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):*

1. /usr/bin/python2.7 /home/qiime/Desktop/Project1/biom\_read\_unifrac.py
2. /home/qiime/Desktop/Project1/study\_77\_011618-113533 found.
3. /home/qiime/Desktop/Project1/figures found.
4. /home/qiime/Desktop/Project1/stats found.
5. Rarefaction check completed.
7. Lean-vs-Lean:
8. Low: 0.09970718
9. Mean: 0.34773411670368204
10. Std: 0.12065306174990073
11. High: 0.7940865500000001
12. Count: 3422
14. Obese-vs-Obese:
15. Low: 0.0721046
16. Mean: 0.37414471930193777
17. Std: 0.12564177289788758
18. High: 0.90631165
19. Count: 34782
21. Lean-vs-Obese:
22. Low: 0.12453261999999997
23. Mean: 0.3645712817076045
24. Std: 0.12259447284488963
25. High: 0.8884746099999999
26. Count: 11033
28. Unifrac distances saved to:
29. Lean-vs-Lean:
30. /home/qiime/Desktop/Project1/stats/LL\_stats\_Unifrac.csv
31. Obese-vs-Obese:
32. /home/qiime/Desktop/Project1/stats/OO\_stats\_Unifrac.csv
33. Lean-vs-Obese:
34. /home/qiime/Desktop/Project1/stats/LO\_stats\_Unifrac.csv
35. Boxplot saved to:
36. /home/qiime/Desktop/Project1/figures/boxplot\_unifrac.png
37. Distributions saved to:
38. Lean-vs-Lean:
39. /home/qiime/Desktop/Project1/figures/LL\_Distribution\_Unifrac.png
40. Obese-vs-Obese:
41. /home/qiime/Desktop/Project1/figures/OO\_Distribution\_Unifrac.png
42. Lean-vs-Obese:
43. /home/qiime/Desktop/Project1/figures/LO\_Distribution\_Unifrac.png
44. Principal coordinates saved to:
45. /home/qiime/Desktop/Project1/stats/weighted\_unifrac\_219\_otu\_table\_even1000\_pcoa.txt
46. Principal coordinates saved to:
47. /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.

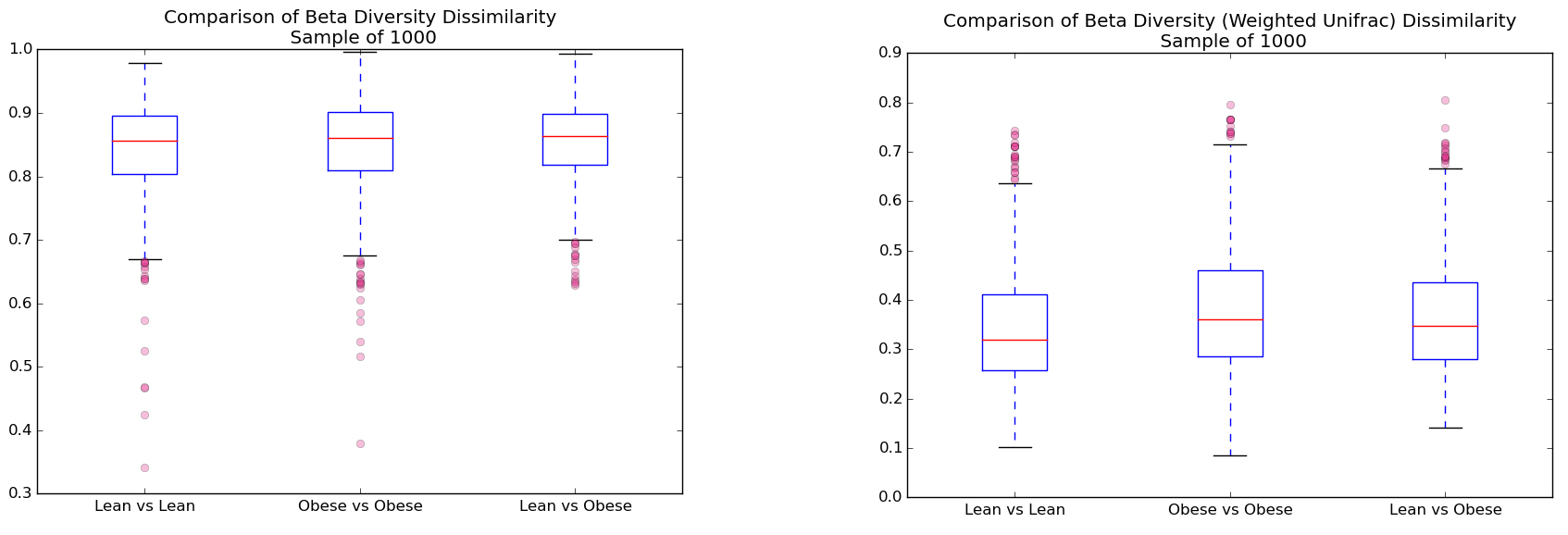
Sample console output (/docs/*firstRun\_output.txt):*

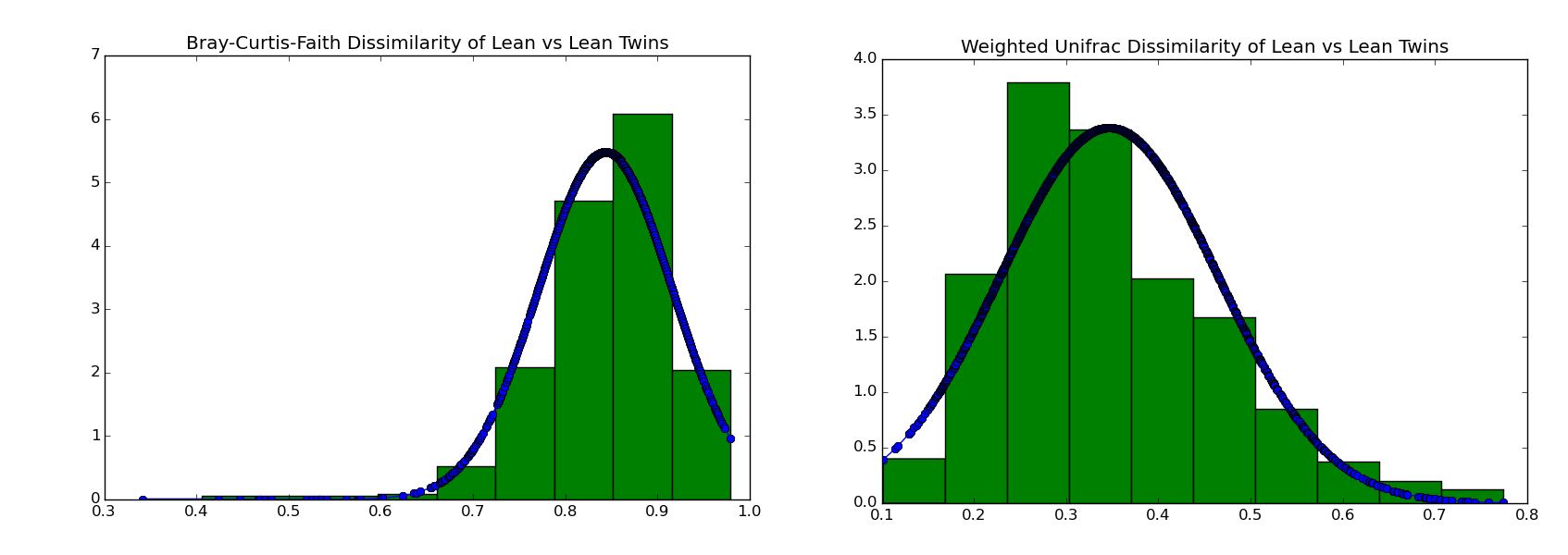
1. /usr/bin/python2.7 /home/qiime/Desktop/Project1/firstRun/biom\_read\_unifrac.py
2. /home/qiime/Desktop/Project1/firstRun/study\_77\_011618-113533 found.
3. /home/qiime/Desktop/Project1/firstRun/figures created.
4. /home/qiime/Desktop/Project1/firstRun/stats created.
5. Rarefaction **in** progress. Calling:
6. single\_rarefaction.py
7. -i /home/qiime/Desktop/Project1/firstRun/study\_77\_011618-113533/processed\_data/219\_otu\_table.biom
8. -o /home/qiime/Desktop/Project1/firstRun/study\_77\_011618-113533/processed\_data/219\_otu\_table\_even1000.biom
9. -d 1000
10. Rarefaction check completed.
11. Dissimilarity Matrix preparation **in** progress. Calling:
12. beta\_diversity.py
13. -i /home/qiime/Desktop/Project1/firstRun/study\_77\_011618-113533/processed\_data/219\_otu\_table\_even1000.biom
14. -o /home/qiime/Desktop/Project1/firstRun/stats
15. -m weighted\_unifrac
16. -t /home/qiime/Desktop/Project1/firstRun/study\_77\_011618-113533/97\_otus.tree
18. Lean-vs-Lean:
19. Low: 0.09436299
20. Mean: 0.3501875217533606
21. Std: 0.12033391562884786
22. High: 0.78030682
23. Count: 3422
25. Obese-vs-Obese:
26. Low: 0.08747150000000002
27. Mean: 0.3766034260836065
28. Std: 0.12658099158464228
29. High: 0.91778851
30. Count: 34782
32. Lean-vs-Obese:
33. Low: 0.11719817
34. Mean: 0.36671111489803315
35. Std: 0.1230288230092367
36. High: 0.8923403599999999
37. Count: 11033
39. Unifrac distances saved to:
40. Lean-vs-Lean:
41. /home/qiime/Desktop/Project1/firstRun/stats/LL\_stats\_Unifrac.csv
42. Obese-vs-Obese:
43. /home/qiime/Desktop/Project1/firstRun/stats/OO\_stats\_Unifrac.csv
44. Lean-vs-Obese:
45. /home/qiime/Desktop/Project1/firstRun/stats/LO\_stats\_Unifrac.csv
46. Boxplot saved to:
47. /home/qiime/Desktop/Project1/firstRun/figures/boxplot\_unifrac.png
48. Distributions saved to:
49. Lean-vs-Lean:
50. /home/qiime/Desktop/Project1/firstRun/figures/LL\_Distribution\_Unifrac.png
51. Obese-vs-Obese:
52. /home/qiime/Desktop/Project1/firstRun/figures/OO\_Distribution\_Unifrac.png
53. Lean-vs-Obese:
54. /home/qiime/Desktop/Project1/firstRun/figures/LO\_Distribution\_Unifrac.png
55. PCoA **in** progress. Calling:
56. principal\_coordinates.py
57. -i /home/qiime/Desktop/Project1/firstRun/stats/weighted\_unifrac\_219\_otu\_table\_even1000.txt
58. -o /home/qiime/Desktop/Project1/firstRun/stats/weighted\_unifrac\_219\_otu\_table\_even1000\_pcoa.txt
59. /usr/local/lib/python2.7/dist-packages/skbio/stats/ordination/\_principal\_coordinate\_analysis.py:107: RuntimeWarning: The result contains negative eigenvalues. Please compare their magnitude with the magnitude of some of the largest positive eigenvalues. If the negative ones are smaller, it's probably safe to ignore them, but if they are large in magnitude, 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.
60. Principal coordinates saved to:
61. /home/qiime/Desktop/Project1/firstRun/stats/weighted\_unifrac\_219\_otu\_table\_even1000\_pcoa.txt
62. Emperor visualization **in** progress. Calling:
63. make\_emperor.py
64. -i /home/qiime/Desktop/Project1/firstRun/stats/weighted\_unifrac\_219\_otu\_table\_even1000\_pcoa.txt
65. -m /home/qiime/Desktop/Project1/firstRun/study\_77\_011618- 113533/mapping\_files/2485\_mapping\_file.txt
66. -b obesitycat
67. -o /home/qiime/Desktop/Project1/firstRun/figures
68. Emperor index.html saved to:
69. /home/qiime/Desktop/Project1/firstRun/figures/index.html

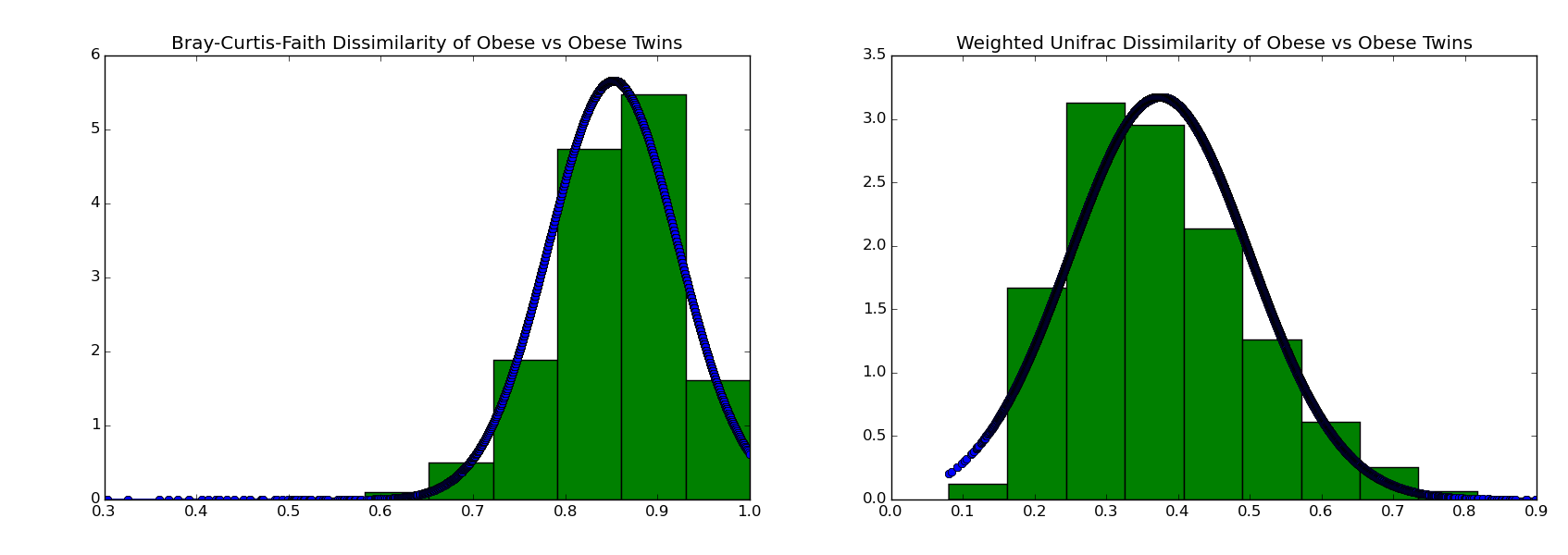
3. Progress, Updates, and Figures

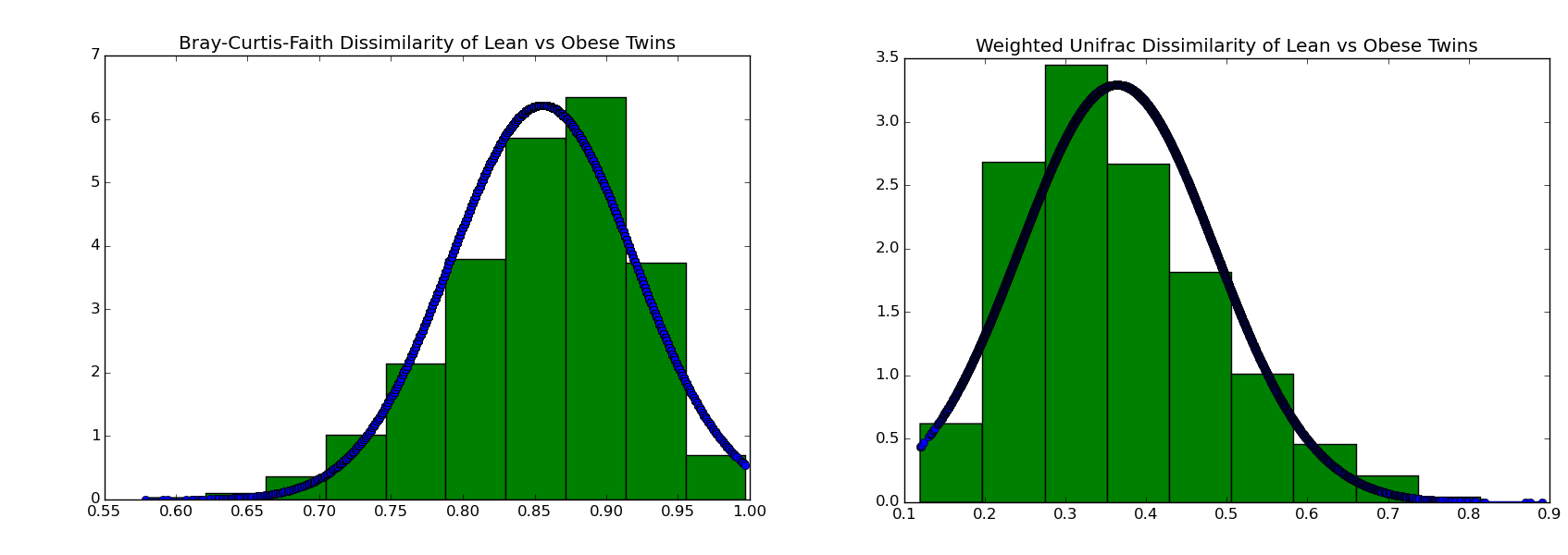
The script has been developed to be implemented in a single-scope, which is something that I plan to address in project 2. I was able to rarify the OTU table using *single\_rarefaction.py*, and compute the dissimilarity matrices using *beta\_diversity.py*. The boxplots and distributions are saved to 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.

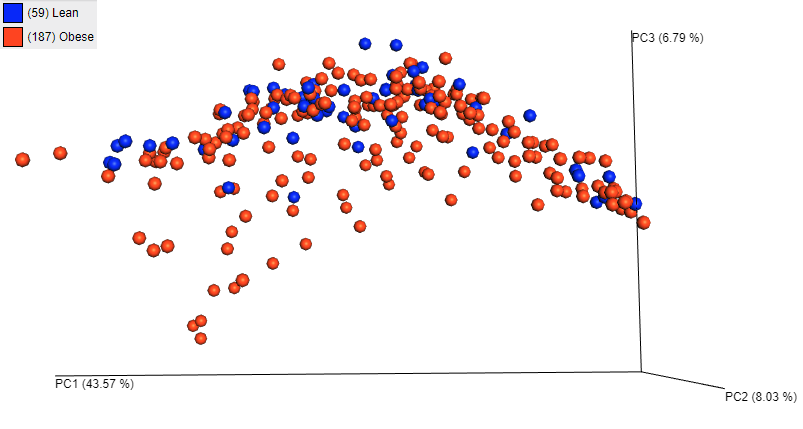


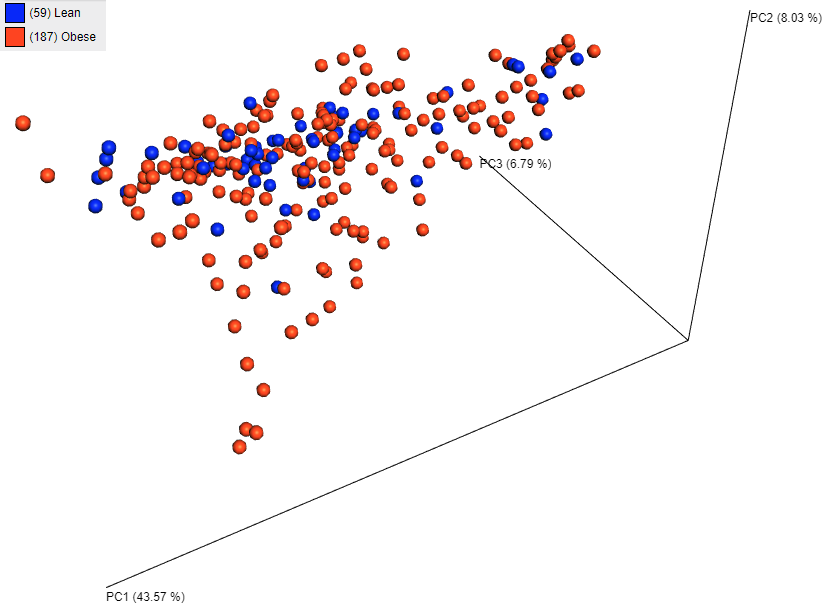


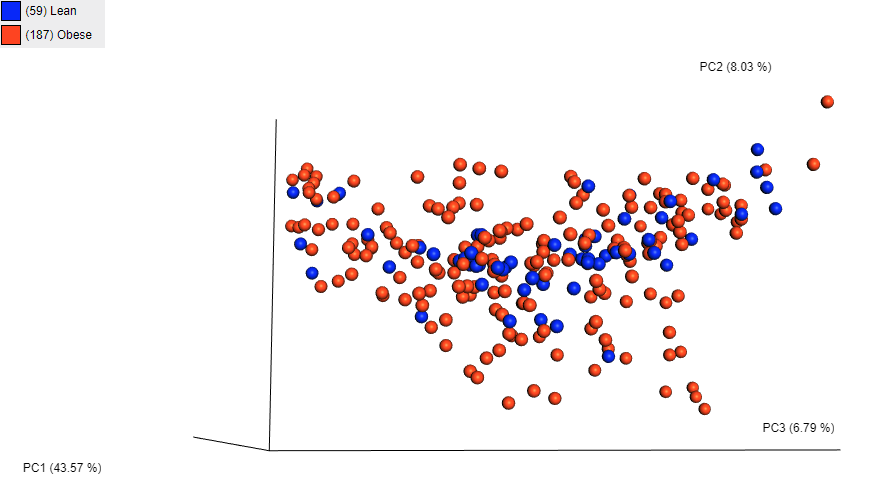




*PCoA Samples:*







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/>