Mano Ranaweera

BF 550 Final Project

Gut microbiomes are important for animal health and can be negatively impacted by certain effects, such as antibiotic treatment and its role in proliferation of pathogens. Prior studies have shown that lower gut microbial diversity is caused just a few days after consumption of antibiotics, and that overuse of antibiotics permanently change our microbiomes, resulting in higher rates of obesity, asthma, diabetes, and even some forms of cancer. The motivation of this study is to evaluate the effects of antibiotic exposure on the size and composition of the honeybee gut communities1. Methods for this included monitoring the survivorship of bees, and determining susceptibility to infection by opportunistic pathogens following antibiotic treatment.

The type of data used started with 16S rRNA gene reads as raw data, which was then processed to generate a counts matrix used for analysis. The counts matrix measured the number of reads of each bacterial organism for each sample(bee), followed by a relative abundance matrix with the percentage representing the counts. Analyses done with the counts matrix include computing alpha diversity of samples and principal coordinate analysis(PCoA) that visualize gut community compositions of control and treatment bees using weighted and unweighted unifrac. Figures 3a(Alpha Diversity), 3c(PCoA- Weighted UniFrac), and 3d(PCoA- Unweighted UniFrac) were replicated from the study for this project.

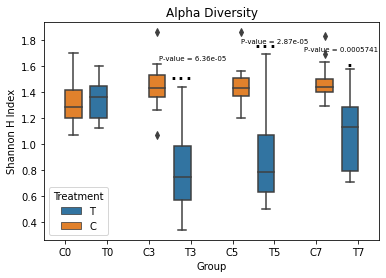
**Methods**

Alpha diversity is the average diversity of species in different sites within one sample2. Shannon’s H index was the parameter used to calculate this. To generate the alpha diversity plot, the counts data of each treatment and control sample’s microbiome were used as raw data in a data frame where the rows were the samples and the columns were the bacterial organisms. An R script, alphaDiv.R, was written to read in the data frame, rarefy the total counts of each sample as the paper did to 5000 total sequence reads, and finally calculate each sample’s alpha diversity. The python skbio package used for calculations like rarefaction of total reads are incompatible with a PC device, which is why this process was done using R. With the alpha diversity values for each sample calculated, the resulting data frame was transferred to a python notebook for plotting. In the python script, “Shannon Index.ipynb”, the data frame was imported via pandas and used to generate a boxplot of the alpha diversity in seaborn, grouped by the control and treatment samples at each timepoint of 0, 3, 5, and 7 days. The control and treatment groups were separated by color, and each timepoint showed that pair of box and whisker plots. In the study, a significance test was performed for each timepoint between the control and treatment group. To see the significance in alpha diversity of each control and treatment group at each timepoint, a mann-whitneyu test in python was performed for each group to generate a p-value. Timepoints showing three dots indicate a p-value less than 0.0001, indicating high significance. One dot indicated a p-value less than 0.05, which is still significant. P-values were added in text to the plot as well in each timepoint.

Principal coordinate analysis(PCoA), also known as Metric Multidimensional Scaling(metric MDS), is defined as a statistical method to convert distance-based data between samples into a map-based visualization3. PCoA plots were generated using R scripts(“Weight\_pcoa\_processing.R” and “Unweight\_pcoa\_processing.R”) as well for weighted and unweighted unifrac data. Just like the alpha diversity calculation, the skbio package in python is also used for this calculation, so that is why this was also done using R. Weighted and unweighted unifrac distance matrices were the data frames that were loaded in the R scripts to produce the PCoA coordinates for the first two dimensions. Columns were added to the initial data frame for the timepoint(day) and group(control or treatment) to group by for the plot. Variances for the first two dimensions were also calculated, and put into a separated data frame. Both data frames were then brought into a python jupyter notebook, in file “PCOA plot.ipynb”. The PCoA data was plotted, with the groups having different colors and the timepoints having different shapes. Percent variance explained for each of the two dimensions were stated on the x and y axes, as was also done in the study.

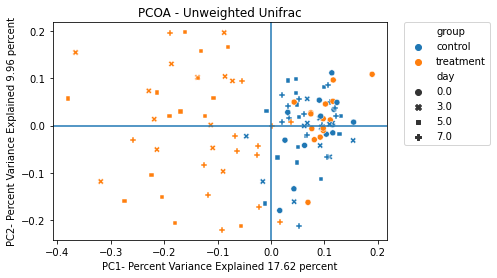
**Results**

Visualization of the alpha diversity of each bee’s microbiome is visualized in Figure 1 below. The Shannon H Index scale is different from the paper’s, meaning the calculated values are different, but the trends visually look very similar when looking at the difference between each treatment group at each timepoint. Levels of significance between the groups, which the dots indicate, are also similar, except in day 7 where the p-value is 0.00057. This p-value is not less than 0.0001, but it was in the study.

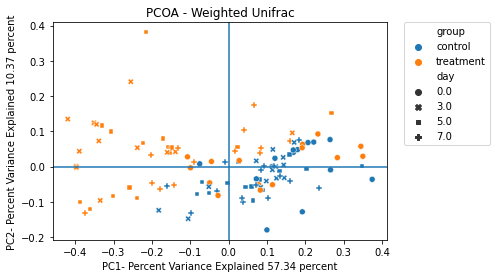


**Figure 1. Difference in alpha diversity(Shannon H Index) between control and treatment bees at each time point.**

The principal coordinate analysis plots seen in figures 2 and 3 below have been replicated from the paper almost identically in terms of the generated PC data coordinates, with data coming from unweighted and weighted UniFrac. Percent variance explained for each dimension are properly calculated by the R scripts, as seen on the x and y axes. These values match the values calculated in the study.



**Figure 2. Principal coordinate analysis using unweighted UniFrac.**

****

**Figure 3. Principal coordinate analysis using weighted UniFrac.**

**Conclusion**

Overall, the purpose of this project was to replicate the alpha diversity and PCOA plots, and master the programming skills to generate these results for any microbiome analysis. There are some differences to address between the replicated plots, and what was produced with the paper.

For the alpha diversity plot in figure 1, the y-axis scale is different. When using the abundance matrix of each sample, the study stated that the counts values were adjusted based on rarefaction, with sequencing depth of 5000 reads per sample. When doing the rarefying with 5000 total reads per sample, computed Shannon index values for alpha diversity still came out differently from the study. From this point, it was unclear what else was done to process the data, which is what ultimately caused the scale to look different as the replicated plot only ranges from 0 to 2. The box and whisker plot ranges with the alpha diversity values were also different, believed to be for the same reason as the other differences. Trends, however, were very similar when looking visually at the study’s figure and this figure 1. In the study, the plot had p-values from the Wilcoxon test being under 0.001 for timepoints at days 3, 5, and 7. Computed p-values done in the python script using the mannwhitneyu function resulted in only timepoints at days 3 and 5 having p-values at less than 0.0001, as denoted by the three dots seen. Having one dot at day 7 was different from the study, which still had a p-value of less than 0.0001, also probably caused by the lack of clarity in processing this data in the study. Ultimately, the code from the R and python scripts for this plot show a fundamental understanding of producing alpha diversity data along with proper data visualization.

PCoA plots were precisely reproduced in terms of the actual coordinate data that was generated, including by the proper corresponding group and timepoint symbolized. Percent variances were calculated properly as well. The only differences seen were in plotting aesthetics.

**Citations**

1. Raymann, K., Shaffer, Z., & Moran, N. A. (2017). Antibiotic exposure perturbs the gut microbiota and elevates mortality in honeybees. *PLOS Biology*, *15*(3). https://doi.org/10.1371/journal.pbio.2001861
2. *The use and types of alpha-diversity metrics in microbial NGS*. The Use and Types of Alpha-Diversity Metrics in Microbial NGS - CD Genomics. (n.d.). Retrieved December 10, 2022, from https://www.cd-genomics.com/microbioseq/the-use-and-types-of-alpha-diversity-metrics-in-microbial-ngs.html#:~:text=Alpha%20diversity%20(%CE%B1%2Ddiversity),diversity%20(%CE%B3%2Ddiversity).
3. Korstanje, J. (2021, July 9). *Principal coordinates analysis*. Medium. Retrieved December 10, 2022, from https://towardsdatascience.com/principal-coordinates-analysis-cc9a572ce6c