Time-Series Analysis of Host Lifestyle and Human Microbiota

Final Report

Timothy Merkel, Max Paszek, Sean Stevens

Drexel University

March 24, 2017

Abstract

We used data provided from the study titled “Host lifestyle affects human microbiota on daily timescales”. The researchers examined the daily human micriobiota and linked them to various life circumstances [1]. Our project goal was to analyze the data collected from a Digital Signal Processing perspective. No members of the group have backgrounds in biology or metagenomics, so tried and trued analyses for time series data were used. These include, but are not limited to, correlation, arima, and periodicity plots. To achieve this, we utilized the programming language R to acquire the data and process it. The methods to manipulate the data will be discussed in greater detail within the Materials and Methods section.

Materials and Methods

Prior to analyzing the data, we first had to acquire the data found in [1]. In order to accomplish this, we first had to familiarize ourselves with R. Thankfully, the markdowns on the Bioinformatics GitHub page were available to use as reference [2]. The first step was to follow the r_sra markdown found on [2]. The first step to download an SRA file to connect to the SRA metadata database, and then download the files from the NCBI. After downloading the SRA files, one must obtain FASTQ files by using the SRA toolkit found on the NCBI's site. A FASTQ file is just a FASTA file that also includes the quality data of the sequence. Recently, FASTQ files have become the standard for storing high-throughput sequencing techniques [3].

After the FASTQ files were obtained, they were then filtered to find higher quality data measures. In [1], the researchers trimmed the first 10 left base pairs off of their sequence data, so we did the same. We also used a truncation length of zero in our program, this yields sequences that are 90 base pairs long. These sequences, of which there are 1,698 were then placed into a sequence table. The sequence table was created by using DADA 2, of which there are very detailed instructions found in the dada2 markdown on [2]. After completing the DADA 2 workflow, the sequence table below was generated. The code that was used to accomplished all of these steps are published on our GitHub page.

	ACAAGCGTTGT	CCGAGCGTTAT	ACAAGCGTTGT	GCGAGCGTTAT	CCGAGCGTTAT	CCGAGCGTTAT	GCGAGCGTTAA	GCGAGCGTTAT	CCGAGCGTTGT	GCAAGCGTTAT	GCGAGCGTTAA	CCGAGCGTTAT	GCAAGCGTTAT	GCGAGCGTTAT	GCAAGCGTTGT	ACAAGCGTTAT	GCAAGCGTTAT
ERR531441.fastq	49920	7792	5220	0	18743	12924	0	10907	0	10855	205	2395	11020	0	3762	0	4062
ERR531442.fastq	0	1960	6329	838	4307	0	0	0	0	20609	0	271	0	0	0	0	386
ERR531443.fastq	26031	23476	14848	0	11673	11677	0	2798	0	6028	210	9608	6695	0	1275	0	1841
ERR531444.fastq	761	15483	466	0	12767	8475	0	547	0	54	4	2879	126	0	64	0	869
ERR531445.fastq	7475	3825	5140	14444	3882	0	0	281	0	250	0	3147	869	0	0	0	311
ERR531446.fastq	5939	4725	3716	12737	1535	0	0	741	0	193	0	3563	643	0	0	0	334
ERR531447.fastq	18704	63718	78358	0	14638	8811	0	3122	0	12160	856	13775	5709	0	3469	0	4597
ERR531448.fastq	1427	1714	314	0	779	386	0	181	0	191	0	550	447	0	207	0	100
ERR531449.fastq	10552	12962	2266	0	1863	8985	0	3807	0	2103	151	924	2502	0	1541	0	1215
ERR531450.fastq	30189	41779	10019	0	16537	36607	0	15820	0	6836	1835	8556	5956	0	7849	23	1678
ERR531451.fastq	11126	19084	6216	0	2886	3983	0	4352	0	658	4064	2031	1346	0	2201	0	873
ERR531452.fastq	3227	3945	1116	0	691	3868	0	1246	0	629	5	200	1101	0	598	0	354
ERR531453.fastq	16316	18006	3005	0	1033	11152	0	7850	0	3587	3888	973	2651	0	708	2	1207
ERR531454.fastq	0	0	0	0	0	0	25730	0	12523	0	14242	0	0	12255	0	6210	0
ERR531455.fastq	0	0	0	0	0	0	17378	0	7333	0	6126	0	0	3321	0	2666	0
ERR531456.fastq	0	0	0	0	0	0	10514	0	7353	0	6735	0	0	2553	0	2118	0
ERR531457.fastq	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ERR531458.fastq	46493	29235	1536	7	19605	11761	0	41100	0	25773	350	19567	30103	0	11339	0	9863
ERR531459.fastq	2375	20191	4017	71477	11003	0	0	75	0	88	0	9495	634	0	0	0	389
ERR531460.fastq	9	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0
ERR531461.fastq	45402	31367	10888	0	12809	14759	0	7226	0	12717	2717	12526	9017	0	6531	0	3016
ERR531462.fastq	29488	38326	9549	0	24084	29300	0	4076	0	5859	195	18554	5061	0	3482	0	1448
ERR531463.fastq	0	0	0	0	0	0	19035	0	12011	0	3628	0	0	8004	0	4288	0
ERR531464.fastq	5179	11494	6389	17969	2900	0	0	405	0	228	0	2238	2013	0	0	0	675
ERR531465.fastq	5716	8137	1235	0	6062	5963	0	847	0	257	0	1366	1535	0	116	0	552
ERR531466.fastq	0	0	0	0	0	0	19300	0	14426	2	13731	0	0	12068	0	7042	0
ERR531467.fastq	0	0	0	0	0	0	13152	0	9408	0	7211	0	0	8519	0	4512	0
ERR531468.fastq	49662	14680	26982	0	1523	4266	0	45613	0	28229	1516	217	17557	0	6283	0	13728
ERR531469.fastq	3	0	0	0	0	0	10662	0	10626	0	17271	0	0	9144	0	6253	0
ERR531470.fastq	0	0	0	0	0	0	10694	0	8524	0	10510	0	0	5831	0	2020	0
ERR531471.fastq	0	0	0	0	0	0	7426	0	10662	0	14269	0	0	3289	0	3822	0
ERR531472.fastq	12262	9156	5794	20245	4465	0	0	1464	0	388	0	4471	2125	0	0	0	810
ERR531473.fastq	12455	40158	2763	0	14153	12425	0	12497	0	8932	145	5598	4280	0	2579	0	2249
ERR531474.fastq	0	0	0	0	0	0	14496	0	7069	0	1987	0	0	4436	0	2538	0
ERR531475.fastq	0	0	0	0	0	0	9868	0	12119	0	10495	0	0	2735	0	2514	0

Figure 1. The sequence table generated from the FASTQ files by using DADA 2.

The sequence table contains the SRA files names as the rows, then the abundances of the sequences as the columns. This is not the end of the process for preparing the data for R, however. Each SRA filename must be matched up with a sample ID from the study. In order to accomplish this, a metadata table was downloaded from [1]. This table is additional file 18, and contains both sample and nutritional metadata. The accession number for the nucleotides are deposited on the EBI/ENA database under accession number ERP006059. While there may be tools to accomplish the mapping of the taxa to the sequences, nobody in the group was comfortable with using such a tool. Therefore, Excel was utilized to manually manipulate the data and construct a table that contained the abundances of each taxa per each sample ID. The data was broken up into the following categories: saliva samples from patient A, stool samples from patient A, and stool samples from patient B. Tables were also generated

which contained the nutrition information for the tables mentioned previously. The results of this pre-processing then allowed the data to be processed in R.

With the data properly formatted, we were able to visualize and analyze our data by using R. The first plots which were constructed where correlation plots between the various phlya across all samples that were assigned to the sequences in the pre-processing stage.

Results and Discussion

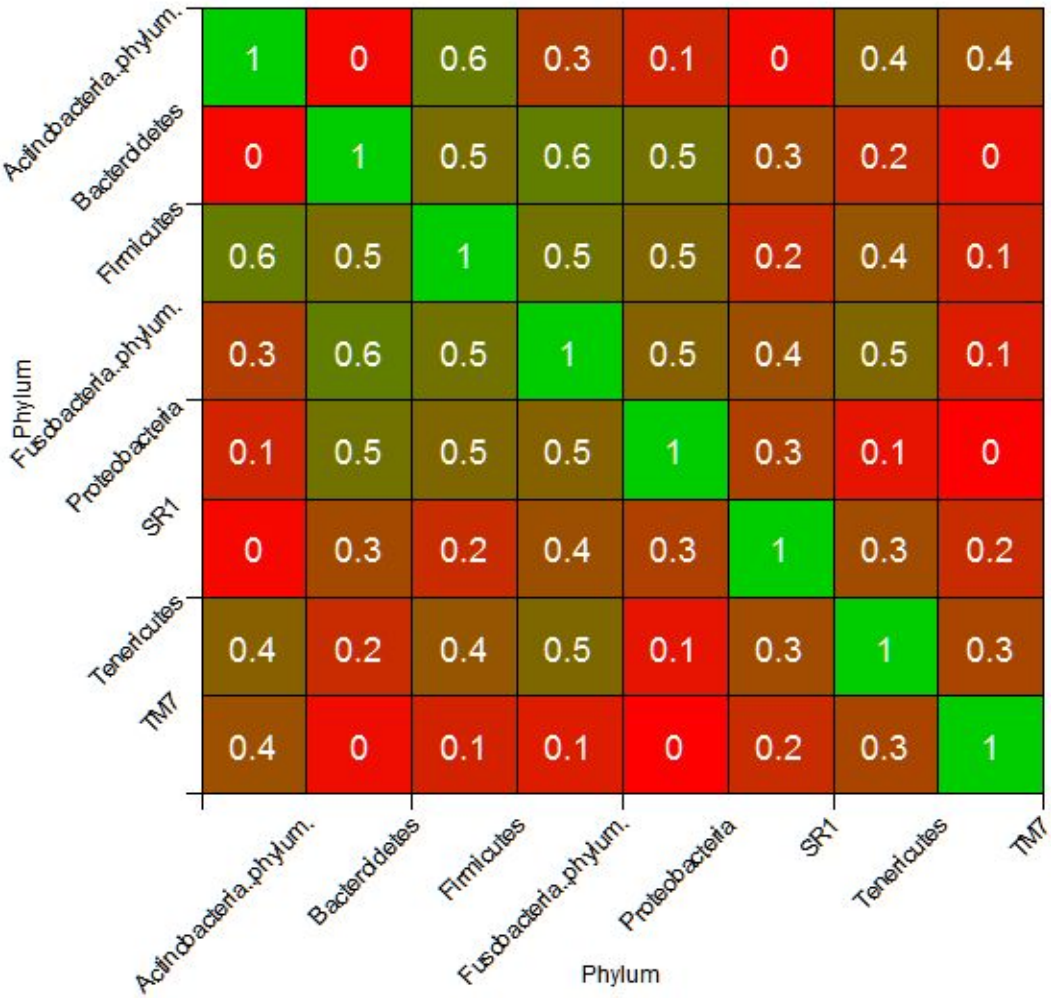

Figure 2. The correlation plot of patient A's saliva data across all samples.

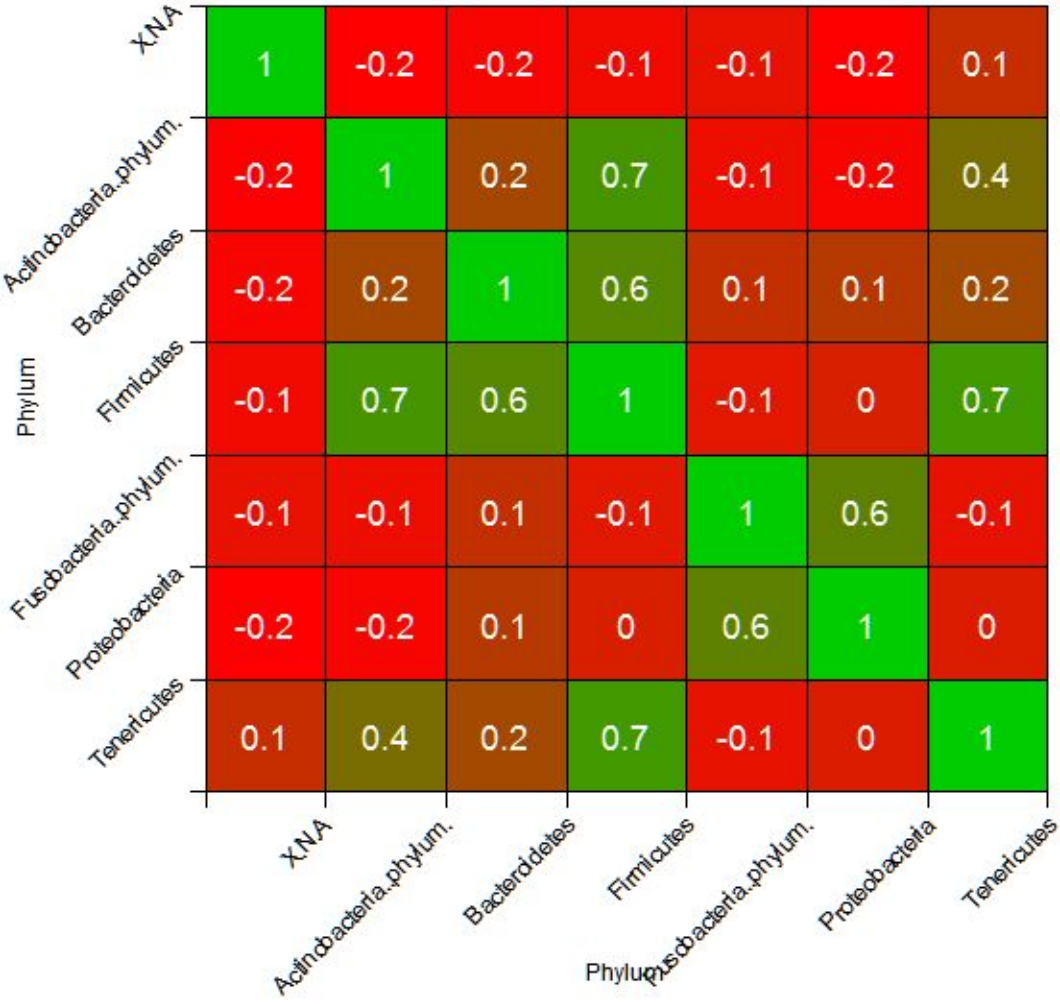

Figure 3. The correlation plot of patient A's stool data across all samples.

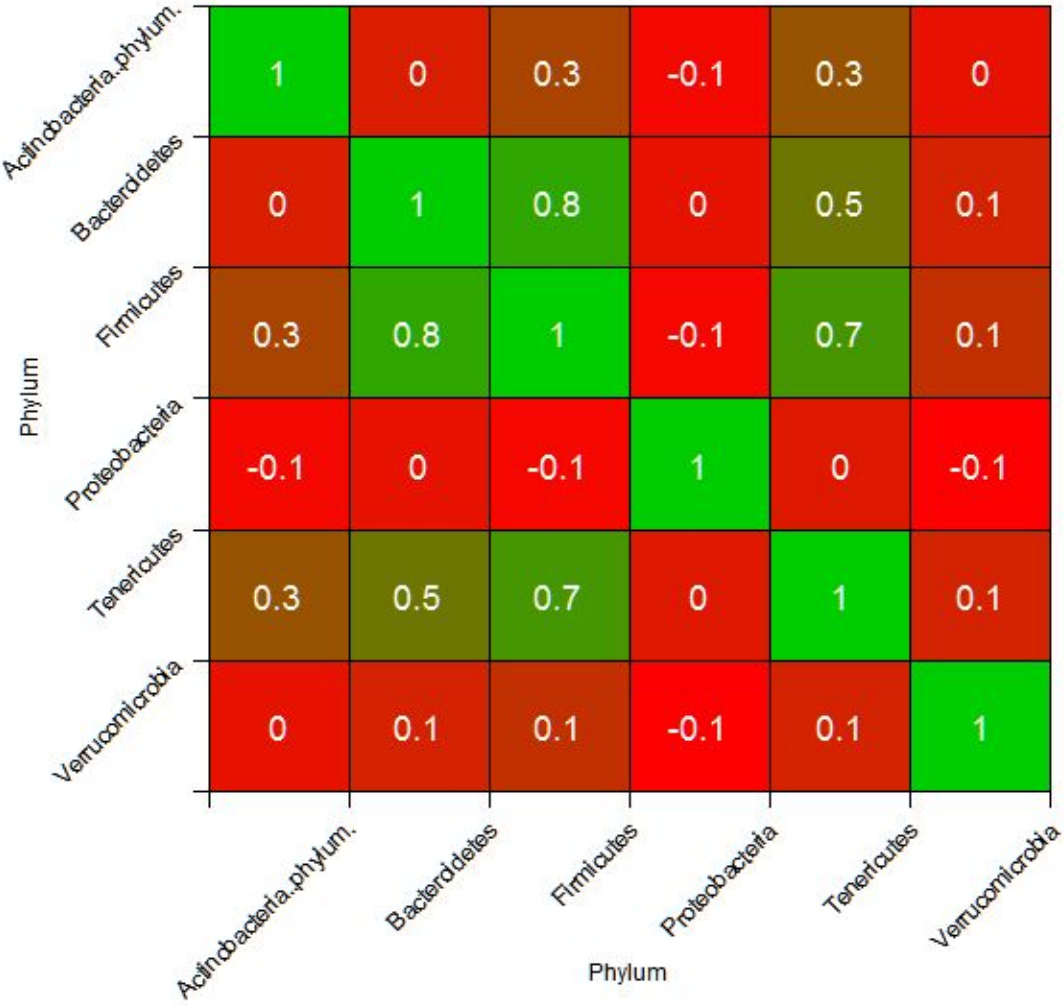

Figure 4. The correlation plot of patient B's stool data across all samples.

The cor() function in R was utilized to generate the correlation matrices, which were then plotted using the color2d.matplot() command.

This data is very interesting because it shows that not only are their correlations between the individual species in the data, but also that taxonomic units even as large as a

phylum have some pretty significant correlations. For instance, in Figure 2 it is clear that there is a strong positive correlation between Actinobacteria and Firmicutes. This specific example however is not consistent between all 3 samples. In fact, in patient B's stool the same pair has the most negative correlation of the whole sample set. One relationship that stands out is the relationship between Bacteroidetes and Firmicutes. This correlation is very positive in all 3 sample sets with a 0.8, 0.6 and 0.5. This must mean that there is some factor which causes these phylums to change at the same rate. Perhaps they have a symbiotic relationship, or perhaps they both just happen to be sensitive to the same nutritional intake.

Next the phylum counts per day were visualized. They are shown below in Figure 5, 6, and 7, and were generated using R's stackpoly() command. In these stacked area charts one of the most interesting observations is that the saliva of patient A and the stool of patient A contain a very different distribution of phylums. Both sets contain the same three phylums but the quantities change drastically. For example in the saliva of patient A it is clear that the majority of the microorganisms belong to the Firmicutes phylum. In the stool of patient A there was a period between 70 and 150 days where Bacteroidetes dominated.

Interestingly, this lines up very well with the time period where the patient went abroad. In the saliva of patient A this period also sees a change in that the bacteroidetes are drastically reduces. Given the correlation data between bacteroidetes and firmicutes

(one of the most correlated in the sample) it is extremely interesting that a decrease in the bacteroidetes in the saliva of patient A corresponds so well with the decrease in Firmicutes in the stool of the same patient over the same period.

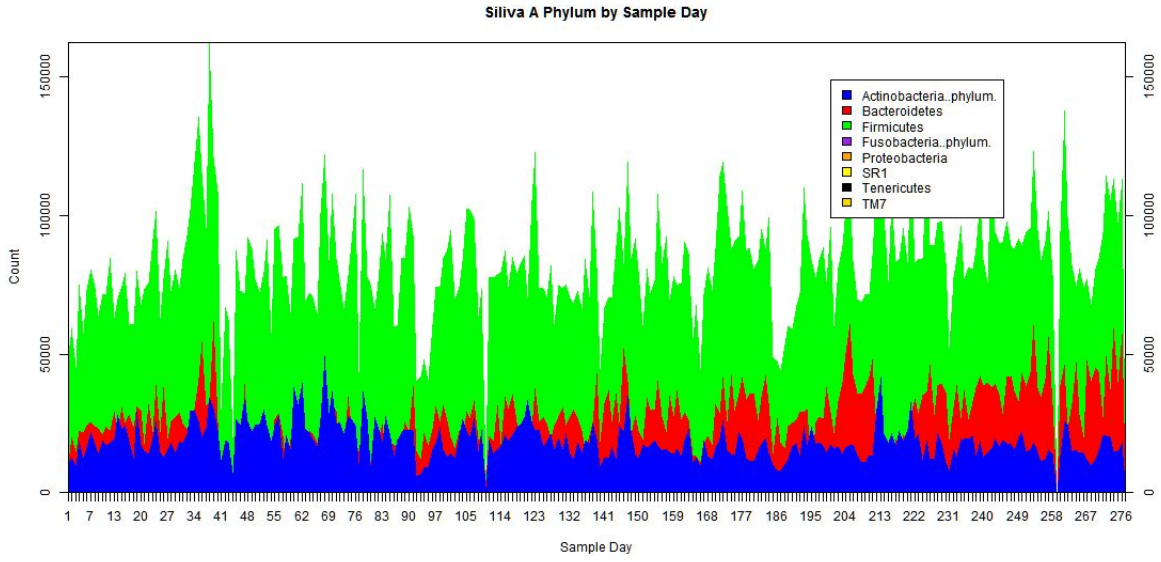

Figure 5. Patient A's saliva samples grouped by Phylum, per day.

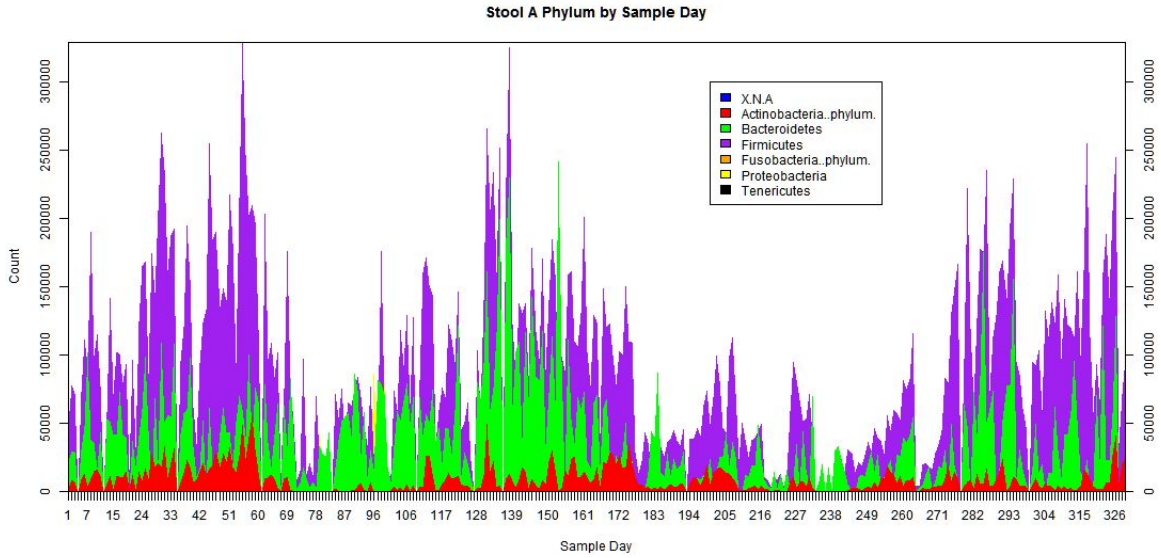

Figure 6. Patient A's stool samples grouped by Phylum, per day.

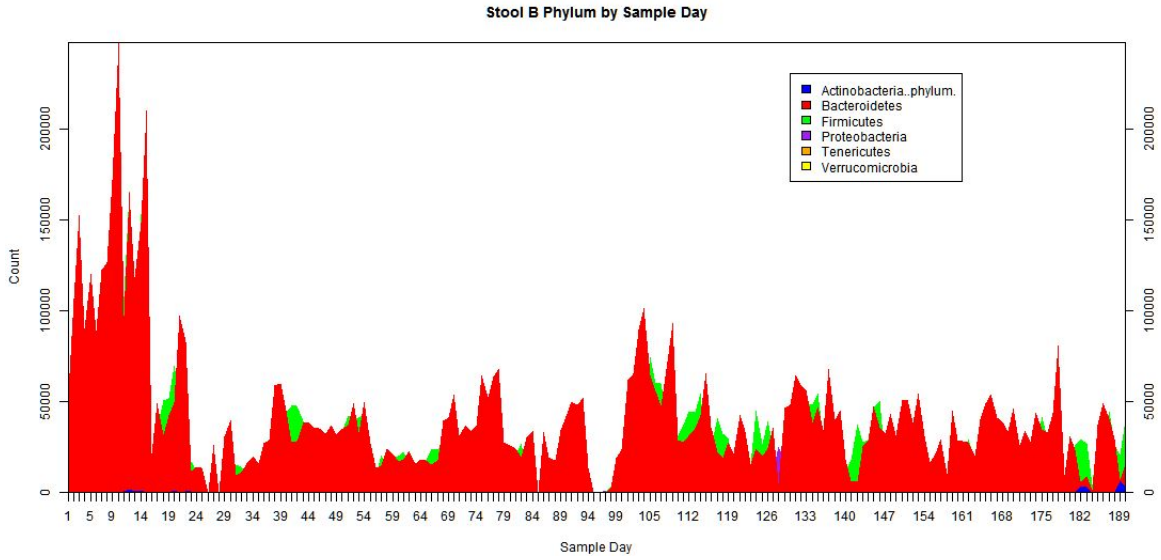

Figure 7. Patient B's stool samples grouped by Phylum, per day.

Arima plots for the three most common phyla in patient A's saliva were generated by using R's arima() command and plotted using the ggplot() command. Two different window lengths were used. In the first set the arima model was trained on the full data set and some future data was guessed. The arima parameters used on this set were (3,1,3) with seasonality parameters (0,1,1). These results can be found in Figure 8, 9, and 10 In the second set the arima model was trained on the first 40 days and then was used to attempt to model the time period when the patient was abroad. The arima parameters used on this set were also (3,1,3) with seasonality parameters (0,1,1).These results can be found in Figure 11, 12, and 13.

For the set that was trained on the full data the arima model seems reasonable. Each one has found a mean value that seems consistent with the current trend

of the data. The upper and lower limit do seem to line up reasonably well with the possible spikes of the data. This data is not incredibly interesting. The only conclusion there is to draw from thi is that the arima model is working relatively well.

The second set is much more interesting. In each of these figures it is clear that the arima model is no correct. The reason for this is that the patient has just gone abroad causing a change in habits, nutrition consumption, and living conditions. As a result the microorganisms inside their saliva became very different than the training set. Upon returning home their habits returned to normal and the arima model trained on the first 40 days would likely work once again.

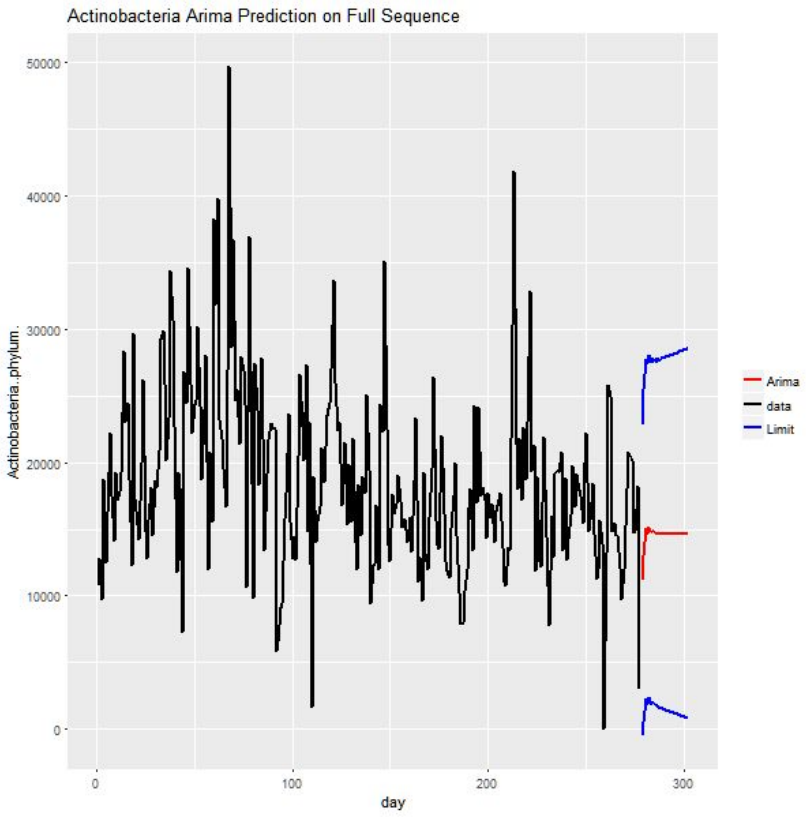

Figure 8. Patient A's saliva actinobacteria count arima forecast.

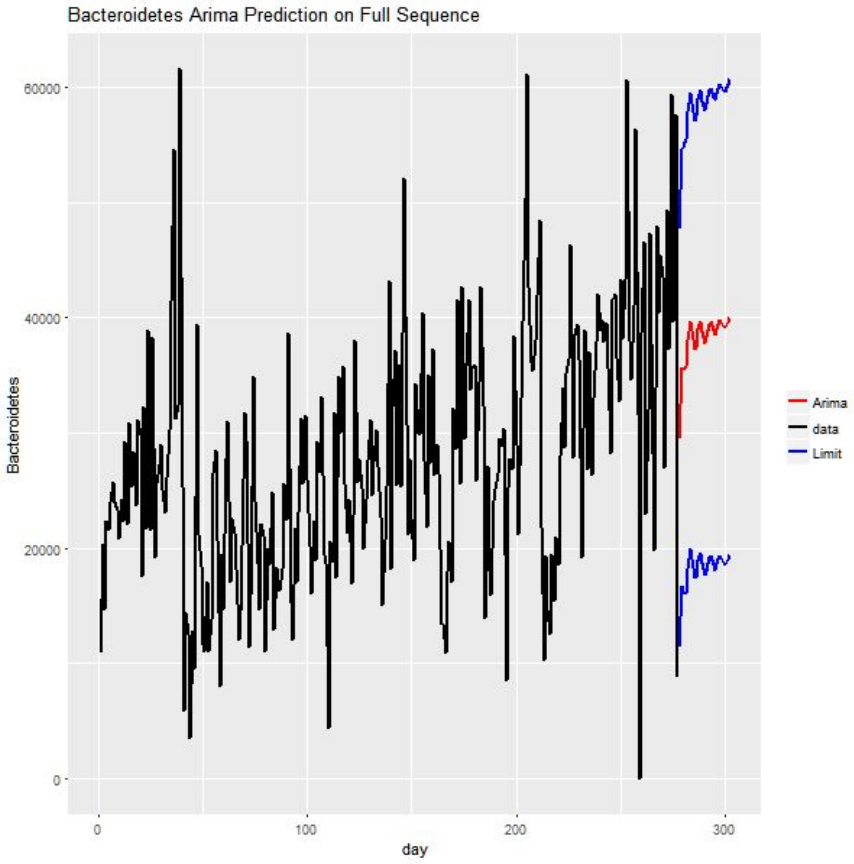

Figure 9. Patient A's saliva bacteriodites count arima forecast.

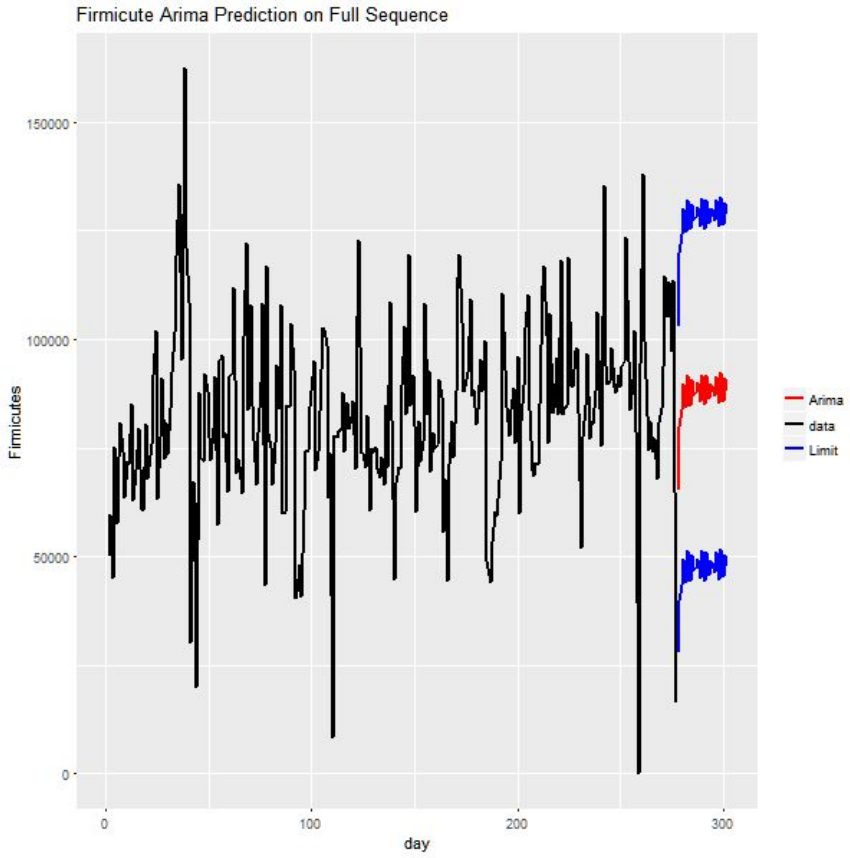

Figure 10. Patient A's saliva firmicutes count arima forecast.

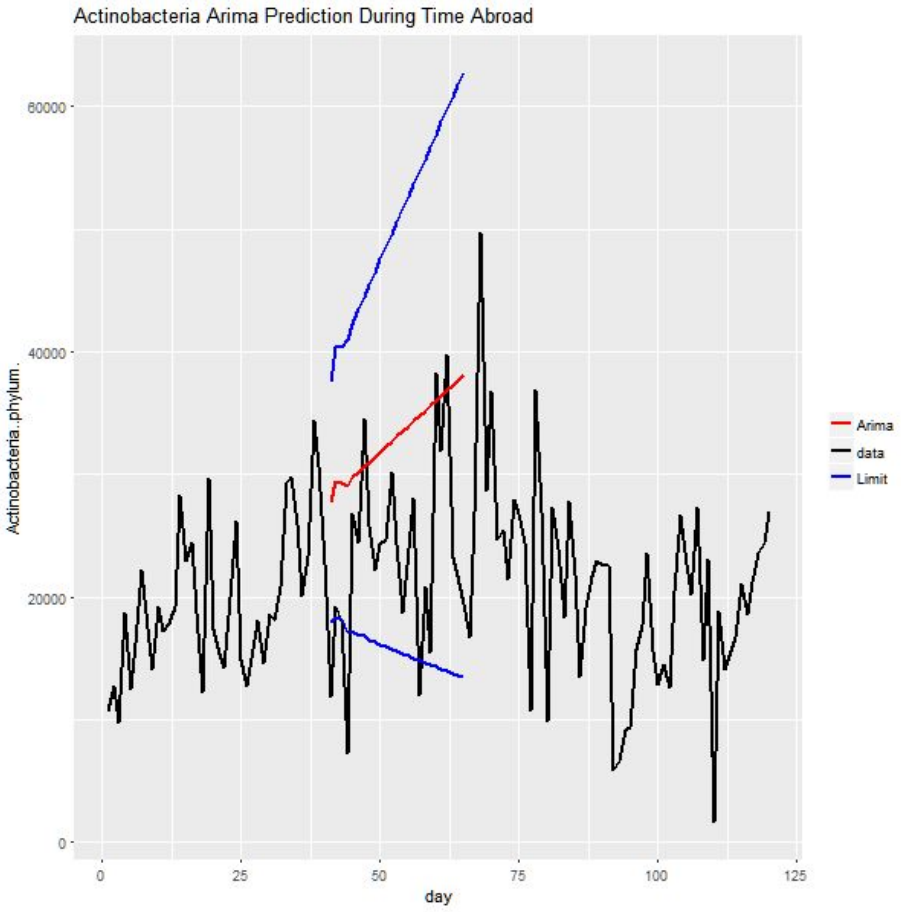

Figure 11. Patient A's saliva actinobacteria count arima forecast against truth while abroad.

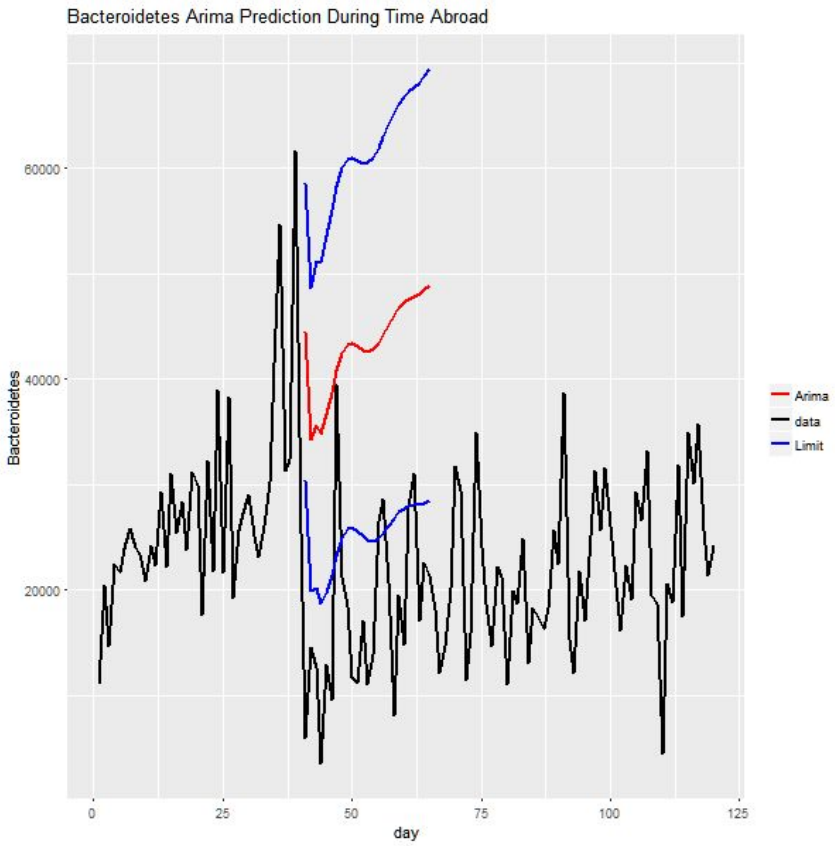

Figure 12. Patient A's saliva bacteriodite count arima forecast against truth while abroad.

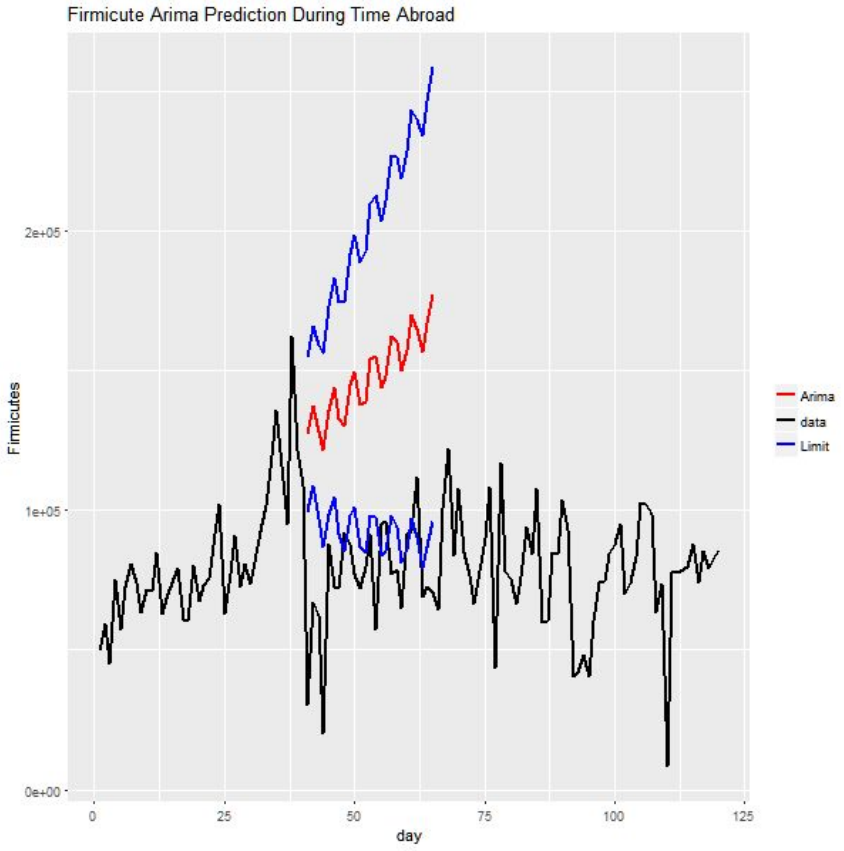

Figure 13. Patient A's saliva firmicutes count arima forecast against truth while abroad.

While all of the above factors are just merely representations of bacterial concentration against time, the data was also correlated with the nutritional intake that was recorded by the participants. Only patient A recorded their nutritional intake, so the correlations could only be found in patient A's saliva and stool. Figure 14 and 15 below show the results of these correlations.

One thing that is interesting about the saliva data (figure 14) is that there are a lot of zeros present in this data set. While there are no strong positive or negative

correlations a zero means that the phylum does not respond at all to that stimulus. One example is that the only phylums which respond to saturated fat are Tenericutes and Firmicutes. All others present in this data set are not affected at all.

The nutritional intake seems to affect the stool of patient A significantly more than the saliva. This makes sense because the stool is a better representation of the stomach than the saliva. Nutritional intake collects in the stomach so those microorganisms will respond to in more vibrantly. Still there are no outstanding correlations. However, as seen in figure 15, it is clear that each of these phylums does have a response to most the nutrition types with the maximum correlation being between Actinobacteria and the previous day's fiber intake.

The phylum level seems to be too high level for this analysis. At the phylum level these correlations are all relatively close to zero. In order to increase the value of this analysis one option would be to look at the data on a more granular scale such as genus or family. This could allow the truly positive and negative correlations to show themselves without having to be held down by the other members of their phylum. Another way to potentially find more interesting results could be to introduce a lag in the system. Perhaps it takes a few days before the effect of the nutrition can really be seen. This would cause the correlations to appear as though nothing is happening. By introducing a lag one could attempt to combat this effect.

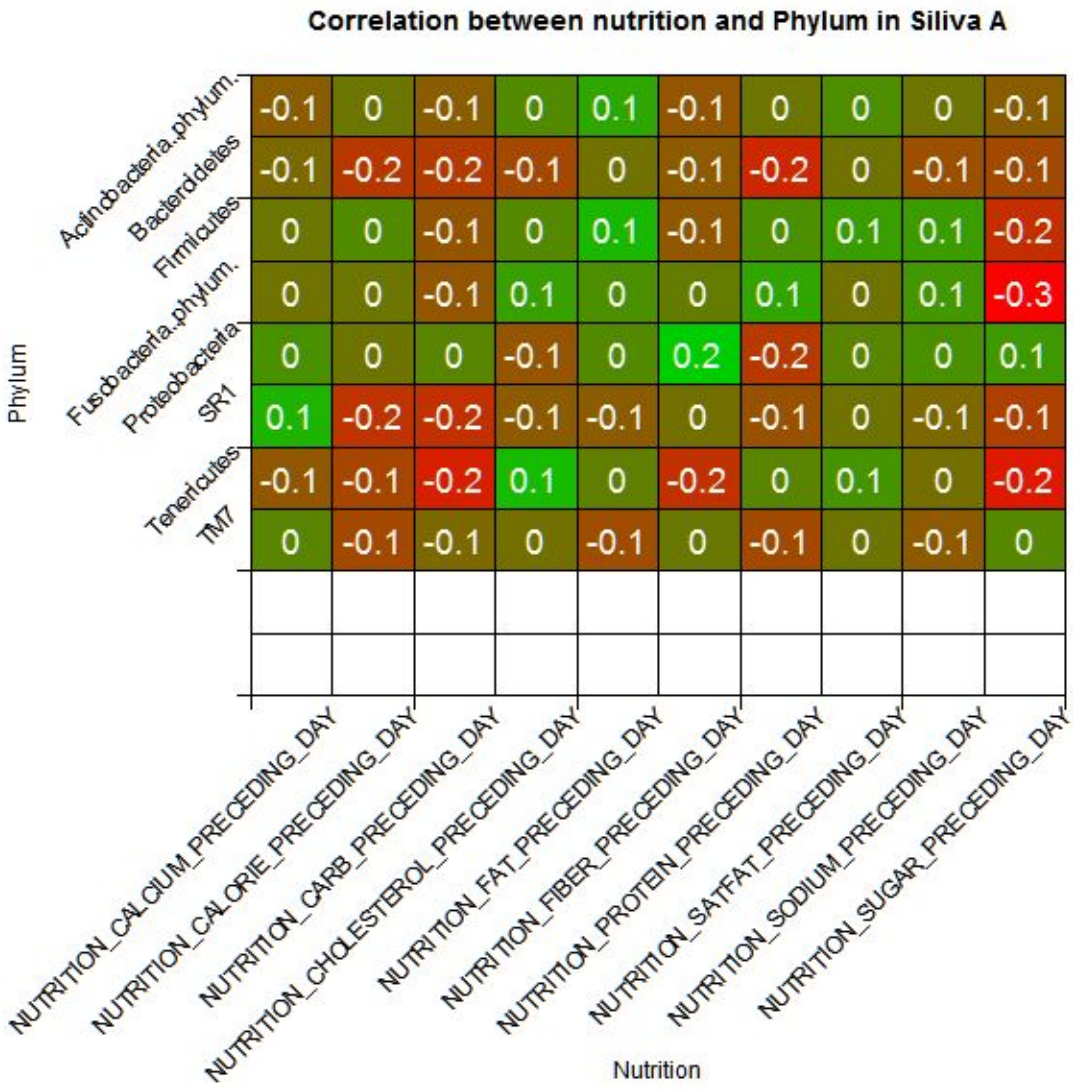

Figure 14. Correlation between nutrition and phylum in patient A's saliva.

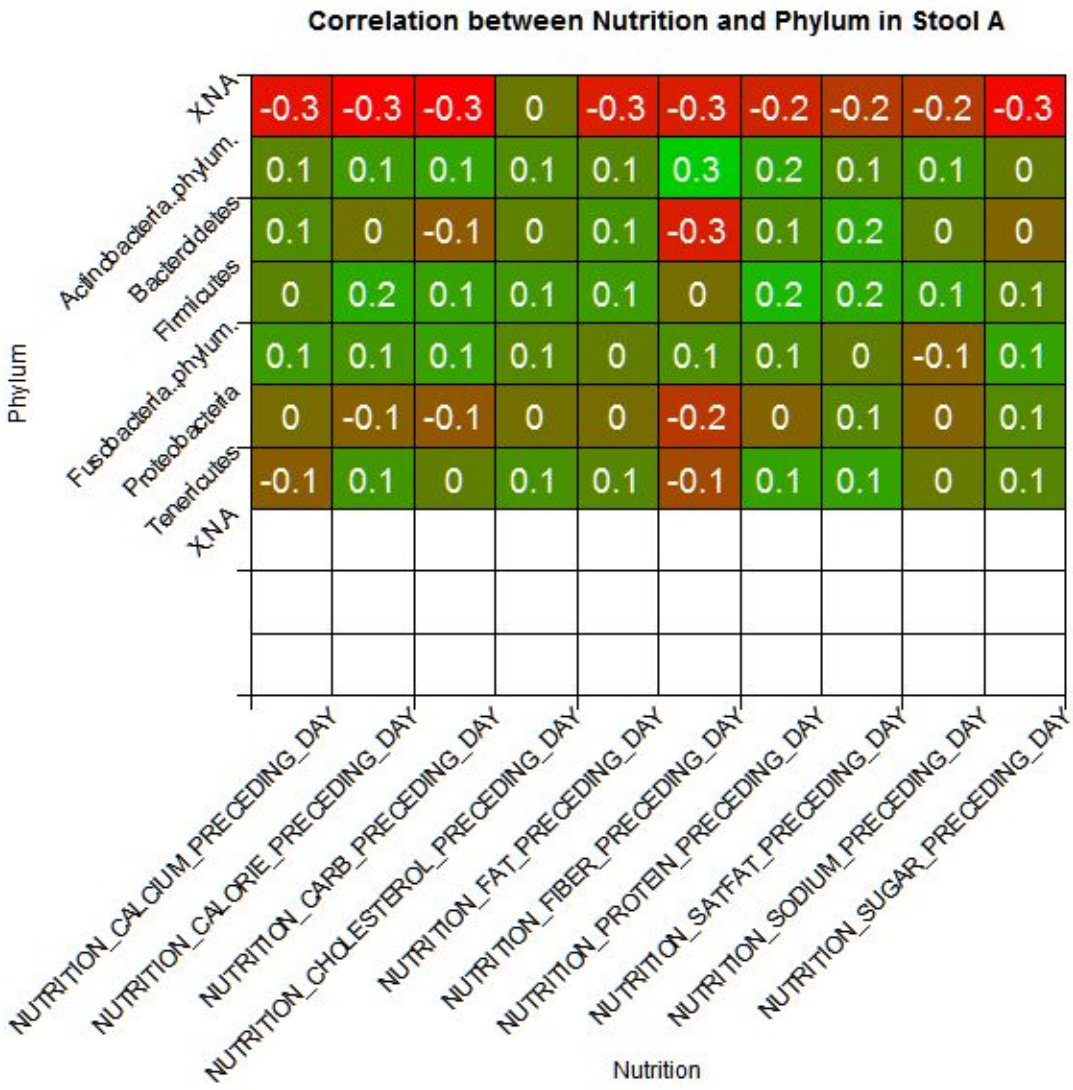

Figure 15. Correlation between nutrition and phylum in patient A's stool.

Conclusion

In conclusion, we were able to visualize the data from the study in various methods. While our biology knowledge limited, we were still able to draw meaningful conclusions from the data. It is easy to see that certain life events such as travel can cause drastic changes in the microbiota of any person. It is clear that there is a correlation between different phyla in both the saliva and stool of the patient tested in this study. It is also seen that there is a correlation between nutritional intake and certain organisms. The analysis of nutritional correlations in this study was likely to high level to see the deeper trends, but there are a few indicators that they are present. For future work, the fft analysis that was starting to be explored in the R code would be fleshed out to be able to draw conclusions. Overall this project helped the team to understand some biological topics as well as develop a deeper understanding of R as a scripting language.

References

[1] David, Lawrence A., Arne C. Materna, Jonathan Friedman, Maria I. Campos-Baptista, Matthew C. Blackburn, Allison Perrotta, Susan E. Erdman, and Eric J. Alm. "Host Lifestyle Affects Human Microbiota on Daily Timescales." Genome Biology 15.7 (2014). Print

[2] Sw1. "Sw1/Bioinformatics." GitHub. 02 Mar. 2017. Web. 22 Mar. 2017.

[3] Cock, P. J. A., C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice. "The Sanger FASTQ File Format for Sequences with Quality Scores, and the Solexa/Illumina FASTQ Variants." Nucleic Acids Research 38, no. 6 (12, 2009): 1767-771. doi:10.1093/nar/gkp1137.