Bioe 131/231
Group 3: Michael Fernandez, Karinna Vivanco, Debbie Pao, Jinho Ko, Katherine Bigelow
12/7/18

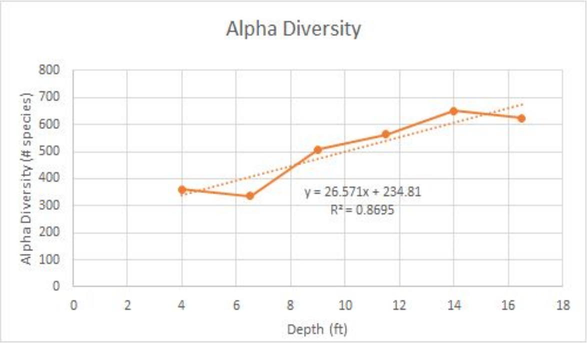
An analysis of Population Diversity and Genetic Variation

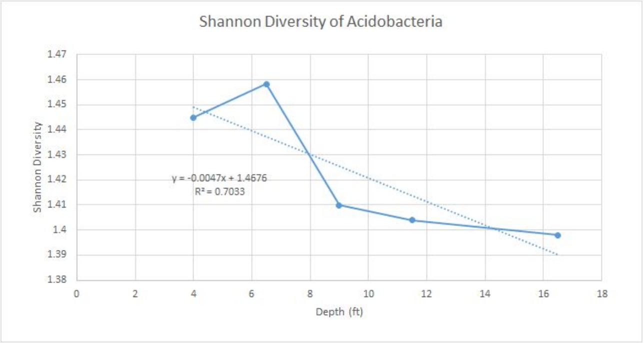
In our analysis, we explored the connection between diversity and depth of sample, where diversity is defined as species abundance and genetic variation. Our hypothesis is that higher diversity in the overall number of species in the population will be correlated with higher genetic variation per species of Acidobacteria. We chose to focus on the phylum Acidobacteria because it is one of the most abundant and diverse phyla on earth, but is difficult to study. On average, Acidobacteria make up 20% of microbial communities in all soil environments but can also be as high as 52%. However, out of the 26 subdivisions of Acidobacteria, only eight have been successfully cultured. Acidobacteria has also been known to play a role in soil recovery, aiding nutrient cycling and promoting plant growth, and prefers low pH environments. We believed choosing this phylum would give us interesting results.

The metrics we used for diversity were alpha diversity and shannon index. Alpha diversity is one of the most popular metrics to calculate diversity of a given environment. In equation, it is done by the equation on left. In MG-RAST, alpha diversity was used as a metric of diversity. We first wanted to calculate diversity for Acidobacteria as well, but soon realized that some notation were hard to be drawn. We instead found Shannon diversity, which is the first-order approximation by setting α=1. The trend of diversity doesn't change by approximation, so we could replace alpha diversity with Shannon diversity in terms of getting diversity of Acidobacteria.

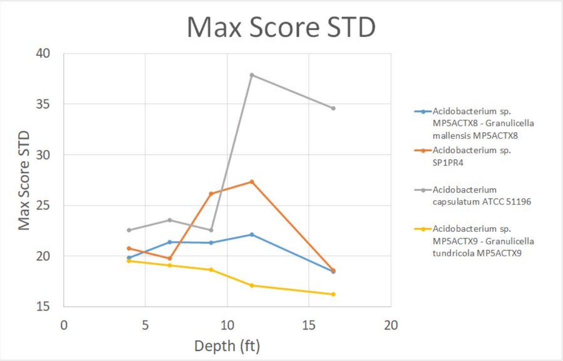
$$\boxed{{}^{q}D_{\alpha} = \frac{1}{\sqrt[q-1]{\sum_{j=1}^{N}\sum_{i=1}^{S} p_{ij} p_{i|j}^{q-1}}}}$$

The purpose of our analysis was to find the diversity of the acidobacteria and compare it with genetic diversity. First we sampled MG-RAST over layer 1,2,3,4, 5 and 6. What we found was that alpha diversity was increasing as depth increased. This data is contradictory to our shannon diversity index we created only sampling our acidobacteria population. However the decrease is not very significant overall. This shannon index can also be converted to effective number of species (ENS) through exp(shannon index). In other words ENS is the number of equally-common species in your population.When we do this for the max and min value we see that the ENS are value for the max value exp(1.46) is 4.30 and for the minimum value the exp(1.4) is 4.05. This is only a 5.8% decrease which means that over the layers the effective number of acidobacteria species were effectively the same. This constancy is most likely due to the average pH of 4.33 across all layers with a standard deviation of 0.22. This is highly acidic.





Next we looked at genetic variation, the subtle differences in DNA within a species. The usual way to measure genetic variation in a population is by relating it through mutation rates. Mutation rates over generations can be used to determine how quickly a population is changing and therefore how different the overall population will be from each other overall. However the samples we were given were not sampled over generations and therefore we needed to create a different metric to get an approximation for genetic variation. In order to do this we first took reads from MG-RAST and separated acidobacteria. The reads had already been assigned to reference genome in the ncbi database. We then used these tags to run reads for each species under each layer to the reference genome they had originally been assigned to in order to see how well they matched the reference genome they had already been assigned to. Each read was matched to some part of the reference genome and was assigned a score by BLAST. With all those scores we took a standard deviation, which told us how varied these scores were in general. By doing this we are determining how different the reads are. If the standard deviation is high then we know the reads themselves are very different, and therefore the the genetic variation is high, vice versa. We understand that this is a crude metric but given the restrictions on our sampling data to 1 generation we were unable to use the usual mutation rate determination of genetic variation. With this metric were able to run 4 different species.



As you can see in this graph the Acidobacterium capsulatum ATCC 51196 spikes up very harshly from a depth of 9ft to 11ft. while the Acidobacterium sp. MP5ACTX9 and Acidobacterium sp. MP5ACTX8 remain flat. Also Acidobacterium sp. Sp1PR4 spiked slightly from depth 7 to 11ft. The reason we think genetic variation may have increased in Acidobacterium capsulatum ATCC 51196 compared to the other strains is because of the increase in diversity of organisms overall. For instance if there's a higher diversity of microbes in a population then Acidobacterium capsulatum ATCC 51196 exists in an environment where more species are represented and therefore more DNA is present. Microbes are constantly taking up DNA from the outside environment and having a larger diversity of other microbes in the soil would allow Acidobacterium capsulatum ATCC 51196 to take up more different DNA. Also given a higher diversity there is a higher effective number of species in the environment also means their is a higher population overall most likely. And given this higher population there very well may be more resources and less pressure on any microbe to adapt. Therefore higher population higher variation in the population as a whole, without a small niche it would otherwise have to fill. However the other species in our population did not exhibit this same trend so further work would need to be done to validate these claims. The current issue is that many of these microbes are unable to be cultured in the lab and therefore without controlled experiments, analysis on a sequence level can only point to places for further investigation.

Some of the issues we ran into during our analysis include memory limitation and defining our metrics. During assembly, we needed to cut down our metagenome sample data to 1gb before being able to use SPAdes due to memory constraints. Additionally, we first wanted to find the alpha diversity using MG-RAST for only Acidobacteria species as a measure of species abundance. However, it was difficult to format the Acidobacteria reads output by MG-RAST into a format that could be re-submitted into MG-RAST. Thus, we were only able to to find the alpha diversity of all species in the sample. We also first wanted to use an SP score as a measure of genetic variation, but had difficulty using MUSCLE to align species-specific reads. This might have occured because our reads were too dissimilar. We instead developed our own metrics to account for species-level and genetic-level diversity.

Currently, our analysis mainly consists of both the MG-RAST data and the statistics which BLAST gives us when we map some of our reads to a reference genome. In the future, we could try assembling more of a genome from our reads and aligning it as a whole to a reference genome and analyze the percent difference between the two; we could also conduct more literature search both on the mathematical basis of our analysis and on genetic diversity metrics. In addition, as our chosen phylum to study is acidobacteria (which tend to thrive in low pH conditions), it could be interesting to analyze how our diversity metrics vary with soil pH and other environmental factors.