Introduction[1]

Ongoing research has linked the human microbiome to many diseases, with many related to gastrointestinal disorders such as irritable bowel syndrome (IBS), Crohn's disease, and nutrient malabsorption. Upper gastrointestinal disorders and abdominal pain affect between 12 to 30% of the worldwide population. This research aims to further the ongoing effort to characterize the microbiomes impact on human disease.

Many studies on humans have investigated large intestine microbial dynamics, however small intestine microbial dynamics are less understood and understudied. The small intestine is involved in immune system functioning and has one hundred times the amount of surface area compared to the large intestine. The small intestine comparatively also has thinner and more permeable mucus membrane, suggesting more interaction with human physiology.

One affliction of the small intestine is the presence of small intestinal bacterial overgrowth (SIBO). SIBO has been implicated with IBS and other gastrointestinal disorders. SIBO is still being investigated and present treatments have their drawbacks. The authors of the study aim to improve the understanding of the community structure of SIBO and the microbial load characteristics they possess.

The study selected 250 individuals from the REIMAGINE study to assess absolute microbial loads of a specific portion of the small intestine, the upper portion, the duodenum. In addition, they also selected 21 individuals from the cohort to investigate the oral microbiome as well, to understand the relation between the small intestine microbiome and the oral microbiome. The authors used their specific digital PCR anchored 16S rRNA gene amplicon sequencing method to tabulate the absolute taxon abundances from each sample. All patients sampled were undergoing esophagogastroduodenoscopy (EGD) without colonoscopy preparation, due to needing medical attention, so no healthy controls were compared. Total absolute microbial loads data were generated via the aforementioned method. This data was then utilized to complete the following data analysis.

Code Structure & Data Analysis

The first figure I aimed to reproduce was Figure 3B, a clustered co-correlation matrix of the top 16 genera ranked by the difference between their maximum abundance and mean abundance. Structurally I divided the code into three main parts. Part one was to prepare the data, part two being the spearman correlation analysis, and part three being the final plot. First all package dependencies to perform the analysis were imported. To begin the analysis, I imported the data from the paper provided data download "pickle_files". The pickle files process the data from the dPCR data into python usable pickle format. From pickle_files I loaded the data via pickle for both the absolute data and the associated metadata for each amplicon sequence variant (ASV). This loaded data contains a list of 6 pandas dataframes. The only relevant data to the analysis in this figure is the 5th data frame, as such I created a variable for the absolute data and the ASV names.

I next began to curate and prepare the data. My goal was to merge the naming data to the PCR abundance data, as they are stored separately with separate orders. The figure also does not make mention of the specific ASV numbers that correspond to

the names of the unique microbes, which is essential to plot the data. To begin I stored the names of each genus name shown in the figure. I then made dictionaries to find the associated “ASV” from each name and stored the result in a list (from the naming metadata). I then took the list and iterated through each species to obtain the data for each ASV involved in the plot. Now having the data, I was able to perform spearman correlation, as was specified in the paper. To do this I used a double loop. I iterated through each ASV, for each ASV and performed spearman correlation each time using scipy.stats spearman correlation function. Furthermore, I processed this into a usable/plottable pandas data frame to then plot.

Finally, I plotted the correlation matrix via seaborn’s heatmap. There were many details on the Figure 3B, the first being the labels. I proceeded to make a list of labels and specify which on the original figure’s labels were bolded. I then iterated through each tick label to achieve this. Of note I also removed the tick visibility and plotted all of the health relevance text. There were also several formatting specifications implemented to emulate the original figure.

The next figure was Principal component analysis (PCA) of absolute microbial abundances at the genus level, with colors for SIBO and non-SIBO, along with specification for non-SIBO and *Lactobacillus*. This was also structurally split into 3 main components: preparing the data, performing the PCA, and plotting the first to principal components.

To begin I imported the relevant data in a similar fashion to the previous, using pickle. The paper specifies that the data followed a log base 10 transformation. I used numpy to transform this initial data as specified by the paper. I then used sci-kit learn’s implementation of PCA to perform Principal Component Analysis on the $\log_{10}$ scaled data. I then labeled each principal components from PC1 all the way to PC250. To achieve the coloring and shapes the original plot possessed, I added the relevant metadata for whether each sample was SIBO or non-SIBO and whether it was *Lactobacillus*. To do this I added each column and then performed a for loop with an if-else statement to add a column for with the corresponding yes/no (1/0) if the read was both non-SIBO and *Lactobacillus.*

Finally, I was able to plot the data after all of the previous preparation and analysis. To plot I used seaborn’s scatterplot. Of note, to color and shape code the data I added the hue and style arguments with the corresponding metadata.

Results/Conclusions

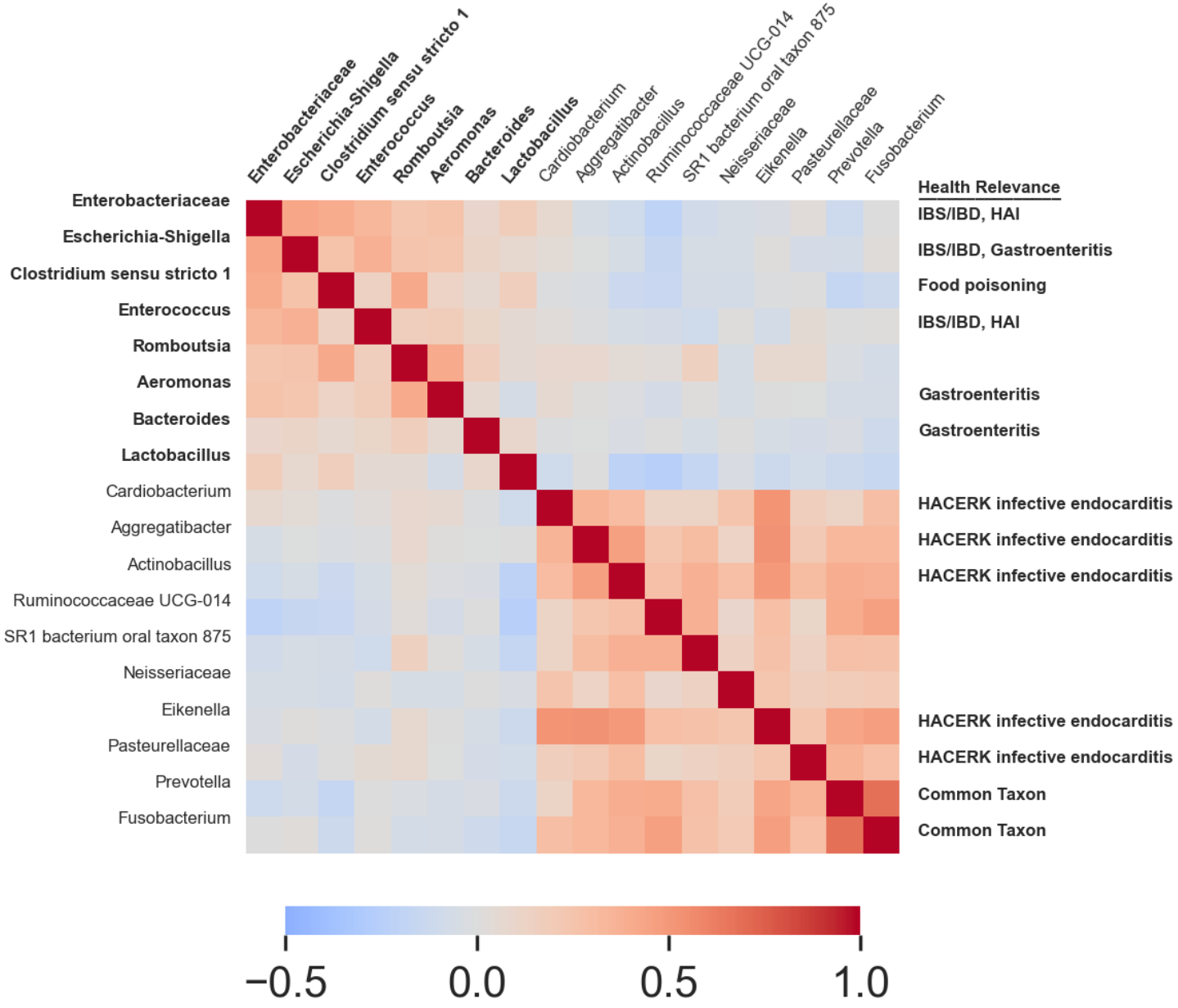

Figure 3B:

The clustered heatmap of the top 16 taxa display two distinct taxonomic signatures. The top left, with higher abundance contained *Enterobacteriaceae* to *Bacteroides*. The second group in the lower right block was found in lower absolute abundance and contained many HACERK infective endocarditis associated taxa. The bolded taxa were labeled as such to provide an indicator of "disrupters" which are associated with higher cases of SIBO and more severe GI symptoms. These disruptors are found in few samples, but when they are present, they tend to dominate and result in high abundance.[1]

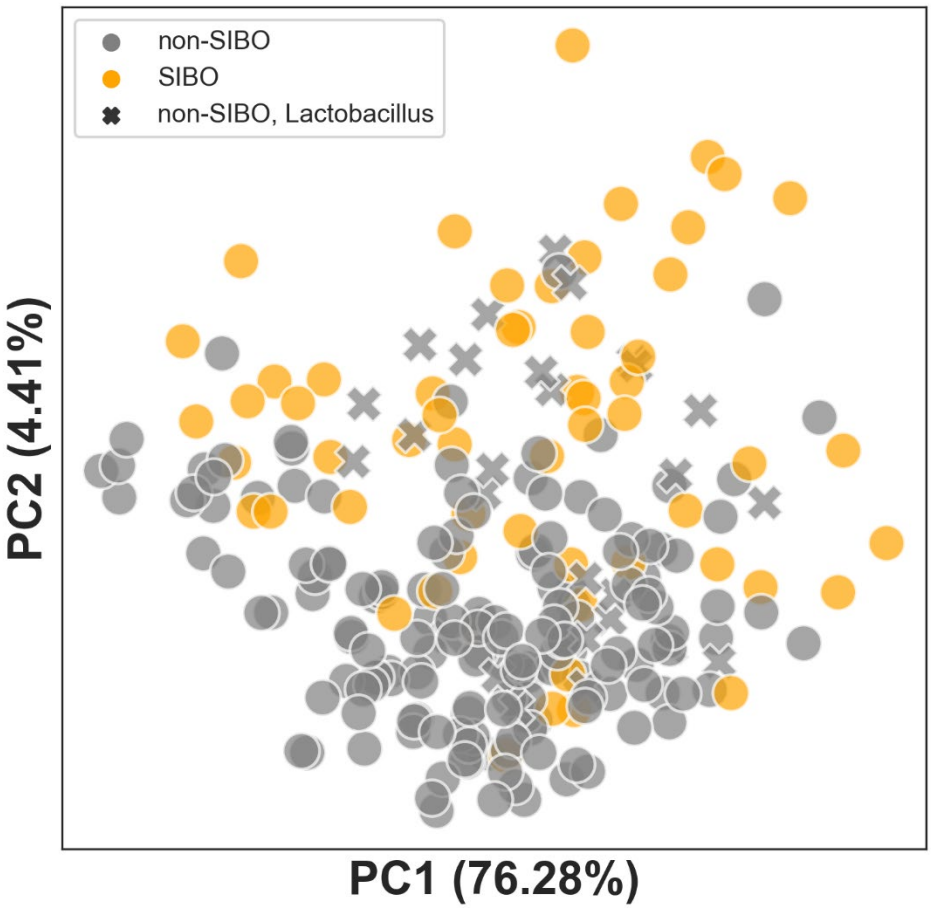

Figure 5A:

The PCA plot is a nice representation of the data as most of the variation can be visualized in the first two principal components. The principal component 1 axis correlates with total microbial load and the PC2 axis correlates with the abundance of disruptor taxa. The data was plotted to investigate whether the disrupter taxa were associated with SIBO. As an initial step the PCA was plotted for the first two principal components which indicates an enrichment in patients with SIBO and in a similar positive direction to the disruptor taxa (*Lactobacillus*). Most of the non-SIBO that clustered with SIBO also contained *Lactobacillus*. Authors mention this clustering may be due to the nature of *Lactobacillus* possibly being anaerobic.

Overall, the analysis suggests that disruptive taxa play a unique role in the small intestinal region. This research provides a step forward in the characterization of the small intestine microbiome.

References

1.) Barlow, J. T., Leite, G., Romano, A. E., Sedighi, R., Chang, C., Celly, S., Rezaie, A., Mathur, R., Pimentel, M., & Ismagilov, R. F. (2021). Quantitative sequencing clarifies the role of Disruptor Taxa, oral microbiota, and strict anaerobes in the human small-intestine microbiome. *Microbiome*, *9*(1). https://doi.org/10.1186/s40168-021-01162-2