Samuel L. Peoples
BBIO 393: Computational Biology
Project 1

Exploratory Analysis of Beta Diversity of Human Gut Microbiome Between Lean and Obese Twins

1. Introduction

This project was completed based on Jeff Gordon's *A Core Gut Microbiome in Obese and Lean Twins* which was used to visualize any noticeable dissimilarity when comparing the Weighted-Unifrac Beta Diversity between Lean-vs-Lean, Obese-vs-Obese, and Lean-vs-Obese Twins. This is completed using the *beta-diversity.py* script from within a *Qiime2* virtual machine. Visualizations were initially created using a Bray-Curtis-Faith metric, but a more accurate result is achieved by introducing the *97_otus.tree* and using a Weighted Unifrac metric.

2. Set-up and use

Parameters:
filepath = path to *study_77_011618-113533* directory

Dependencies (Qiime2 scripts):
beta_diversity.py
single_rarefaction.py
principal_coordinates.py
make_emperor.py

Sample console output (/docs/*subsequent_output.txt):*

```
/usr/bin/python2.7 /home/qiime/Desktop/Project1/biom_read_unifrac.py
/home/qiime/Desktop/Project1/study_77_011618-113533 found.
/home/qiime/Desktop/Project1/figures found.
/home/qiime/Desktop/Project1/stats found.
Rarefaction check completed.

Lean-vs-Lean:
 Low: 0.09970718
 Mean: 0.34773411670368204
 Std: 0.12065306174990073
 High: 0.7940865500000001
 Count: 3422

Obese-vs-Obese:
 Low: 0.0721046
 Mean: 0.37414471930193777
 Std: 0.12564177289788758
 High: 0.90631165
 Count: 34782

Lean-vs-Obese:
 Low: 0.12453261999999997
 Mean: 0.3645712817076045
 Std: 0.12259447284488963
 High: 0.8884746099999999
 Count: 11033

Unifrac distances saved to:
     Lean-vs-Lean:
        /home/qiime/Desktop/Project1/stats/LL_stats_Unifrac.csv
     Obese-vs-Obese:
        /home/qiime/Desktop/Project1/stats/OO_stats_Unifrac.csv
     Lean-vs-Obese:
```

```
        /home/qiime/Desktop/Project1/stats/LO_stats_Unifrac.csv
Boxplot saved to:
     /home/qiime/Desktop/Project1/figures/boxplot_unifrac.png
Distributions saved to:
     Lean-vs-Lean:
         /home/qiime/Desktop/Project1/figures/LL_Distribution_Unifrac.png
     Obese-vs-Obese:
         /home/qiime/Desktop/Project1/figures/OO_Distribution_Unifrac.png
     Lean-vs-Obese:
         /home/qiime/Desktop/Project1/figures/LO_Distribution_Unifrac.png
Principal coordinates saved to:
    /home/qiime/Desktop/Project1/stats/weighted_unifrac_219_otu_table_even1000_pcoa.txt
Principal coordinates saved to:
    /home/qiime/Desktop/Project1/figures/index.html
```

There is an additional *firstRun* zip directory included which can be used to check that all set-up is completed properly.

Sample console output (/docs/*firstRun_output.txt):*

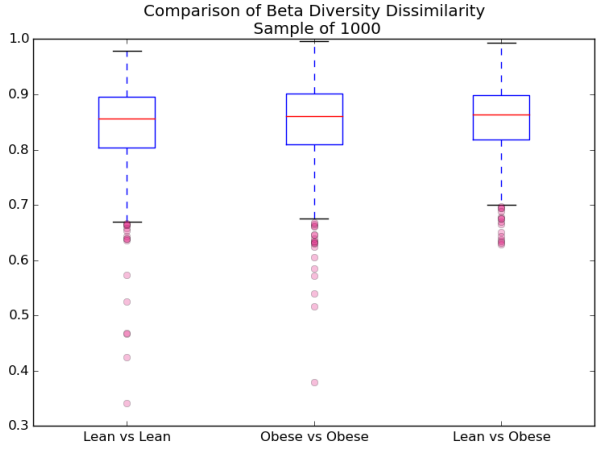
```
/usr/bin/python2.7 /home/qiime/Desktop/Project1/firstRun/biom_read_unifrac.py
/home/qiime/Desktop/Project1/firstRun/study_77_011618-113533 found.
/home/qiime/Desktop/Project1/firstRun/figures created.
/home/qiime/Desktop/Project1/firstRun/stats created.
Rarefaction in progress. Calling:
    single_rarefaction.py
        -i /home/qiime/Desktop/Project1/firstRun/study_77_011618-113533/processed_data/219_otu_table.biom
        -o /home/qiime/Desktop/Project1/firstRun/study_77_011618-
113533/processed_data/219_otu_table_even1000.biom
        -d 1000
Rarefaction check completed.
Dissimilarity Matrix preparation in progress. Calling:
    beta_diversity.py
        -i /home/qiime/Desktop/Project1/firstRun/study_77_011618-
113533/processed_data/219_otu_table_even1000.biom
        -o /home/qiime/Desktop/Project1/firstRun/stats
        -m weighted_unifrac
        -t /home/qiime/Desktop/Project1/firstRun/study_77_011618-113533/97_otus.tree

Lean-vs-Lean:
 Low: 0.09436299
 Mean: 0.3501875217533606
 Std: 0.12033391562884786
 High: 0.78030682
 Count: 3422

Obese-vs-Obese:
 Low: 0.08747150000000002
 Mean: 0.3766034260836065
 Std: 0.12658099158464228
 High: 0.91778851
 Count: 34782

Lean-vs-Obese:
 Low: 0.11719817
 Mean: 0.36671111489803315
 Std: 0.1230288230092367
 High: 0.8923403599999999
 Count: 11033

Unifrac distances saved to:
     Lean-vs-Lean:
        /home/qiime/Desktop/Project1/firstRun/stats/LL_stats_Unifrac.csv
     Obese-vs-Obese:
        /home/qiime/Desktop/Project1/firstRun/stats/OO_stats_Unifrac.csv
     Lean-vs-Obese:
        /home/qiime/Desktop/Project1/firstRun/stats/LO_stats_Unifrac.csv
```

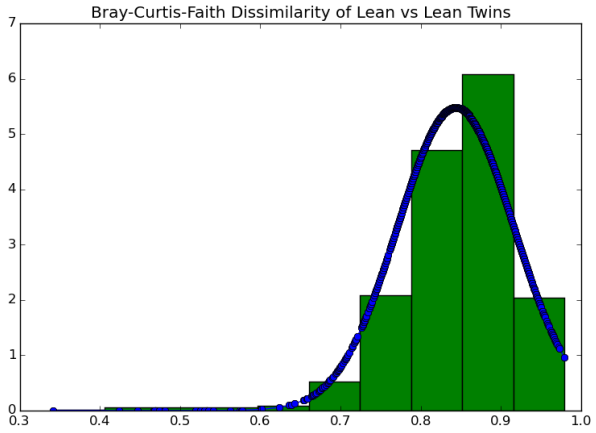
```
Boxplot saved to:
     /home/qiime/Desktop/Project1/firstRun/figures/boxplot_unifrac.png
Distributions saved to:
     Lean-vs-Lean:
          /home/qiime/Desktop/Project1/firstRun/figures/LL_Distribution_Unifrac.png
     Obese-vs-Obese:
          /home/qiime/Desktop/Project1/firstRun/figures/OO_Distribution_Unifrac.png
     Lean-vs-Obese:
          /home/qiime/Desktop/Project1/firstRun/figures/LO_Distribution_Unifrac.png
PCoA in progress. Calling:
    principal_coordinates.py
        -i /home/qiime/Desktop/Project1/firstRun/stats/weighted_unifrac_219_otu_table_even1000.txt
        -o /home/qiime/Desktop/Project1/firstRun/stats/weighted_unifrac_219_otu_table_even1000_pcoa.txt
/usr/local/lib/python2.7/dist-
packages/skbio/stats/ordination/_principal_coordinate_analysis.py:107: RuntimeWarning: The result contains n
egative eigenvalues. Please compare their magnitude with the magnitude of some of the largest positive eigen
values. If the negative ones are smaller, it's probably safe to ignore them, but if they are large in magnit
ude, the results won't be useful. See the Notes section for more details. The smallest eigenvalue is -
0.229622495618 and the largest is 10.2807894039.
Principal coordinates saved to:
    /home/qiime/Desktop/Project1/firstRun/stats/weighted_unifrac_219_otu_table_even1000_pcoa.txt
Emperor visualization in progress. Calling:
    make_emperor.py
        -i /home/qiime/Desktop/Project1/firstRun/stats/weighted_unifrac_219_otu_table_even1000_pcoa.txt
        -m /home/qiime/Desktop/Project1/firstRun/study_77_011618-
113533/mapping_files/2485_mapping_file.txt
         -b obesitycat
         -o /home/qiime/Desktop/Project1/firstRun/figures
Emperor index.html saved to:
    /home/qiime/Desktop/Project1/firstRun/figures/index.html
```

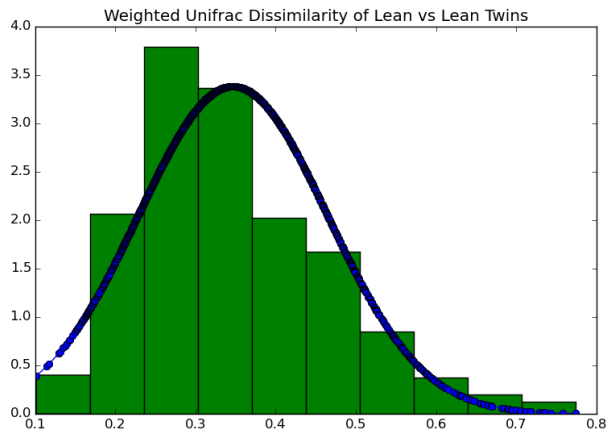
3. Progress, Updates, and Figures

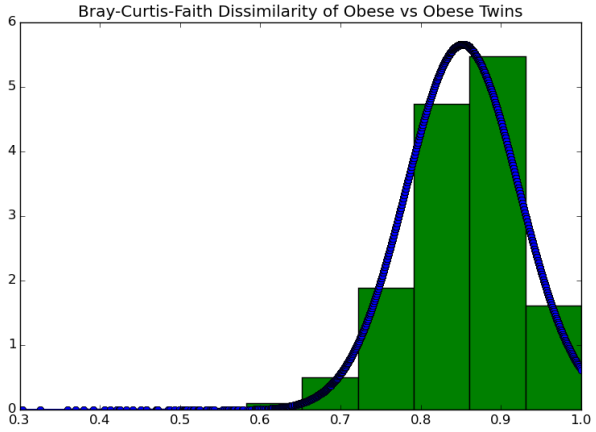
The script has been developed to be implemented in a single-scope, which is something that I plan to address in project 2. I was able to rarify the OTU table using *single_rarefaction.py*, and compute the dissimilarity matrices using *beta_diversity.py*. The boxplots and distributions are saved to an output directory, and displayed below.

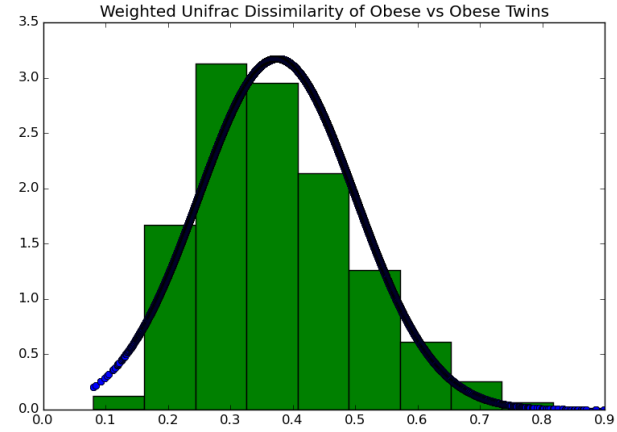
The script also develops an interactive PCoA plot, which displays the distribution of the obesity categories. This is saved to *index.html* and samples are included below.

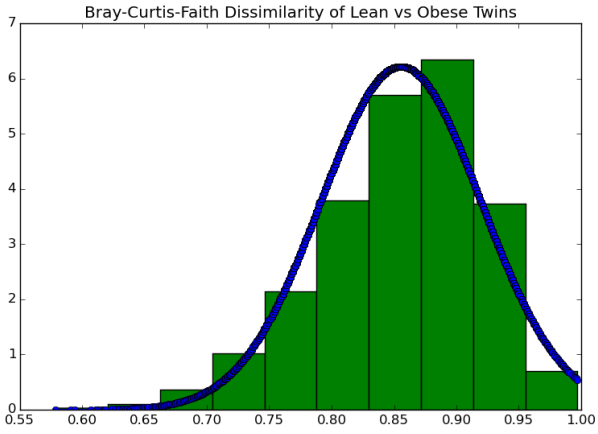


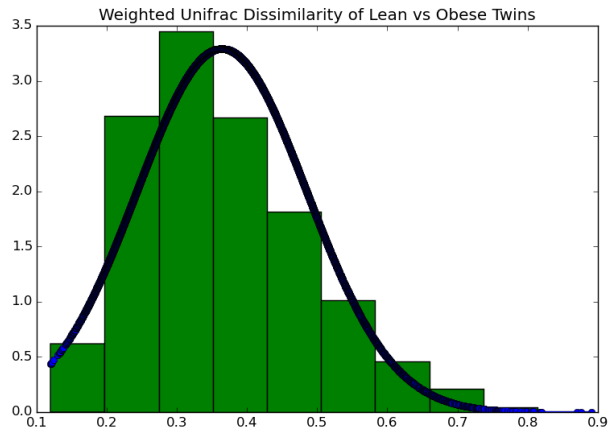

Bray-Curtis-Faith Dissimilarity of Lean vs Lean Twins


Weighted Unifrac Dissimilarity of Lean vs Lean Twins

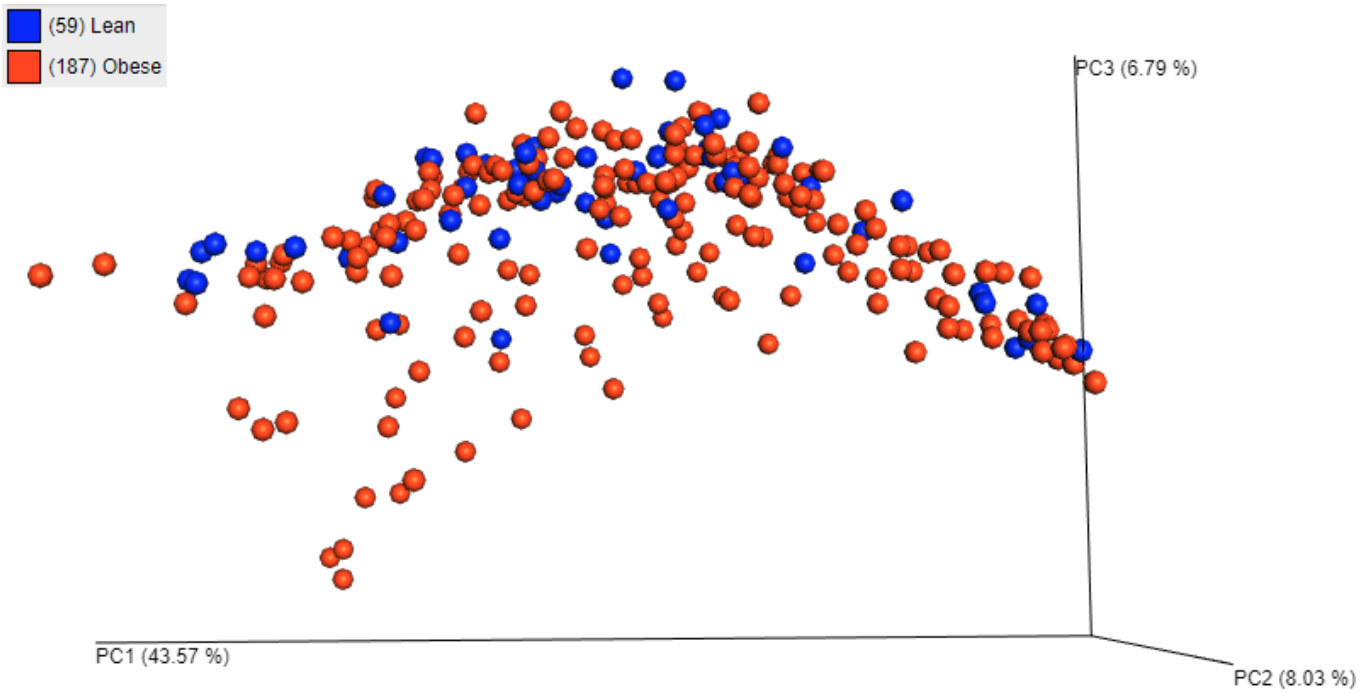

Bray-Curtis-Faith Dissimilarity of Obese vs Obese Twins

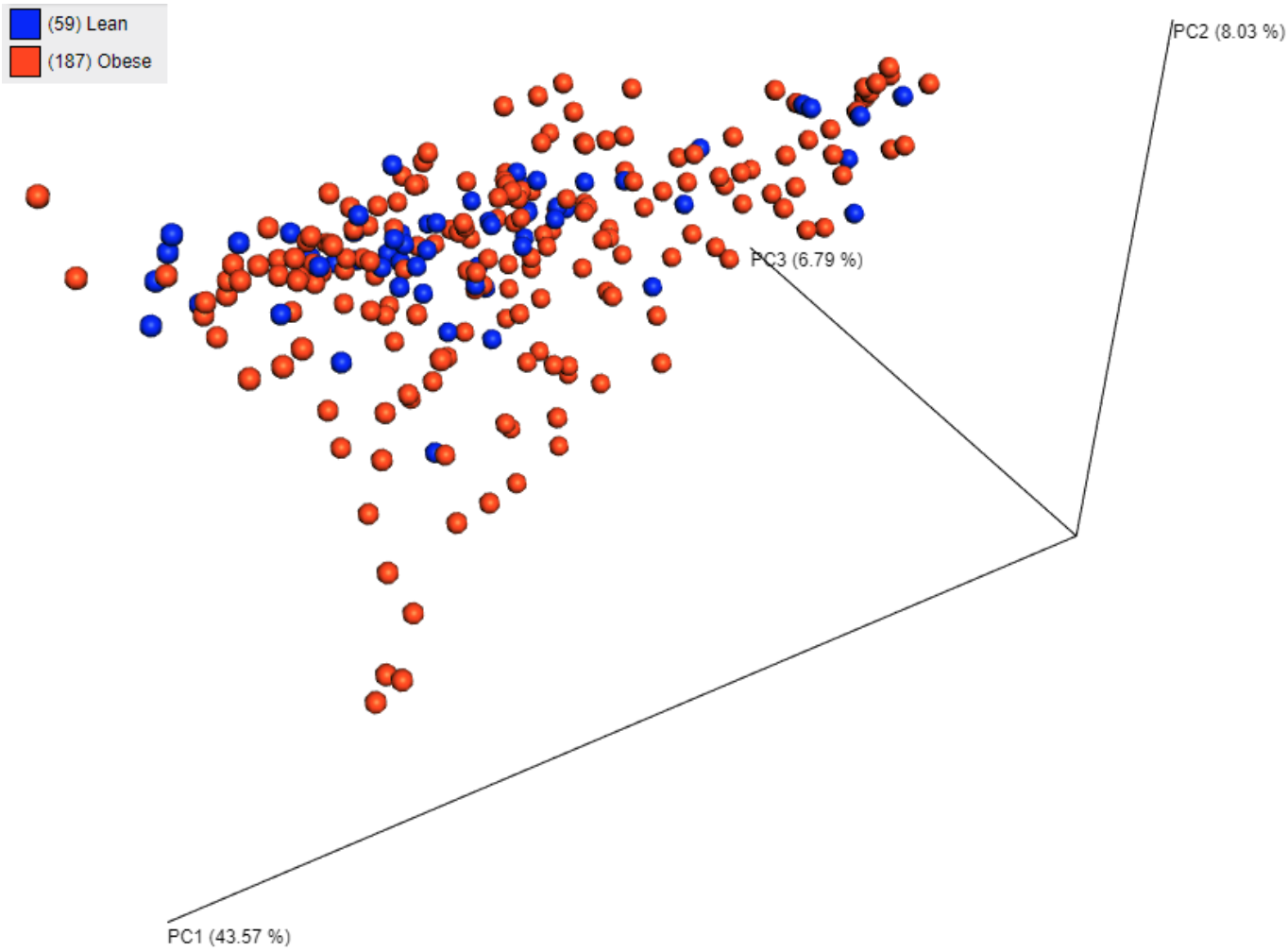

Weighted Unifrac Dissimilarity of Obese vs Obese Twins

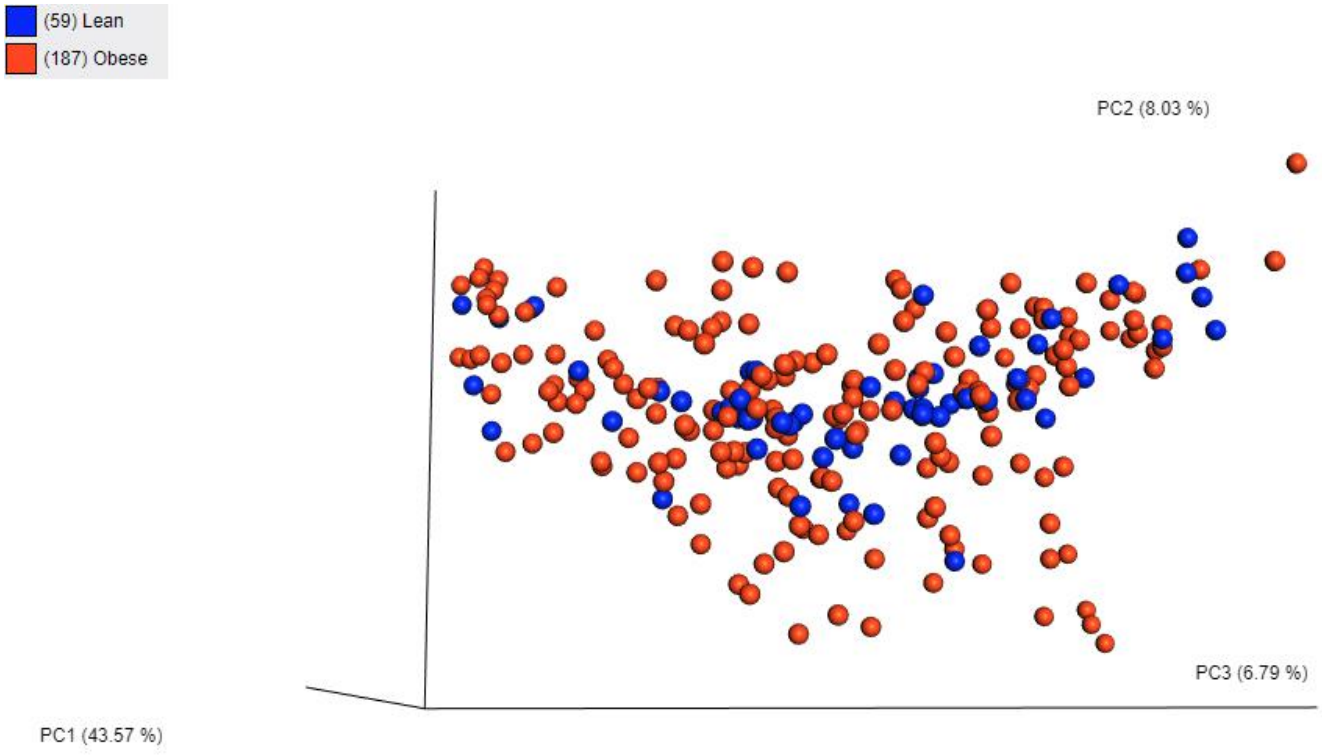

Bray-Curtis-Faith Dissimilarity of Lean vs Obese Twins


Weighted Unifrac Dissimilarity of Lean vs Obese Twins

PCoA Samples:







3. Conclusion

Viewing the figures side-by-side lets us see just how significant the changes are when using the *weighted-unifrac* metric. In project 2, I plan to increase the scope of the script, where I will make more of the options parameterized, because just the *filepath* to the study folder is parameterized at this time. I also plan to explore the significance of the Unifrac distances, and implement a monte-carlo simulation to better sample the data.

The differences that can be seen between the distributions shows the significant changes that occur from the *Weighted-Unifrac* metric, which is much more accurate than the *Bray-Curtis-Faith* metric. In the *Unifrac* boxplot, there are more differences that can be seen other than outliers, which can provide exciting opportunities for the second project. In particular, comparing the beta diversity of Lean-vs-Lean patients shows a mean of .348, while the mean of Obese-vs-Obese and Lean-vs-Obese are .375 and .365 respectively. The middle 50% of the data is aggregated lower for Lean-vs-Lean than the other two categories, and the maximum (excluding outliers) is highest when comparing Obese-vs-Obese patients.

In the PCoA plots, samples labeled obese seem to deviate from the band of blue lean samples. This suggests that the obese samples have more variance in their diversity because the script is developed from the output from the *weighted-unifrac* beta diversity measure. These findings are incredibly useful, and leave a lot to be explored in the subsequent projects.

These differences suggest that there does exist a measurable difference in the beta diversity of the human gut microbiome based on obesity, and the statistical significance of these differences as well as a more detailed analysis can be completed as a part of the second project. Visualizing these differences can be a great way to begin exploratory analysis on the data, and can direct our questioning while simply expressing any trends that may exist.

4. References

Qiime: http://qiime.org/
Qiita: https://qiita.ucsd.edu/
Bray-Curtis-Faith: http://readiab.org/book/0.1.3/3/1#4.1.1
Gut Microbiome Dataset: https://qiita.ucsd.edu/study/description/77
Biom-Format: http://biom-format.org/documentation/biom_format.html
Lozupone/ Knight's UniFrac: http://aem.asm.org/content/71/12/8228.full
Qualitative/ Quantitative: https://www.ncbi.nlm.nih.gov/pubmed/17220268
Weighted UniFrac: https://liorpachter.wordpress.com/2013/09/18/unifrac-revealed/