community, simple co-occurence of bacteria is not sufficient to establish meaningful microbial exchange (REF). Using the correlation of abundances in two sites is also not adequate, since many people may not

have the OTU present in either or both sites due to the inherent variability of microbial communities between people.

Thus, we will use both co-occurence and correlation to quantify microbial exchange across sites, which we call p_s. We will first identify which microbes are shared using the abundance correlation, and then quantify each microbe's degree of sharedness by its co-occurence rate in patients. For each microbe, we will calculate the non-parametric Spearman rank correlation of the log10(relative abundances) in the two sites, using only abundances from patients with the microbe present in both sites. If this correlation is greater than 0.5, the microbe is considered to be shared. Our metric p_s is then defined as the percentage of patients who have the microbe present in both sites (Fig. 1).

3.1.2 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 how much exchange occurs between sites. We will calculate p_s for all a-priori defined exchanged microbes (Section 3.1.1), stratified by the clinical factor. Furthermore, we will investigate whether the total abundance of exchanged microbes in the different sites and beta diversity community similarities differ between our case and control patient groups.

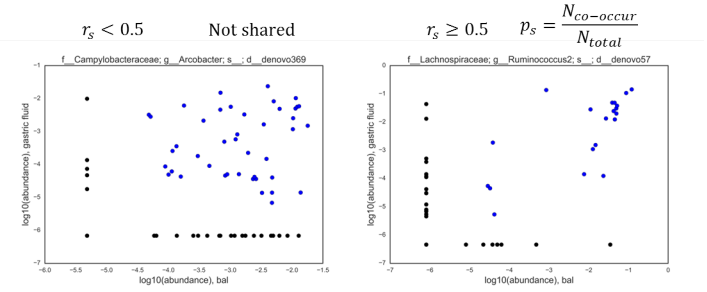

Figure 1: Determining p_s

Our first hypothesis is that aspirators will have a stronger connection between their throat and lungs. Additionally, because many of these patients have GERD, we also expect a slight increase in the connection between the stomach and lungs. Our 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 sharing between the stomach and lung communities. Because are interested in GERD that may modulate the stomach-lung connection, we define "severe GERD" as reflux in which more than 50% of events are full-column reflux events. With this definition, we have 99 patients with severe GERD and 26 patients without. In addition to the binary severe/not severe comparison, we may regress each of the continuous measures of GERD severity onto the abundance of microbes in each patient's lung and stomach communities. We will also investigate how PPIs modulate the connection between aerodigestive sites, since PPIs are often prescribed for respiratory disease but have unclear clinical impact. The dataset has 114 patients on PPIs and 85 patients not on PPIs.

3.1.3 Additional considerations

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 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.

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. Follow up studies focus on patients who develop respiratory infections, or patients who frequently have GERD- or aspiration-associated respiratory infections.

Finally, an important caveat to note when discussing the human lung and gastric microbiomes in general is that our basic understanding of these communities in healthy humans is still quite rudimentary. For example, the extent to which lung microbial communities are composed of surviving and replicating microbes versus transient members is unknown. However, because we do see similarities in microbiota across patients, we can conclude that at least some selection is happening in these sites, and even if these communities are not thriving they may still be physiologically relevant.

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. diabetes and obesity).

3.2.1 Compile and process gut microbiome datasets

To perform a comprehensive 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. See TABLE for exclusion and inclusion criteria for studies to be considered.

We will process these datasets using a standardized in-house pipeline developed by Thomas Gurry and to which I have contributed. 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.

Because studies which sequence different 16S regions will have different sequences corresponding to the same bacteria, we can not used sequence-based open-reference approaches to compare OTUs across studies. After assigning Latin names based on OTUs within each study, we will collapse OTUs to the genus level and compare these across studies.

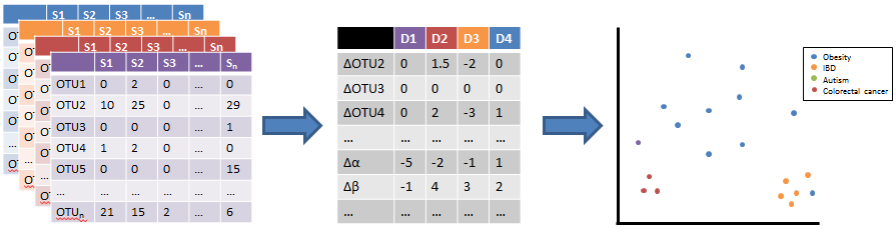

Figure 2: Defining microbial signatures

3.2.2 Identify microbes consistently associated with diseases

Once we have processed all datasets in a standardized way, our first goal is to identify consistent markers of health and disease. We will analyze each dataset with microbiome 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).

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 disease and the direction of change of these microbes. 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". (Fig. 2)

If a disease has a strong impact on or association with the gut microbiome, then we would expect its signatures from multiple studies 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 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 experiments to follow up on mechanism or causal relationships.

3.2.4 Additional considerations

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 (REFS)). Developing robust methods to overcome technical batch effects in 16S studies is not within the scope of this work. However, if standard meta-analysis methods fail, we can try some naive correction methods like a linear correction for the principal components which correlate closely with technical artifacts like read depth (FIGURE?). We could also consider using a phylogenetics-based approach rather than taxonomically-assigned closed-reference OTUs. We could identify associations with disease using open-referenced OTUs and then compare their phylogenetic relationships across all studies. Finally, we could also use tools like PiCRUST to assign functionality to our observed taxonomies, and approach our meta-analysis from a functional point of view (REF).

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 *a priori*, we will enable enrichment analyses similar to GSEA (**??**, REF) for microbiome data. Enrichment analyses will allow for better biological understanding of individual studies' results as well as more consistent comparisons of results across individual studies in the literature.

3.3.1 Define microbe sets based on known biological relationships

No existing databases annotate microbes in a way that can be mapped to 16S sequencing studies. Our first task is therefore to curate and define *microbe sets* for use in enrichment analyses of microbiome datasets. We will begin with ImG, an existing database with approximately 10,000 annotated microbial genomes. About half of these genomes are human-associated, and half of those have annotations for categories like disease association, sporulation, body site habitat, and gram staining. We will extract the 16S sequences for all the annotated human-associated microbes in this dataset to determine whether they adequately span the phylogenetic diversity we expect to find 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 ImG data and to add our own fields of interest. We will first perform a comprehensive literature review

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 2: Possible approaches to define microbe sets of interest

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
Immune system	Rheumatoid arthritis, allergy, Crohn's disease, Graft-versus-host disease vs. controls
Liver	NASH, MHE (all hepatic diseases) vs. controls
Metabolic syndrome	T1D, T2D, obesity, metabolic syndrome vs. controls

Table 3: Classification tasks to identify groups of phenotype-associated microbes

to identify all existing annotated microbial databases. We may also pursue unsupervised text mining of literature results to infer trait-associations for well-studied bacteria (REF). As we populate the microbes in our microbe sets with annotations, we will simultaneously collect their 16S sequences to put them all on a tree. We will apply phylogenetic and phylogenomic inference models to infer the missing traits of leaf organisms when possible.

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

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 distinguished diseases or broader phenotypes in (Section 3.2.3). We will also pursue machine-learning driven approaches to identify novel disease- or phenotype-associated *microbe sets*. We will define *microbe sets* based on the most discriminating features of various comparisons, as described in Table 3.

3.3.3 Develop collaborative tool for interpreting microbiome studies

We will make the microbe set annotations available to researchers for further annotation and development. Through our literature searches, we will pay attention to the databases researchers have found most useful and strive to package our 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.

We will also package our *microbe set* annotations into a tool that researchers can use to interpret the results of their 16S studies. Our software will take as input an OTU table with taxonomically assigned microbes and disease-state labels for samples. It will perform enrichment analysis on the OTU table and return the results to researchers, similar to the Broad's GSEA tool (REF).

All of this work will be done using public open-source tools like GitHub to encourage collaboration and dissemination of our findings.

3.3.4 Additional considerations

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 be just the beginning of systematic grouping of phenotypically-associated microbes, and so we will ensure that the format of the database we develop is easily accessible and modifiable by other researchers.

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?"

4 Preliminary studies

Three to four pages

4.1 Aim 1

4.1.1 Microbiome community sharedness

Look what I can do ma.

4.1.2 Modulators of sharedness

And here!

4.2 Aim 2

4.2.1 Collect datasets

Check out what I got!

4.2.2 Find consistent microbes

PCA, alpha diversity, comparing microbial signatures, consistency of significant OTUs.

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.

References

[1] 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`.

[2] 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`.

[3] 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`.

[4] 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`.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[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] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.