TIME SERIES ANALYSIS OF HUMAN MICROBIOTA

ECES T480/680 Final Project

Rebecca Cargan, Bryan Featherstone, and Bairavi Venkatesh

Winter 2017

Abstract

The study of microbiomes is an exciting field that has gained support and interest from the scientific community because of the integral roles they play in digestion, metabolism, production of vitamins, and immune response (Guarner and Malagelada 2003). In addition, studies have shown bacterial communities are affected by diet, geography, habitat richness, phylogeny and physiology (Amato *et al.* 2013; Grover et al. 2014; Gomez et al. 2015; Ley *et al.* 2008). However, few studies investigate temporal dynamics of host lifestyle changes and their effects on microbiomes. Here we use David et al.'s work that uses 10,000 longitudinal measurements of host lifestyle to identify correlations with saliva and stool microbiota dynamics. We have identified dysbiosis events in both saliva and stool sites when the subjects either traveled abroad or became ill. Interestingly, the gut dynamics were temporarily altered during travel, but were permanently altered after an illness. FFT analysis showed that some bacteria, specifically within the Lachnospiraceae family, appeared in increased abundance at constant time intervals over the course of the year regardless of environment changes, whereas bacteria such as Ruminococcaceae *Ruminococcaceae* appeared with a regular period only when the environment was unchanged. ARIMA analysis on two distinct groups, *Clostridiales* and *Bacteroidales,* revealed that while *Clostridiales* is significantly affected by changes in host lifestyle, *Bacteroidales* relative abundances remain consistent throughout the duration of the experiment.

Background and Introduction

The concept of studying the change in human microbiota over time, and across varying environments, is described at length in the paper published by David et al, "Host lifestyle affects human microbiota on daily timescales." In this study, samples are collected daily from two subjects, Person A and Person B, based in the United States over the course of a year. Person A donated samples of both stool and saliva each day, whereas Person B donated only stool. Both Person A and B were expected to record their nutritional intake on a daily basis to better identify trends between diet and bacterial abundance. The initial study found large deviations in the abundance of certain gut bacteria clusters when Person A traveled to a developing nation in Southeast Asia and when Person B was infected with salmonella.

David et al. uses a 97% sequence similarity threshold to bin sequences into like groups called operational taxonomic units (OTUs). This method is used to help solve the problem of sequencing errors. However, studies have shown the OTU method doesn't provide fine scale sequence resolution which may provide additional insight into species or strains of bacteria present in the metagenome (Callahan, 1). Our project moves away from clustering bacteria into OTUs to exploring a finer sequence resolution method that identifies unique sequences within our samples based on an Illumina sequencing error model using DADA2.

Methods

David et al. provided demultiplexed Fastq files for our analysis using Illumina GAIIx. Prior to downstream analysis of these data, the DADA2 pipeline was used to prime each sample by filtering, performing sample inferencing, and then generating both sequence and taxa tables which could ultimately be merged with the metadata provided by David et al. First, parameters for filtering the sequenced reads were determined by visual inspection of quality plots (Figure 1).

Samples were truncated at a length of 95 nucleotides to disregard cycles with quality scores less than 30 (Figure 2). Then the first 10 nucleotides were trimmed on each sequence which is suggested because of common sequencing errors in Illumina Sequencing Technology. Sample inferencing was performed post-filtration by applying DADA2's parametric error model to the data. This algorithm begins with an initial guess and assumes the maximum possible sequencing error rates in the data. This is followed by alternating estimation of error rates and the composition of the samples until a joint solution is converged on (Callahan, 1). A subsample of 2 million reads were used to first learn the error rates, then model the inherent sequencing error, and finally remove any chimeras. The final output from the DADA2 pipeline was a sequence table of 1560 sequences and their relative abundances across 750 samples. Sequence variants were then assigned taxonomies within the DADA2 package by using the Silva Database (v.123).

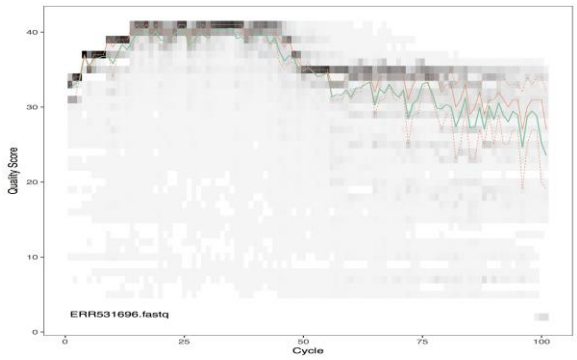

Figure 1-- Quality Score Plot of Sequences Pre-Filtering

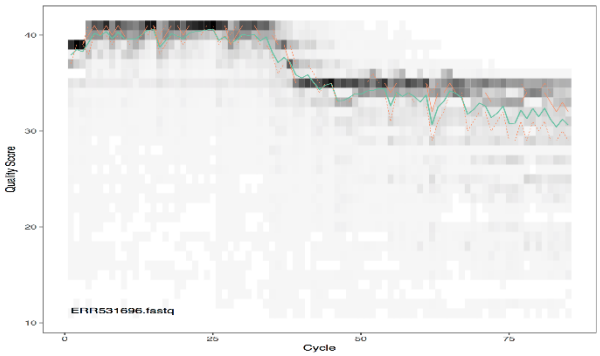

Figure 2 – Quality Score Plot of Sequences Post-Filtering

In conducting the time series analysis of Subject A and B's stool and saliva samples, three separate methods were used. Namely, diversity and ordination plots were created, fast Fourier (FFT) analysis was performed, and autoregressive moving average models (ARIMA) were developed. These methods were chosen as a means of understanding the data from different, yet complementary vantage points in an attempt to extend the study conducted by David et al.

Abundance, Diversity Measures and Ordination

To visualize our taxonomic table from DADA2 we calculated relative abundances for all taxonomic rankings. We filtered out any samples that weren't able to assign taxonomy to the Phylum level and lower unassigned taxonomic rankings were put into an 'Other' category for these analyses. To enhance visualization of the most abundant taxa present, we took the top 8 taxons present at each site and plotted this over the study period of 360 days for all taxonomic rankings. Next a Bray Curtis Dissimilarity measure was calculated between each sample to represent bacterial diversity. This distance matrix was then ordinated using a principle coordinate analysis (PCoA). Finally, a Jensen Shannon Divergence metric (JSD) was calculated using the relative abundance distribution for each sample within a site. The square root was taken to form a distance matrix and a heatmap was made to visualize the similarities between samples. A value of 0 represents exact similarity and 1 represents a complete dissimilarity. The heatmap was arranged from the beginning of the sampling period, day 1, to the end of the

sampling period, day 360. Several samples within our dataset were discarded after they showed complete dissimilarity with all other samples. These samples represented a 1 day time period and, when removed, had no effect on downstream analyses.

Fast Fourier Transform (FFT)

Fourier analysis was used to find and visualize periodicities in the data. The Fast Fourier Transform (FFT), an efficient version of the Discrete Fourier Transform (DFT), used the measured daily abundances of each sequence as discrete time steps to convert the signal in the time domain to its frequency components. Each measured sequence was analyzed separately over distinct time periods, typically two months long, in order to capture monthly recurrences while still differentiating between times when the subjects were healthy or sick and home or abroad. Each time period was the same length, or a multiple of that length, so that the resolution of the FFT would stay relatively constant for easy comparison of periodicities found over the course of the year. The FFT algorithm requires a defined amplitude at each discrete time point, so linear interpolation of missing points was used to ensure all time points contained a signal. Since the Fourier transform is able to identify multiple frequencies within a single time series, the top three most prominent “peaks,” representing the strongest frequency components, were identified.

ARIMA

The objective in performing ARIMA analysis was to identify trends in the time series data in an attempt to model both future and past values of the fluctuation exhibited by different bacterial populations. While there exist nonseasonal linear models that rely on three basic parameters (autoregressive, differencing, and moving average degrees), the ARIMA model used in the following analysis was multiplicative for one main reason. The analysis done with the FFT method (Figure 7) gave reason to believe that several bacterial populations exhibit periodicities in their abundances throughout the duration of the experiment. Therefore, using ARIMA models that exploit this inherent characteristic within the data were chosen.

Prior to modeling the data, preprocessing steps were taken. As discussed earlier, the output of the DADA2 pipeline yielded a sequence table with 1560 sequences. However, a closer look at the data showed that approximately 35% of the sequences had no values of abundance across any sample. These sequences were thus eliminated from analysis. Clustering was performed on the remaining sequences based on similarities in relative abundances across samples. The main steps in deeming sequences “similar” involved a maximum normalization step, followed by generating a distance map, and finally a threshold selection step.

The maximum normalization step was performed by summing the abundances of each sequence across all samples and dividing by the maximum value. Doing this step ensured that those data with only one or two high spikes in abundance would not confound results when generating the distance map. The Euclidean distance was then used to find the pairwise distances between sequences. Using these distances, visual inspection was used to select a threshold. As shown in Figure 3, sequences with varying levels of distances were plotted. Sequence similarity increased with decreasing values of difference and a threshold of 0.7 relative abundance units was chosen. All sequences with a difference less than 0.7 were thus

deemed similar to each other and clustered in groups. One representative sample from each cluster was chosen for ARIMA analysis to be performed.

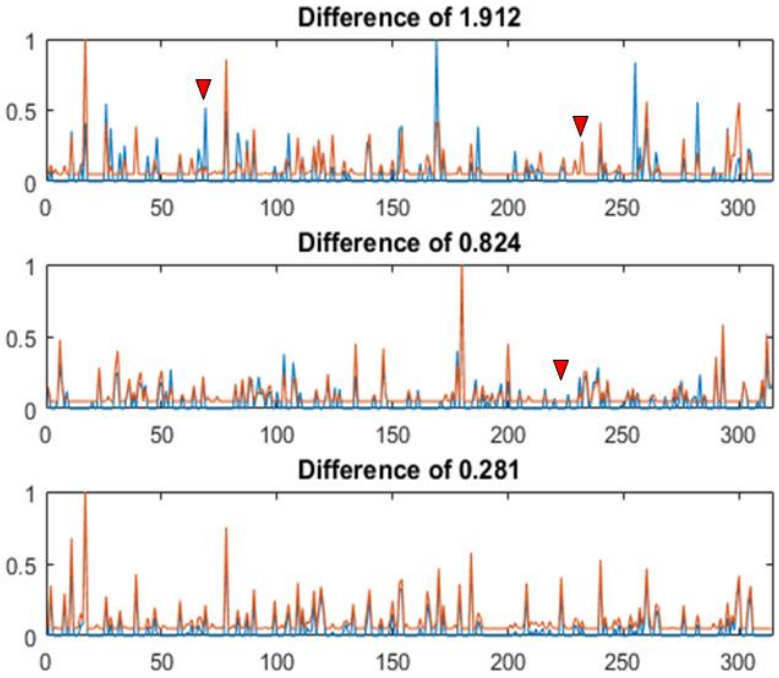

Figure 3 -- Similarity was assessed with Euclidean distance. Indications (red) emphasize differences in normalized sequences. Threshold for deeming sequences similar was chosen at a value of 0.7

Two main clusters, Order *Clostridiales* and Order *Bacteroidales*, were focused on for analysis. ARIMA prediction and comparison were performed for Subject A's stool and saliva samples. The main parameters that were adjusted in the models were those of seasonality and lags. Seasonality affects the periodicity with which the model approximates the true data, whereas the lag parameters affect the coefficients of the autoregressive polynomial which the model creates. The Kolmogorov-Smirnov (KS) test (α=0.1) was used to determine whether the fitted model and the raw data are derived from populations with the same distribution.

Results

I. DADA2 Taxonomic Table

The top 8 most abundant bacterial Orders in all sample sites were Bacteriodales, Clostridiales, Erysipelotrichales, Enterobacteriales, Neisseriales, Lactobacillales, Selenomonadales, and Pasteurellales (Figure 4). Gaps within the bar graphs represent days without samples. Donor A shows a more complete sampling period throughout the study while Donor B lacks data for the later part of the study. Because of this we only explored Donor A's saliva and stool samples for downstream analyses. Our project focuses on bacterial Order because a dramatic decrease in Taxonomic assignment was seen with lower taxonomic rankings. However, we attempt to use the finest scale resolution when conducting our downstream analyses without sacrificing sample size.

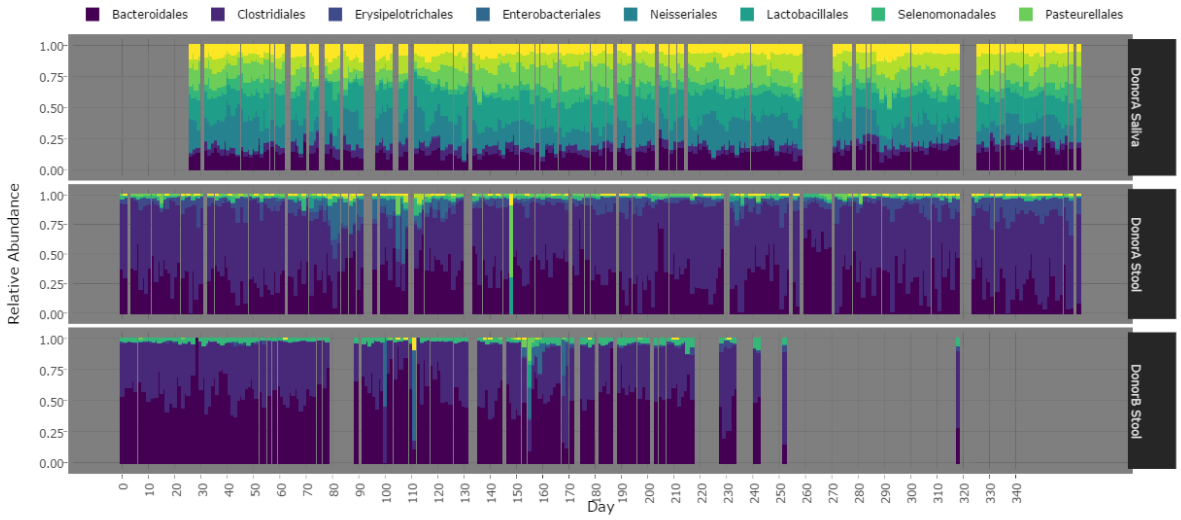

Figure 4 - Relative abundance of the top 8 bacterial orders for each sample site. The x-axis represents the collection day of a given sample and the y-axis represents the relative abundance of given order of bacteria. Donor A saliva, Donor A stool, and Donor B stool are grouped together from top to bottom respectively. Each vertical line represents a sample from one day and areas with no bars are days when no samples were collected from Donors.

II. Diversity and Ordination

We compared diversity metrics between all samples and within sites to better understand the similarity between microbiota and time periods. These analyses were used to identify time periods that were useful for our FFT and ARIMA analyses. First, a Bray Curtis Dissimilarity metric was calculated between all samples to create a distance matrix. This matrix was then ordinated using a PCoA to maximize the variation between sample points (Figure 5). Axis 1 explained 51.2% of the variation between our sample points which separated our saliva samples and stool samples. This agrees with other studies findings that bacterial composition are more similar within sites than between sites. Axis 2 explained 13.8% of the variation and separated Donor A stool samples from Donor B stool samples. It is important to note two things; 1. Other axes were present in this analysis, however, the percent of variation explained was minimal. 2. More variation was seen within the stool samples compared to the saliva sample. This variation could be explained by two reasons. The gut microbiome has been shown to be the most diverse site within the human body and the bacterial composition fluctuates or changes over small time scales.

Next we performed a JSD metric to our relative abundance distributions. JSD values range from 0 to 1 with 0 representing exact similarity and 1 representing complete dissimilarity. This distance matrix was then visualized using a heatmap to represent the similarity of two sample distributions within a site over time (Figure 6). Donor A Saliva and Donor A Stool heatmaps suggest a dysbiosis from Days 70 to 120. This change can be attributed to a 50 day time window in which Donor A traveled abroad. In addition, there is a short time window at day 225 where a dysbiosis occurs again within both sites of Donor A. However, there is a time lag between Donor A saliva and Donor A stool at this dysbiosis event. We were not able to correlate this phenomenon to any of the 10,000 measurements provided by David et al. Donor B Stool heatmap shows a change in bacterial composition on day 150. This change was see after Donor B became sick with a salmonella outbreak. Surprisingly, the gut dynamics were temporarily altered during travel, but were permanently altered after a salmonella outbreak.

This suggests that daily lifestyle changes don't have long term effects on microbiota. However, illness or pathogens seem permanently alter the gut composition.

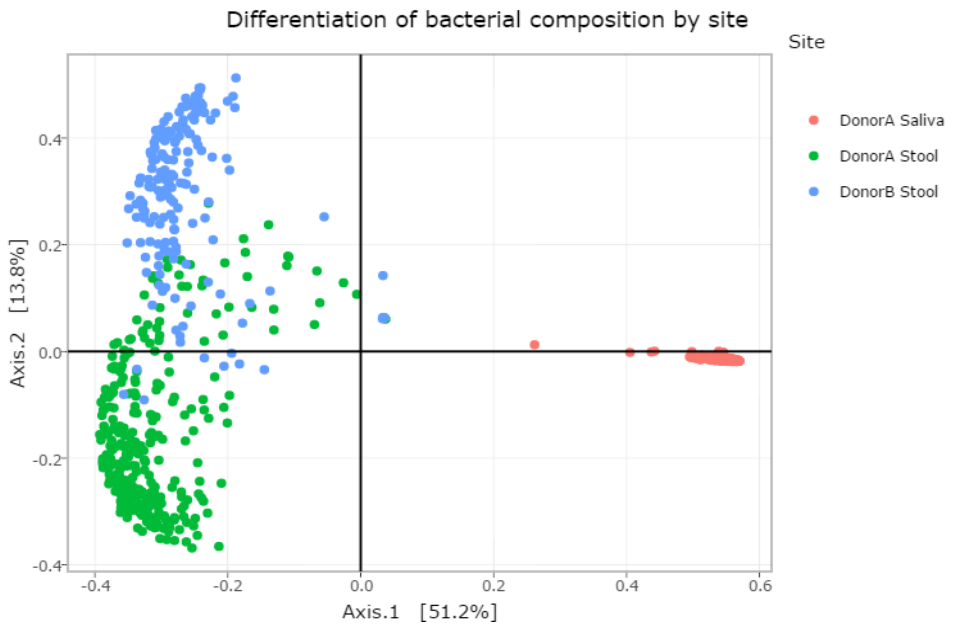

Figure 5 - Differentiation of bacterial composition by site. Each dot represents a single sample on a given day with Donor A saliva, Donor A stool and Donor B stool represented by red, green, and blue respectively. Axis 1 explains 51.2% variation among samples (horizontal separation) and Axis 2 explains 13.8% (vertical separation).

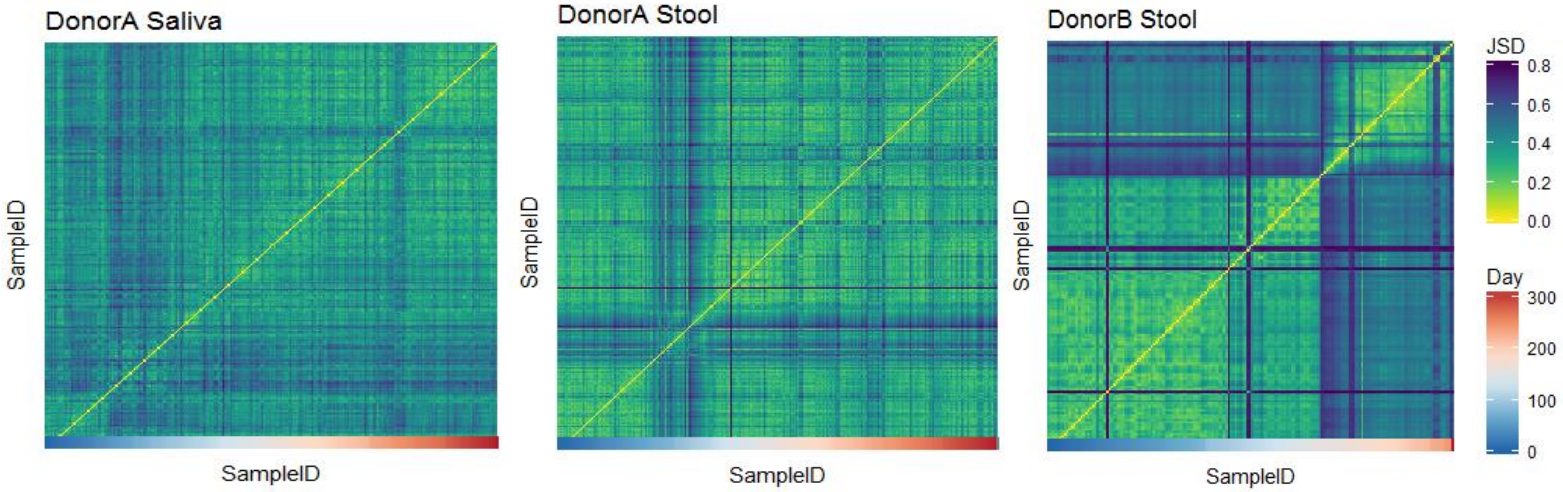

Figure 6 -Jensen Shannon Divergence Index of relative bacterial abundance within sample sites. Samples are labeled on both the x-axis and y-axis with the day samples represented by a bar at the bottom of the graphs. JSD values range from 0 to 1 with 0 representing exact similarity (yellow) and 1 representing complete dissimilarity (blue).

.

Further studies focusing on these effects of pathogens and microbiomes could provide insight on the mechanics of illnesses or diseases within a microbiome or allow for personalized treatment and the use of fecal transplants to recover from a dysbiosis. Overall, Donor A stool

and Donor B stool show a distinct dissimilarity when dysbiosis events occur which contrasts with Donor A saliva. This suggests the gut microbiome is more easily altered by host lifestyle changes compared to saliva. Thus, future temporal studies should focus on the gut microbiome.

III. FFT

Although all sequences were run through the FFT algorithm, there were a few of particular interest. Sequences identified as belonging to the family "Lachnospiraceae" showed particularly strong similarities to each other in terms of periodicity. Table 1 shows the three most prominent periodicities, in days, with which each of four genuses within the family appear. It can be seen that the genuses Anaerostipes, Blautia, and Fusicatenicater have a common period throughout most of the year. The genus Coprococcus 2, on the other hand, does not seem to follow the same pattern as the other genuses or even follow an identifiable pattern throughout. The three primary genuses also seem to maintain their periodicities when Person A goes abroad and returns, which is consistent with studies of the occurrence of this family which show that abundances of Lachnospiraceae within the gut of humans are specific to that human's genetic makeup rather than the environment they are in (Eren,1) (Meehan,1).

Table 1 -- Periodicity of High Bacterial Abundance within Family Lachnospiraceae

Genus	Periods Found Days 1-61 (days)	Periods Found Days 61-121 (days)	Periods Found Days 121-181 (days)	Periods Found Days 181-361 (days)
Anaerostipes	20,7.5,2.4	2.86,6,10	2,6,60	2.43,5.62,60
Blautia	20,3,6	2.86,10,6	4,5,2	2.91,5.62,60
Fusicatenicater	20,2,7.5	10,60,2.86	2.86,5,8.57	5.62,2.43,3.91
Coprococcus	6,7.5,15	2,15,2.61	3.33,6.67,2.14	5.81,12.9,6.92

Other bacteria, such as the genus "Ruminococcaceae: *Ruminococcaceae*" have strong periodicities that seem to change with their environment. In Figure 7 below, the periodic nature of this particular genus is seen before Person A travels, while he is away, and then upon his return. The period of note in this case is 20 days, which disappears during his trip days 71 through 122. Also of note is that there is a periodicity of approximately 3 days throughout the duration of the entire year, which may imply that the abundance of this bacteria is not entirely dependent on environment.

Fourier analysis of Saliva, on the other hand, showed similar but inconsistent periodicities across families, genuses and time steps, and therefore was not used in the comparison of sequence results. Data from Person B was also excluded from the analysis due to the large number of days without recorded sample data, which caused entire periods of data to disappear.

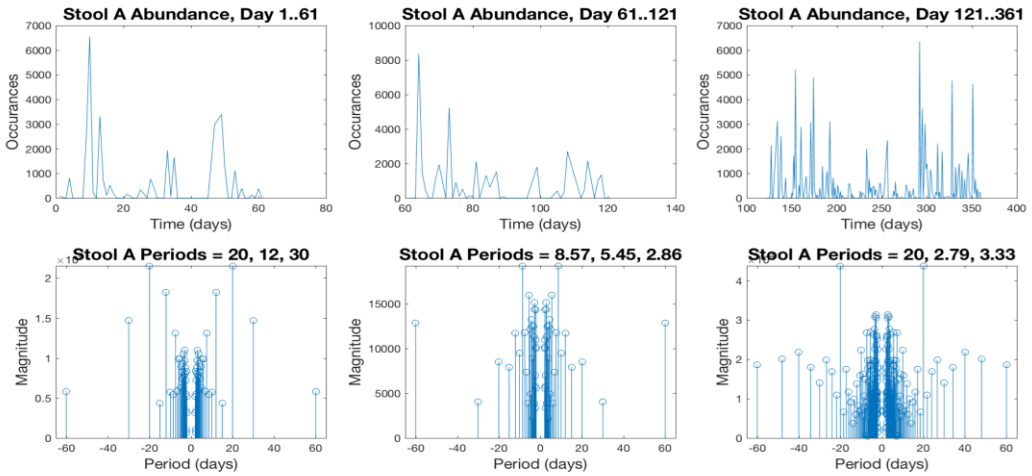

Figure 7-- Abundance and Periodicity of Ruminococcaceae: Ruminococcaceae Genus Over One Year

IV. ARIMA

Results from performing ARIMA on two clusters across all taxa revealed that there are distinct differences between *Clostridiales* and *Bacteroidales*. Figure 8 shows prediction performed on a range of data before Subject A traveled abroad. Data post-travel (specifically days 150 to 250) were used as a training set. Predictions for both of these clusters yielded an output of 0 from (α=0.1) from the KS test, implying that both the raw and fitted data are from the same distribution. This suggests that Subject A's gut and saliva compositions of the *Clostridiales* bacterial group remains consistent when a normal routine is assumed. However, the periodicity of flare ups in this bacteria are more frequent in the saliva (roughly every 10 days) than they appear to be in the gut (every 25 days). This finding parallels Segata et al's study in which significantly less abundance of *Clostridiales* were found in the stool as compared to in the saliva (Segata et al, 7). Using an ARIMA model to predict the data between days 70 and 120, while the subject was traveling, yielded a prediction curve with a KS value of 1 and a mean squared error of 5E7, suggesting that the population trends in *Clostridiales* are dramatically affected with a change in the host lifestyle.

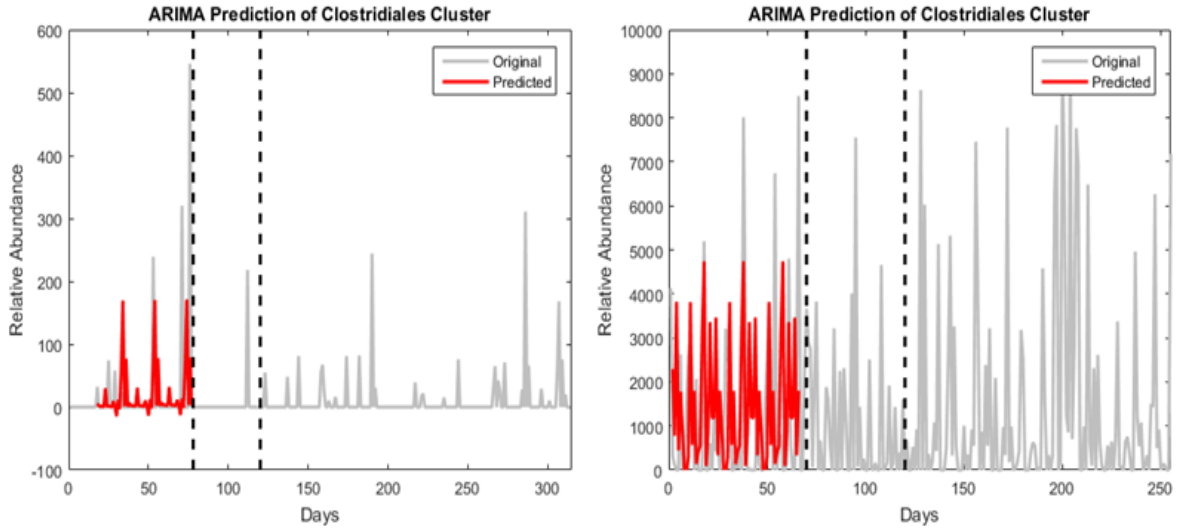

Figure 8-- ARIMA model prediction of "normal" data pre-travel using post-travel "normal" data in Subject A's stool (Left) and saliva (Right) samples in Clostridiales cluster.

The *Bacteroidales* cluster, on the other hand, behaves quite differently from the *Clostridiales* group. When ARIMA models were used to predict trends in the data, it was found that a single model could be used relatively well on the entire data set (Figure 9). This implies that the population abundance of *Bacteroidales* does not seem to be highly affected with changes in host lifestyle, either in the gut or saliva. Yet, like the *Clostridiales* cluster, the *Bacteroidales* cluster also seems to be more abundant in the saliva than in the stool.

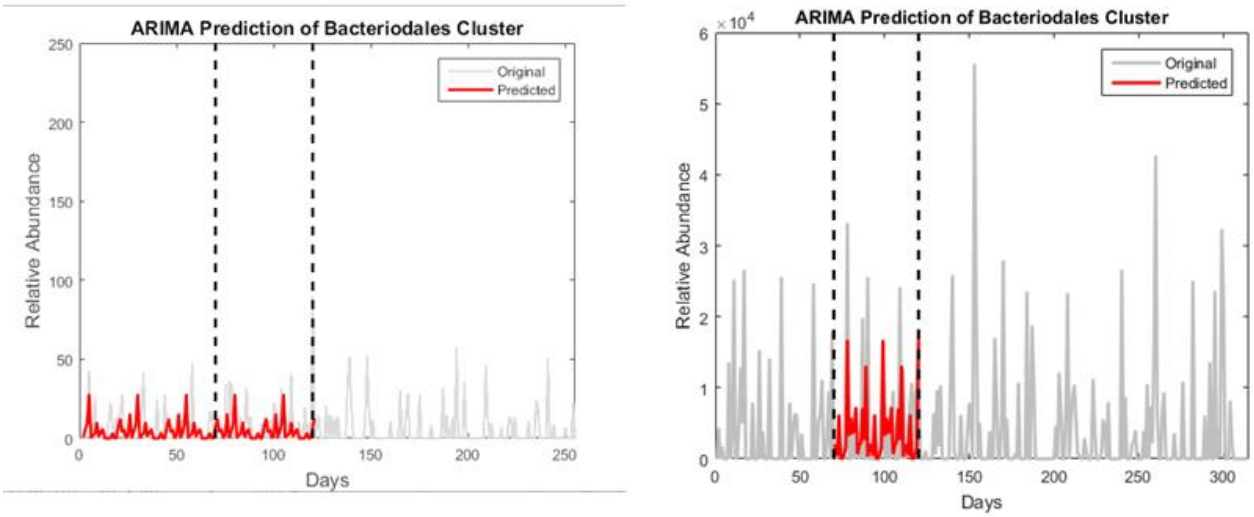

Figure 9-- ARIMA model prediction of "normal" data pre-travel using post-travel "normal" data in Subject A's stool (Left) and saliva (Right) samples in Bacteroidales cluster.

Some limitations worth noticing in using an ARIMA model implemented through the 'arima' function in MATLAB is that the function only allowed for integer values for the seasonality parameter. Values with more precision, like the periods found in the FFT analysis may have perhaps yielded predictions of the data that were more representative of the true signal. The use of the ARIMA model is also sensitive to missing data which can confound and negatively affect model predictions. For this reason, Subject B's data, which was missing approximately 15% of data was not analyzed. Additionally, the ARIMA model requires that at least 50 observation points be used for forecast, however Subject B fell ill only for 9 days, which would not have been a big enough sample to make predictions from.

Conclusion

Our project was designed around David et al.'s study using a time series analysis approach to complement their research and gain additional insight into host lifestyle and its effects on the microbiome. The major difference in our analyses is that we use a fine scale sequence resolution approach, relying on the DADA2 pipeline, to generate a sequence and taxonomic table compared to the commonly used OTU method. In addition, we use an FFT and ARIMA analysis to find periodicities and trends within our data and thus correlate changes in bacterial composition to those in host lifestyle. Furthermore, we attempt to model these changes over time and use this model to accurately predict the fluctuations within the microbiome.

Sources

Amato, Katherine R., et al. "Habitat degradation impacts black howler monkey (Alouatta pigra) gastrointestinal microbiomes." The ISME journal 7.7 (2013): 1344-1353.

Callahan, Benjamin J., et al. "DADA2: high-resolution sample inference from Illumina amplicon data." Nature methods (2016).

Earnest, A., et al. "Using Autoregressive Integrated Moving Average (ARIMA) Models to Predict and Monitor the Number of Beds Occupied during a SARS Outbreak in a Tertiary Hospital in Singapore." BMC HEALTH SERVICES RESEARCH, vol. 5-1 (2005).

Eren AM, Sogin ML, Morrison HG, Vineis JH, Fisher JC, Newton RJ, et al. “A single genus in the gut microbiome reflects host preference and specificity.” ISME J. 2015;9:90–100.

Gomez, Andres, et al. "Gut microbiome composition and metabolomic profiles of wild western lowland gorillas (Gorilla gorilla gorilla) reflect host ecology." Molecular ecology 24.10 (2015): 2551-2565.

Grover, Madhusudan. "Role of gut pathogens in development of irritable bowel syndrome." Indian Journal of Medical Research 139.1 (2014): 11.

Guarner, Francisco, and Juan-R. Malagelada. "Gut flora in health and disease." *The Lancet* 361.9356 (2003): 512-519.

Meehan, Conor J., and Robert G. Beiko. “A Phylogenomic View of Ecological Specialization in the Lachnospiraceae, a Family of Digestive Tract-Associated Bacteria.” Genome Biology and Evolution 6.3 (2014): 703–713. PMC.

Segata, N., et al. "Composition of the Adult Digestive Tract Bacterial Microbiome Based on Seven Mouth Surfaces, Tonsils, Throat and Stool Samples." GENOME BIOLOGY, vol.13- 6, (2012)