Project 2: Final Report **Merai Dandouch**

Introduction

Although an ongoing debate continues on whether microbiota of different species can generalize the human microbiota, the trends and patterns in the microbial biome of Baboons remain significant to help support findings in human microbiota composition. In a paper by Miller et al., 2017, scientists studied the vaginal microbiota of baboons across various reproductive states during specific ovarian cycles to better understand the underlying patterns of microbiomes where lactobacilli do not dominate. Fifty-two vaginal swabs were taken from 48 wild baboons, where 4 individuals were sampled twice. Baboon subject samples were studied in the Amboseli ecosystem of Kenya. Microbial analyses of the swabs were completed by amplification of the V4 regions, followed by OTU identification, and alpha and beta diversity predictions. Samples were taken from baboons that were recently pregnant (P), in postpartum amenorrhea (PPM), undergoing ovarian cycling (C), or miscarrying (M). The ovarian cyclic baboons were placed into 4 subcategories containing anestrus (A), swelling (S), peri-ovulating (O), and deturgescent (D). The significance of this paper is to replicate the analytical methods in Miller et al., 2017 and report findings, limitations, and results. Statistical analysis was completed using Python 3.9.7 on Jupyter IDE and all code can be found at https://github.com/meraidandouch.

Methods

To start off, bacterial DNA was extracted from swab samples using PowerSoil DNA Isolation kit with modifications to account for the cotton. To avoid sequencing human host DNA, the V4 region of the 16s rRNA gene, both a conserved and variable region of bacterial DNA, was amplified and barcoded using PCR. The samples were then pooled together and sent to Illumnia HiSeq2000, where the pooled DNA was sequenced in technical triplicates across three lanes of the flow cell. After obtaining sequence reads, researchers discarded sequences with Phred quality scores less than or equal to 25 using PRINSEQ lite v.0.20.4. Subsequently, the first round of OTU-picking and clustering was filtered using USEARCH v.7.0 with 2% sequencing depth and clustered using the UPARSE algorithm. The matched OTUs were searched against QIIME and assigned a taxonomic identity and underwent multiple rounds of identification using Greengenes and PyNast. To control for differences in sequencing depth between samples, the authors normalized read counts using cumulative sum scaling (CSS) in the metagenomeSeq R package. The normalized OTU table was used to calculate beta diversity and for differential relative abundance (alpha diversity).

During data organization, I created a subset of three datasets from the original OTU counts data file to categorize counts by phylum called **countmi**, count of reads per sample called **otu_rich**, proportion of reads or relative abundance called **dfmi**. The relative abundance dataset was used to calculate alpha diversity to measure diversity within a specific habitat. In contrast, beta diversity which accounts for the differences between multiple habitats, was computed using Bray-Curtis Dissimilarity and a module from scipy library called scipy.spatial.distances where tools such as pdist and squareform were used to calculate PCOA.

OTU raw read file was normalized using cumulative sum scaling. After obtaining the cumulative sum of each sample, it was then divided by the original OTU counts data. Proportion of reads were calculated by summing the reads of count data by phyla group and dividing by the total. I used the OTU table to obtain the differential relative abundance numbers for each sample. I put each row of the OTU table in a dictionary with the keys are phyla groups and the list of read counts for each sample as the value. I then iterated over the dictionary and summed the columns to total sum read counts for each phyla if the keys were not unique. I converted the dictionary into a Pandas Dataframe and then transposed the matrix so that all samples (V1, V2, …V53) were on rows and each phyla were on the columns. I multi-indexed the dataframe to package the metadata. The index information included reproductive state (pregnant(P), postpartum amenorrhea (PPA), miscarrying (MC)) and ovarian cycling (anestrus(A), swelling(S), periovulation(O), and deturgescence(D)) states. I then filtered the data down to phyla that have a phyla abundance mean greater than 1%. In contrast, phyla abudance less than 1% were grouped as low abundance. The structure of my code can be followed as OTU richness matrix gathering, stacked bar chart plot of relative abundance, Shannon's index calculations, alpha diversity box plot graphs, brays-curtis dissimilarity matrix calculations, beta diversity PCoA plots.

Results

Analyses on alpha and beta diversity were calculated using the cumulative sum scaling read counts matrix. As shown below, the relative abundances are nearly identical to that of the authors and the differential relative abundances are consistent. Only phyla with a relative abundance mean >1% are shown, and phyla with a relative abundance mean <1% are grouped as "low abundance". This rule did not apply to Archaea kingdom, unclassified groups and unclassified bacteria denoted as "k_Bacteria", "k_Archaea", and "Unclassified". Similarities between the abundances are that Firmicutes, and Fusobacteria bacteria dominated in peri-ovulating samples and deturgescence samples. Additionally, Proteobacteria are dominant in miscarrying samples on the right. The authors excluded unclassified bacteria, labeled as "k_Bacteria", but I chose to include it in my analysis.

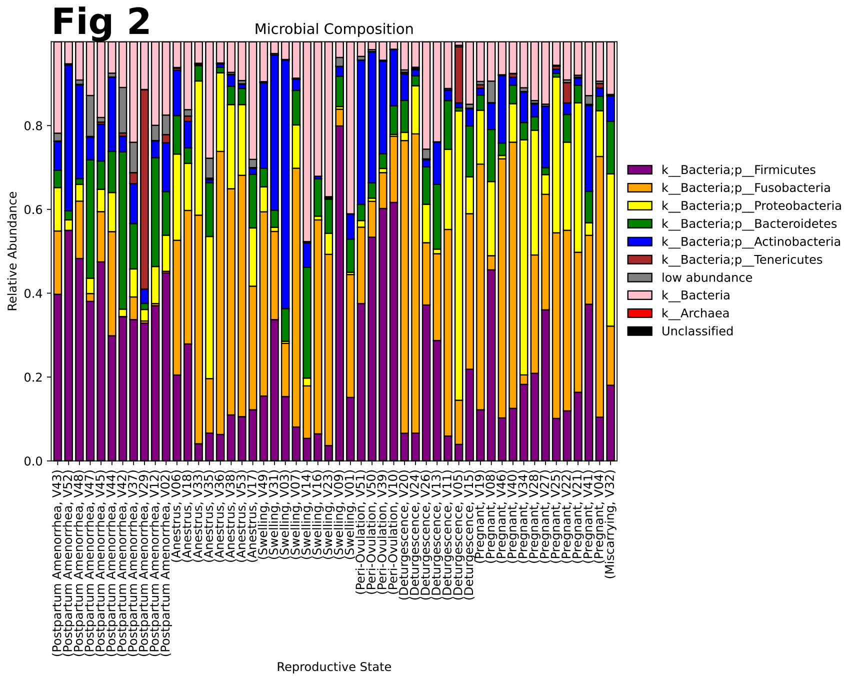

Figure 2: Stacked barplot showcasing species abundance across all phyla (Firmicutes, Fusobacteria, Proteobacteria, Bacteroidetes, Actinobacteria, Tenericutes, low abundance, k_Bacteria, k_Archaea, Unclassified) in the data. Baboon subjects in Postpartum Amenorhhea and Peri-ovulation reproductive state displayed a greater diverisity of spceis abudnace. Meanwhile, baboon subjects in the Anestrus, Deturgescence, and Pregnancy state were dominated by Fusobacteria.

Authors conducted statistical analyses for alpha diversity, including a multivariate linear regression model for OTU richness and Shannon's diversity index. This examined the fixed effect of read count on reproductive state vs ovarian cycles. I worked to plot the variation of the entropy of each sample for different groups such as C, P, and PPA instead. To do this, I iterated over the relative abundance data frame and calculated Shannon's index for each sample. I multiplied the list by the natural log transformation of the list, and then summed up all the values in the list. I then negated this value and raised it to the natural exponent number. I performed this calculation for the remaining samples. I then plotted the Shannon numbers for samples that belonged to ovarian cycling(C), pregnant(P), postpartum amenorrhea (PPA). I then plotted the Shannon numbers for anestrus(A), swelling(S), periovulation(p), and deturgescence(D). I obtained OTU richness by counting the non-zeros in each row for each sample. As shown in the graph below, my Shannon Diversity Index numbers were close to the paper. An observation was that low Shannon numbers around 4.0 matched the paper, but the high Shannon numbers that were greater than 6 differed by a point in my dataset. This could be due to normalizing process in previous steps. Yet, the ovarian cycling (cyan) box plot in fig 3a is between 3 - 4.5 Shannon indices is similar to the author's paper. Additionally, the Shannon index number for postpartum amenorrhea (purple) is found to be higher than all remaining reproductive states such as pregnant samples (orange), and ovarian cycling (cyan). Similar methods and results were obtained for OTU richness shown to the right (Fig 3b).

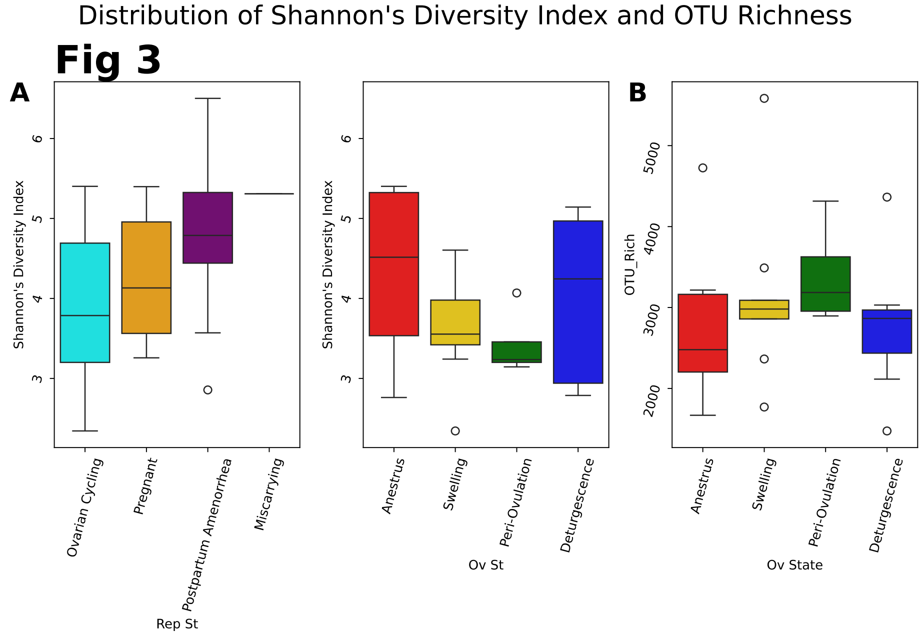

Figure 3A&B: Boxplot diagram displaying distribution of species diversity and quantification using Shannon's Diversity Index(SDI) and OTU richness. (3a&b) SDI for reproductive state is between 3-5.2 units where baboon subjects in Postpartum Amenorhhea displayed a greater quantity of microbiota composition spread than ovarian cycling. Specifically, baboon subjects in the Anestrus cycle displayed higher SDI units at 5.4 units. (3c) While peri-ovulating baboons showed lower diversity, they had the highest OTU richness, indicating a greater number of distinct OTUs.

To study the variation and difference of species composition between varying reproductive and ovarian states, the researchers performed PCoA and PERMANOVAs using Bray-Curtis dissimilarity and weighted UniFrac distance. I omitted weighing UniFrac and calculated PCOA using Brays-Curtis instead. To start this procedure, I used the relative abundance data frame and compared each row (sample) to all 52 samples in the data. I kept doing this until all pair wise distances were computed between all samples. In the end, I obtained square matrix (52 x 52) elucidating the dissimilarity between microbiota in samples. Although I did not obtain a comparable PCOA plot to that of the authors, certain patterns are distinguishable. For example, it can be noted that peri-ovulation and swelling samples are clustered in Fig 4B to the top left corner. This relationship is also seen in the author’s paper. There is a clear differentiation between peri-ovulation (green) samples from the rest of the ovarian cyclic samples.

Fig 4

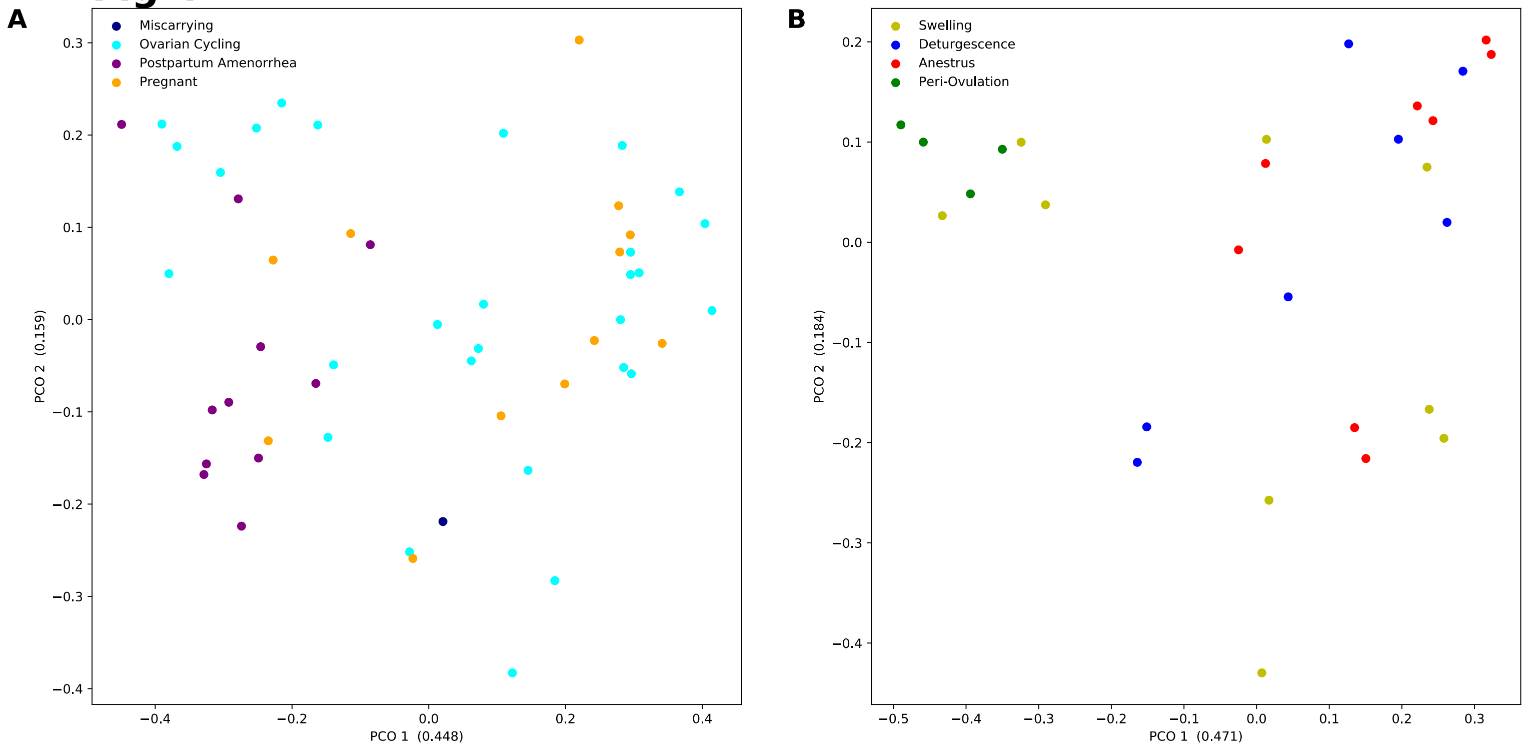

Figure 4: PCoA coordinate plot to showcase beta diversity baboon subjects in varying reproductive and ovarian states. A close cluster can be found in the top left corner in (b) where swelling and peri-ovulation share a similar microbial profile.

The goal of this report was to replicate analytic approaches in an academic journal to get a better understanding of python analyses methods. The authors used a variety of analytical tools such as cumulative sum normalization, relative abundance calculations, entropy (Shannon’s index), and Brays-Curtis dissimilarity matrices. A major significance in data analyses lies in data organization, a crucial step for allowing successful results to occur. While reproducing the author’s statistical approaches, it was found that packages provided by scikit bio were not reliable. In figure 4b&c, mainly Firmicutes bacteria dominated in peri-ovulating samples, resulting in lower diversity compared to the remaining ovarian cycle state samples. In contrast, postpartum amenorrhea baboons had higher species abundance than the remaining reproductive states as shown in figure 2, and figure 3a. In conclusion, differential abundances for each phyla were obtained and low microbiota diversity for peri-ovulating baboons can be seen in Fig 3A and Fig 4B.

Miller, E.A., Livermore, J.A., Alberts, S.C., Tung, J. and Archie, E.A., 2017. Ovarian cycling and reproductive state shape the vaginal microbiota in wild baboons. *Microbiome*, *5*(1), pp.1-14.