

Analyzing the Clinical Microbiome
Biological Engineering Thesis Proposal

Claire Duvallet

October 11, 2016

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.

Contents

Overall objectives and specific aims

Overall objectives

My overall objective is to expand 16S microbiome research along the diseases, batches, and groups of microbes axes.

Specific Aims

Aim 1 Establish a baseline for 16S analyses by deeply characterizing one patient cohort in one disease state.

- (a) Apply current statistical and biological methods to identify microbial community characteristics associated with gastro-esophageal reflux and aspiration.
- (b) Characterize the under-studied lung and gastric microbiomes.
 - (a) Determine how lung, gastric, and throat microbial communities are related and shared.
 - (b) Identify clinical modulators of these microbial communities, such as diseases state or medication use.

Aim 2 Expand 16S analyses to multiple diseases and patient cohorts to elucidate microbial characteristics consistently associated with health and disease.

- (a) Collect and process all existing case-control gut microbiome 16S datasets in the literature.
- (b) Identify general relationships in gut microbial communities within and across different categories of diseases
- (c) Discover consistent microbes associated with health and/or individual diseases
- (d) Determine best practices for analyzing case-control clinical 16S microbiome studies

Aim 3 Define *microbe sets* to enable generalizable interpretations of case-control microbiome studies.

- (a) Use *a priori* knowledge about biological relationships between microbes to define biologically meaningful *microbe sets*.
- (b) Use machine-learning techniques to extract disease-associated groups of microbes from datasets collected in Aim 2.

- (c) Use the defined *microbe sets* to perform enrichment analyses on a broad spectrum of case-control microbiome studies.

Proposal

Background and significance

Three to five pages

The microbiome

The microbiome is a hot hot hot field and we know some stuff but we don't know a lot of stuff too.

Microbiome of the upper gastrointestinal and respiratory tracts

We don't really know anything bc they weren't included in the HMP and clinicians are bad at big data womp womp.

Aspiration, gastro-esophageal reflux disease, and pulmonary infections

It's all super complicated, yo. And also super connected wat who knew.

Gut microbiome in disease

Lots of interest, not that many conclusions.

Existing meta-analyses?

Haven't found much and haven't been great.

Current practice in case-control microbiome studies

Eh there isn't really a current practice. Use QIIME and see what it tells you? Common methods: Alpha diversity, JSD, t-tests, etc Data processing: different methods (OTU calling, Latin name mapping) lead to drastically different results Data analysis: lots of multiple corrections to do

Enrichment analysis

GSEA is commonly used in RNAseq data! Gene databases exist and have been curated into gene sets. No curated microbe sets. Some existing similar tools: SourceTracker, ImG, ...?

Research design and methods

Six to eight pages

Aim 1

The dataset

From clinical collaborator, really cool dataset, etc

Defining shared microbes across sites

Come up with a metric based on correlation of abundances

Finding clinical modulators of these relationships

Aspiration, reflux severity, PPI usage.

Aim 2

The datasets

Identified through literature search. Processed using an in-house pipeline that I've also contributed to.

Finding general patterns of disease-microbe relationships

Process and analyze each dataset individually. Compare results.

Look at: alpha diversity, beta diversity for H-Dis, significant genera.

Also try to see how *disease types* are related - either by comparing their microbial signatures or something...

Correcting for batch effects

Do stuff here

Aim 3

Define microbe sets - a priori from biology

Using an undergrad and lots of lit searching.

Define microbe sets - unsupervised, from results from Aim 2

And #machinelearning stuff bam wow wow.

Perform enrichment analysis on datasets in Aim 2

See if results jive with what we found from individual genus-based analyses, if we uncover new "signatures of general types of diseases", etc.

Preliminary studies

Three to four pages

Aim 1

Microbiome community sharedness

Look what I can do ma.

Modulators of sharedness

And here!

Aim 2

Datasets

Check out what I got!

Preliminary analyses

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

Conclusion