BioE131/231
Layer 2 (Michael Fernandez, Karinna Vivanco, Debbie Pao, Jinho Ko, Katherine Bigelow)
11/28/2018

Checkpoint 2: MG-RAST Summary and Project Proposal

MG-RAST Summary:

In order to analyze the diversity and genomic content of our sample, our assembled metagenome was processed and annotated using MG-RAST. The sample our group received was FWB-306-03, which was taken from soil at a depth of 5-8 feet. Sample reads were directly submitted to One Codex for an initial evaluation of metagenome diversity. Due to memory constraints, we reduced our metagenome data to one gigabyte of memory before assembling it using SPAdes. The resulting contigs were then directly submitted to MG-RAST for analysis.

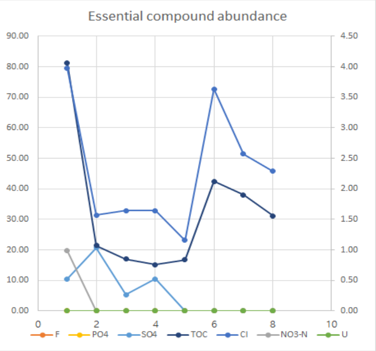
MG-RAST noted a 0% failure of quality control in our submission, which means that all of assembly was included in the analysis. The majority of our sequences (78.26%) was found to feature protein, while the remainder (21.74%) contained unknown sequences. Additionally, the featured protein included a distribution of 0.09% rRNA, 40.39% annotated protein,and 59.52% unknown protein. This finding suggests that MG-RAST is predicting the protein-coding regions without a reference as to what those proteins are. According to the sequence analytics, the contigs also seem to be relatively short with an average a length of 200 bp but were still sufficient for quality control.

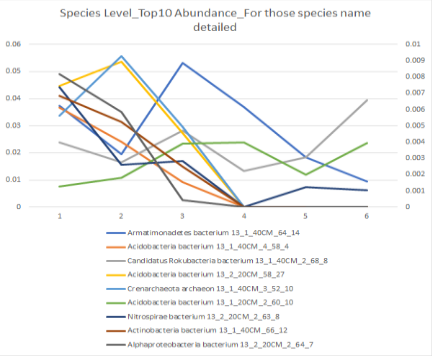
The databases with the most amount of hits for our system were RefSeq, TrEMBL, and GenBank, likely due to the size of those databases. Our samples' proteins then were categorized according to several different orthological categorizations. First, when categorized by COG (Clusters of Orthologous Group (ie proteins that "share a high level of sequence similarity")), a sizeable percentage fell under metabolism (43.06%), followed by cellular processing and signalling. A sizeable amount (18.30%) was still poorly categorized. However, when categorized by NOG (Nonsupervised Orthologous Group), the vast majority of the proteins were "Poorly Characterized," with less falling under metabolism, cellular processes/signalling, and information storage/processing. When characterized via KEGG Orthology (based off of the Kyoto Encyclopedia of Genes and Genomes), a solid 60.63% of the proteins categorized were for metabolism. Somewhat surprising for a soil sample at our sample's depth was that 1.41% of the samples were categorized under "human disease." However, this could be simply due to the fact that many disease-causing bacteria did not necessarily evolve as pathogens, but simply happen to have a negative impact on human health.

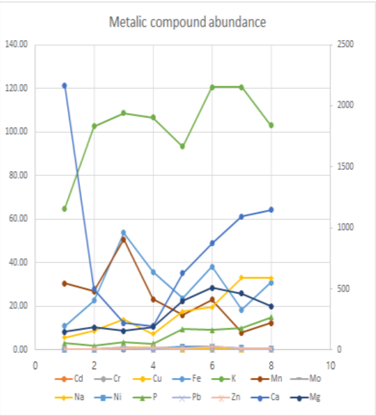
The most noteworthy subsystems in our sample consist of, ordered from the largest percentage to smallest: carbohydrates (15.35%), clustering-based subsystems (13.44%), amino acids and derivatives (11.36%), and protein metabolism (7.04%). Our sample's domain was analyzed to include mainly bacteria of 95.32% and then 4.05% of archaea. For the phylum, our sample had more of an even spread of Proteobacteria (27.41%), Chloroflexi (23.24%), Acidobacteria (18.45%), and Actinobacteria (9.59%). For the class, the most noteworthy in our sample were: Ktedonobacteria (23.01%), Actinobacteria (11.04%), and Alphaproteobacteria (9.78%). For the order, the most noteworthy were: Ktedonobacterales (24.83%), Actinomycetales (10.46%), and Solibacterales (6.49%). For the family, the most noteworthy were: Ktedonobacteraceae (27.08%) and Solibacteraceae (7.08%). For the genus, the most noteworthy were: Ktedonobacter (28.41%) and Candidatus Solibacter (7.43%). Thus, it is apparent that our sample is dominated by two different bacteria: Ktedonobacter is the most common bacteria in our sample, as seen from the data, and the next most common bacteria is Candidatus Solibacter. Ktedonobacter is known to be the largest prokaryotic genome.

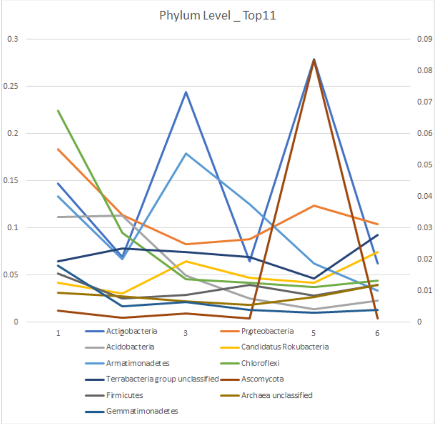
Project Proposal:

The goal of our project to to compare sediment data with specisis and phylum data in order to try to understand the differences in species abundance as the depth increases. We are going to use this data to identify trends and back those trends up with species specific research.









In these figures we are displaying sediment analysis overlaid with specisis and phylum population. As you can see somes species of bacteria spike at certain depths but then quickly lower in abundance at other levels. This type of change from surface level depths such as 1 or to deeper depths such as 5 or 6 demonstrate that depth adversely affects the populations of bacteria. An example of this is where the phylum armatimonadetes peak at level 3 while manganese also peaks at level three. Also for the remainder of the graph the manganese levels continue to fall while armatimonadetes abundance continue to fall as well. This trend in sediment data to bacteria species population points to a trend between armatimonadetes and manganese abundance. According to “Manganese Homeostasis and utilization in pathogenic bacteria” by Dr. Juttukonda, manganese is an important cofactor for many enzymes that make life possible. This may be the reason why armatimonadetes is more abundant at that level, in order to take advantage of this vital resource. This type of analysis is what we will focus on through our project. We will attempt to draw correlations between soil composition and bacteria abundance.

Reference: Juttukonda, L. J., & Skaar, E. P. (2015). Manganese homeostasis and utilization in pathogenic bacteria. Molecular microbiology, 97(2), 216-28.