


8-2022

A bioinformatics analysis of microbial diversity and its correlation with human lifestyle, diet, and health variables

Alivia Ankrum
aankrum@unomaha.edu

Kate Cooper
University of Nebraska at Omaha, kmcooper@unomaha.edu

Follow this and additional works at: https://digitalcommons.unomaha.edu/university_honors_program

 Part of the [Bioinformatics Commons](#), [Biology Commons](#), [Food Science Commons](#), [Genetics and Genomics Commons](#), and the [Nutrition Commons](#)

Please take our feedback survey at: https://unomaha.az1.qualtrics.com/jfe/form/SV_8cchtFmpDyGfBLE

Recommended Citation

Ankrum, Alivia and Cooper, Kate, "A bioinformatics analysis of microbial diversity and its correlation with human lifestyle, diet, and health variables" (2022). *Theses/Capstones/Creative Projects*. 186.
https://digitalcommons.unomaha.edu/university_honors_program/186

This Dissertation/Thesis is brought to you for free and open access by the University Honors Program at DigitalCommons@UNO. It has been accepted for inclusion in Theses/Capstones/Creative Projects by an authorized administrator of DigitalCommons@UNO. For more information, please contact unodigitalcommons@unomaha.edu.

A bioinformatics analysis of microbial diversity and its correlation with human lifestyle, diet, and health variables

Alivia Ankrum
School of Interdisciplinary
Informatics
College of Information Science

and Technology
Omaha, NE
aankrum@unomaha.edu

Dr. Kate Cooper
School of Interdisciplinary
Informatics
College of Information Science

and Technology
Omaha, NE
kmcooper@unomaha.edu

Abstract—The abundant impact of microbiota on human physiology suggests a need for exploration into their impact on human health and disease. The American Gut Project (AGP) was established to aggregate microbiome sequencing data as well as health, diet, and lifestyle metadata. This study proposes to identify taxonomic species and build a phylogenetic tree representation from the AGP participant sample collection as well as find their respective alpha and beta diversity of all metadata variables based on patient questionnaire data. Additionally, this study will involve a chimeric sequence extraction from the 16S rRNA sequences of the AGP. The expected results are hypothesized to identify the Actinobacteria's *Bifidobacterium* and the Firmicutes' *Lactobacillus* as dominant genera, as well as significant correlation between digestive or intestinal diseases and the microbial diversity due to pathogenic species often present in the microbiome. The dominant phyla were found to be *Bacteroidetes*, *Firmicutes*, and *Proteobacteria*. In contrast to predictions, the two dominant genera were found to be *Bacteroides* and *Faecalibacterium*. The subset of metadata variables that had a statistically significant correlation between both alpha and beta diversity were found which included variables relating to lifestyle habits, geographic location, diet habits, medical diagnoses, and environmental factors. (Abstract)

Keywords—microbiome, American Gut Project, QIIME2, bacteria, bacterial phylogeny, phylogenetic tree, alpha diversity, beta diversity

I. INTRODUCTION

A. Background

The human microbiome is made up of over 100 trillion microbes, or microbiota, including that of bacteria, fungi, protozoa, and viruses which live in adjacency to surrounding eukaryotic cells [1]. Although, these microbial cells outnumber the number of eukaryotic cells ten to one, which translates to 200 times more microbial genes than genes in the human genome [1]. The microbial cells in our body, usually in the large intestine, have many functions including digestion, regulation of the immune system, protection against disease-causing bacteria, production of vitamins, and more [1]. Because of these diverse and abundant functions that the microbiome performs, they thus have the ability to impact human physiology, and therefore can influence health and disease [2].

As technological advancements have developed, the human microbial components have undergone culture-independent study through the use of 16S rRNA-encoding gene sequencing

paired with alignment to bacterial reference sequences [2]. This gene is commonly utilized to differentiate between microbiota of varying bacterial phylogeny and taxonomy due to its almost universal presence in bacteria coupled with its conservation of function over time [3]. 16S rRNA sequencing produces genus identification over 90% of the time with 65 to 83% of the genus identifications also assigned a species identification [3].

There have been significant initiatives to collect data on the human microbiome, including the Human Microbiome Project (HMP) and European Metagenomics of the Human Intestinal Tract (MetaHIT), which revealed 200 times the amount of microbial DNA sequences previously reported, allowing for a significant increase in the known diversity of the human microbiome [2]. This increased microbial diversity has subsequently led to increased research into correlation between the presence of certain microbiota with human health. The microbiome diversity has thus been linked to physiological changes within the body, thought to possibly contribute to the pathogenesis of diseases, such as rheumatoid arthritis, colorectal cancer, obesity, diabetes, cardiovascular disease, irritable bowel disease, inflammatory bowel disease, and more [2]. There have also been significant research that supports the idea that modulation of one's diet has a significant benefit on their microbiome, and in turn greatly effects their health. There are numerous ways to positively impact health by altering ones' microbiota makeup, such as through taking probiotics from genera *Bifidobacterium* and *Lactobacillus*, or other methods that increase the diversity of microbiota in one's intestine [4].

One of the few large-scale studies for human microbiota data comes from the American Gut Project (AGP) which gathers stools samples and patient meta-data in the form of a questionnaire to analyze the diversity of microbiota and how these correlate with diet, lifestyle, and health metadata [5]. The intention of this study is to identify and analyze the taxonomical diversity from the 16S rRNA sequencing data published by the AGP. Additionally, microbiota species identification and diversity analysis will be correlated with the patient's meta-data such as allergies, diagnoses, and diet. This will be used to determine reasonable association between specific microbiota and patient health, diet, and lifestyle metadata. This study will also involve a chimeric sequence extraction from the 16S rRNA sequence aggregate, and further analyze these results and apply a comparison to the original taxonomic classifications. This project ultimately aims to correlate microbiome diversity with

patient lifestyle, diet, and health data in order to provide a reasonable method of association in future patient management.

B. Hypothesis

After execution of the taxonomic classification, metadata analysis, and diversity analysis, an expected outcome of this project involves identifying the dominant genera as *Bifidobacterium* and *Lactobacillus*, as previous analysis of the American Gut Project had concluded from ASV inferences [9]. Because microbial species within both of these genera are often used in probiotics, it is expected that the metadata involving probiotic use and diet regimens will influence the diversity of microbial taxa diversity more than that of lifestyle or health variables. Although, it is expected that health data involved with that of digestive or intestinal diseases will also greatly affect the microbial diversity due to the innate state of disease being linked to lack of homeostasis brought on by the lack of presence or excessive presence of certain pathogens.

II. MATERIALS AND METHODS

A. Data Availability

The data from the American Gut Project is publicly available via the European Nucleotide Archive (ENA) at EMBL-EBI using accession number PRJEB11419 and will be downloaded locally [6]. The FASTQ files of the first 998 samples were downloaded using enaBrowserTools with the project accession on the command line. The corresponding metadata files were downloaded via a Python script utilizing wget to interact with the ENA REST APIs. All corresponding documentation and code can be downloaded at the GitHub Repository corresponding to this project at https://github.com/aliviaankrum/BIOI_capstone¹.

B. QIIME2 Analysis

QIIME2's end-to-end microbiome analysis tools was used to perform species identification, phylogenetic tree construction, and diversity analysis using questionnaire results submitted by participants of the American Gut Project [7]. The first 998 microbiome sequence files from PRJEB11419 were imported into QIIME2 using a manifest file in 'SingleEndFastq-ManifestPhred33V2' format. Firstly, DADA2 was used to perform denoising of the single-end sequences. To perform taxonomic classification, the QIIME2 feature classifier plugin was used to classify the previously produced ASVs alongside a naive Bayes machine-learning classifier using the Greengenes 13_8 reference set to identify taxa. The resulting taxonomic classifications were then used to create a phylogenetic tree using the fragment insertion plugin which inserts the sequences into the Greengenes 13_8 99% identity reference tree backbone. The chimeric sequence identification was performed using the 'vsearch uchime_ref' method, which uses the UCHIME de novo program alongside the QIIME2 'vsearch' analysis pipeline tool based on the sequencing abundance data [8]. Alpha and beta diversity was calculated using the QIIME2 diversity plugin. The alpha diversity was calculated, which includes Faith's phylogenetic diversity and Pielou's evenness value, as well as the beta diversity using a PERMANOVA test of variance. The alpha diversity metrics were then used to understand the diversity within samples and

how this relates to each metadata variable. Beta diversity will be used to analyze differences in microbiome diversity between the samples and therefore, how each metadata variable involving health, diet, and lifestyle is correlated with the identified differences in taxa classification. Additionally, the chimeric sequence extraction will be done

III. RESULTS

A. Quality Control

After performing QIIME2's importation command, a visualization was produced to summarize sequence statistics (Table 1) as well as a quality control plot (Fig. 1). Table 1 reveals a wide range of forward sequence counts, though only 93 sequences ranged from one to 57 base pairs in length, while a significant jump from 57 to 2952 subsequently occurs.

Measurement	Forward reads
Minimum	1
Median	30257.5
Mean	30556.885772
Maximum	76651
Total	30495772

TABLE I. DEMULTIPLEXED SEQUENCES SUMMARY

Table showing the minimum, median, mean, maximum, and total number of forward reads calculated from the 998 FASTQ files imported into QIIME2.

Fig. 1 shows a box and whisker plot for each base pair within the 150 base pair forward read. Each quality measure per sequence base was calculated by taking a random sample of 10,000 sequences out of the total 3,0495,772 forward reads. The lowest quality score measured at the 9th percentile was 12 at position 150. Though, the next highest 9th percentile quality score was 27 at positions 139 and 140. At all but 13 base positions, the median or 50th percentile was between quality scores of 37 and 39. The full TSV file detailing all quality scores for each sequence base position can be found on the GitHub corresponding to this project at https://github.com/aliviaankrum/BIOI_capstone¹.

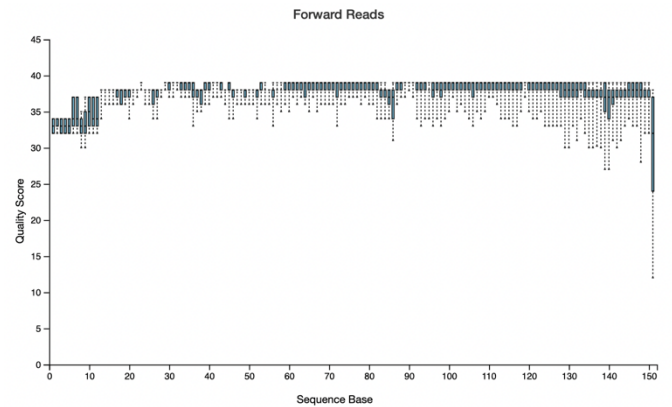


Fig. 1. Interactive quality control plot visualized using QIIME2View. At each sequence base, a box and whisker plot portrays the 9th, 25th, 50th, 75th, and 91st percentile quality scores for a random sample of 10000 forward reads.

¹This is a private repository, for access, email aankrum@unomaha.edu

B. Taxonomical Classification

C. Phylogenetic Tree Construction

directly from the center. Additionally, the ‘align-to-tree-mafft-fasttree’ pipeline was utilized to construct another phylogenetic tree, this time, portraying all 32940 unique features detected in the 998 sequences (Fig. 4). This tree more accurately shows the relative phyla distribution based on the imported microbiome sequences. These phylogenetic trees as .qza artifacts can be downloaded from the GitHub corresponding to this project at https://github.com/aliviaankrum/BIOI_capstone1.

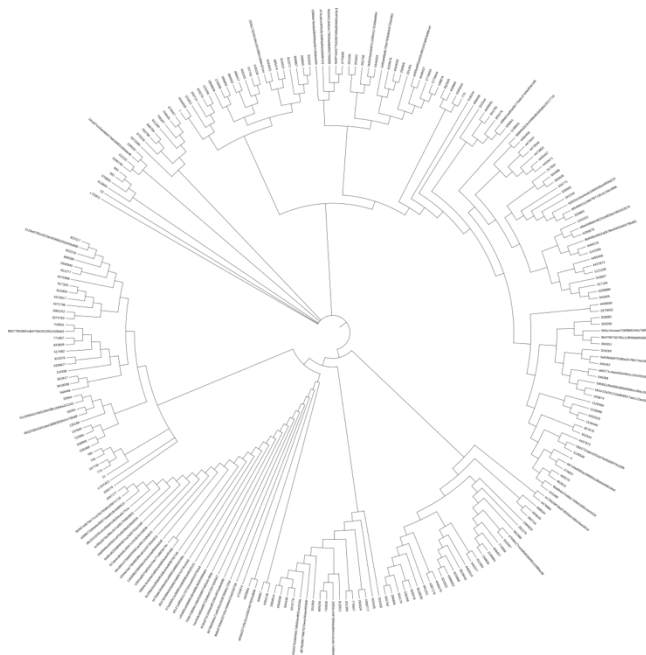
[illegible]

Fig. 2. Taxonomical classification bar plot showing relative frequencies of each microbial organism at a phylogenetic level of 7, denoting species identification, for each patient sample seen on the x-axis. This image and legend represents a small portion of the full visualization, and therefore does not represent all samples or taxonomical classifications, respectively.

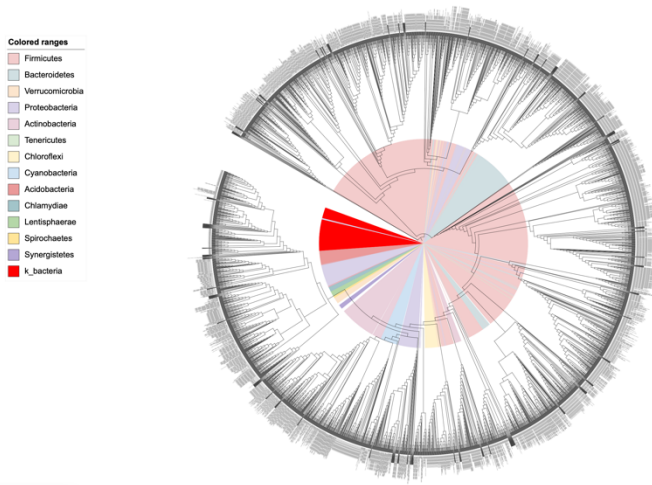


Fig. 4. Feature phylogenetic tree mapping all 32940 OTUs detected in the 998 FASTQ files imported into QIIME2. Shaded portions represent phylum classification, aside from the red section which represents an undetectable phyla and OTUs only defined within the kingdom Bacteria, viewed using ITOL.

D. Alpha and Beta Diversity

The alpha diversity was first calculated using the ‘core-metrics-phylogenetic’ pipeline. Based on the results of the resulting DADA2 table artifact, the sampling depth was set at 5047, because of its distinguished separation from the uncharacteristically low number of sequences in a few samples, which also allowed most of the samples to be retained. The Faith Phylogenetic Diversity, a measure of alpha diversity, revealed which metadata variables were statistically significantly associated with species richness, or how much of the phylogenetic tree is represented in the sample. For each alpha diversity measurement, a p-value is calculated that reveals whether the association between species richness and the metadata variable is statistically significant. When visualized in QIIME2View, they produced a box and whisker plot, like the one shown in Fig. 5. All metadata variables with a Faith’s Phylogenetic alpha diversity p-value less than 0.05 is listed in Table 2 with the corresponding p-value.

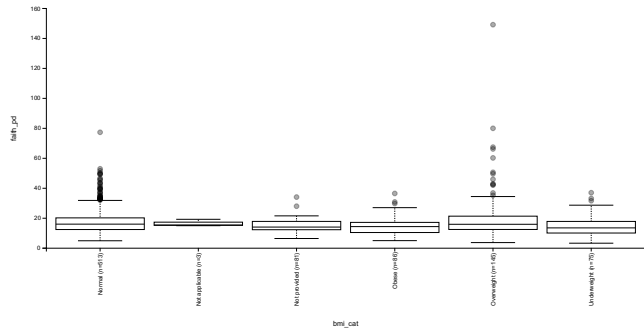


Fig. 5. Sample Faith Phylogenetic Diversity boxplot representing the alpha diversity, or species richness, present in one of the categorical variables tested in the American Gut Project. This image shows alpha diversity measures of the ‘bmi_cat’, a categorical variable for the patient’s Body Mass Index. As shown, there is significant difference in alpha diversity in those who are overweight and normal.

Metadata Variable	P-value
Age_cat	0.0032698629643783685
Alcohol_consumption	0.0001943406975716111
Alcohol_frequency	0.000016268361573957885
Antibiotic_history	0.00039896899756371434
Appendix_removed	0.019592670823693464
Birth_year	0.021122750674523724
Bmi_cat	0.00010883129994906723
Body_habitat	2.776623161224216e-51
Body_product	2.776623161224216e-51
Body_site	2.0575774719245235e-50
Bowel_movement_frequency	0.025832436239938417
Census_region	0.01701165501941924
Collection_date	0.019239404315396724
Collection_month	1.4012767014455236e-8
Collection_time	0.0281505078576528
Csection	0.0028991712289122462
Drinking_water_source	0.023431988283273076
Economic_region	0.000034583728625899345
Env_material	2.776623161224216e-51
Env_package	2.776623161224216e-51
Exercise_frequency	0.0006819892336315701
Exercise_location	0.0013139094435320334
Height_cm	0.0338452873503027
Ibd	0.003133324358146945
Latitude	0.0423326043103398
Longitude	0.021990268160364368
Milk_cheese_frequency	0.037599110648606736
Sample_type	1.4411950703055883e-49
Skin_condition	0.02362689149906728
Sleep_duration	0.031170752400950213
State	0.0013042506756748959
Subset_age	0.0005736253664513524
Subset_antibiotic_history	0.00010382988424618728
Subset_bmi	0.000004931894070504741
Subset_healthy	7.665407219457153e-10
Subset_ibd	0.0033952496073359154
Types_of_plants	0.00032147286037028294
Weight_change	0.008164336542164086

TABLE II. ALPHA DIVERSITY P-VALUES

Table showing the metadata variables with an alpha diversity p-value less than 0.05, indicative of a statistically significant correlation between species richness and each variable listed.

The beta diversity was later calculated using a PERMANOVA test with the ‘beta-group-significance’ command in QIIME2. The beta diversity command measures the similarity of microbiome composition between samples and based on the resulting p-value, reveals whether the association between microbiome diversity and metadata variable is statistically significant. Each categorical metadata variable underwent pairwise permutation tests to reveal these statistically significant associations between instances of the metadata variable and their microbiome composition. Table 3 shows the metadata variables tested and corresponding p-value if less than 0.05.

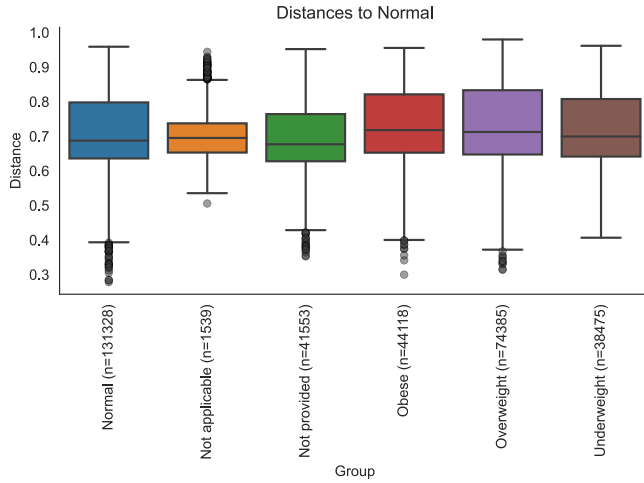


Fig. 6. PERMANOVA group significance plots for the metadata variable ‘bmi_cat’, a categorical variable for the patient’s Body Mass Index, representing the distances between participants who are within ‘normal’ BMI to other responses, a measure of beta diversity.

IV. DISCUSSION

A. Taxonomical Distribution

Based on the results of the taxonomical classification and phylogenetic tree construction, the majority of bacterial phyla represented by all 998 microbiome samples were found to be *Firmicutes*, *Proteobacteria*, and *Bacteroidetes*. Additionally, the phylogenetic distribution of all microbiota represented in this study is mainly broken up between two major clades. As seen in both Figs. 3 and 4, the middle of the graph separates into two branches, that subsequently contain the majority of microbiota or OTUs. One clade is more diverse than the other, which contains mostly *Firmicutes*. Therefore, it can be concluded that the majority the human microbiome analyzed in these samples is composed of bacteria classified as *Firmicutes*, *Proteobacteria*, and *Bacteroidetes*.

B. Diversity Analysis

After performing alpha diversity measures, several metadata variables were shown to have a statistically significant association with species richness, or the amount of microbiota present in one microbiome sample. To better understand how the microbiome affects health outcomes, alpha diversity is often calculated and found to be correlated with human health variables, such as alcohol consumption [10]. This could be due to complex microbiota interactions with each other as well as a

Metadata Variable	P-value
Age_cat	0.001
Age_corrected	0.001
Age_years	0.001
Alcohol_consumption	0.001
Alcohol_frequency	0.001
Alcohol_types_red_wine	0.021
Alcohol_types_unspecified	0.016
Allergic_to_unspecified	0.037
Antibiotic_history	0.001
Autoimmune	0.048
Birth_year	0.001
Bmi	0.001
Bmi_cat	0.001
Body_habitat	0.001
Body_product	0.001
Body_site	0.001
Cat	0.011
Census_region	0.001
Chickenpox	0.048
Collection_date	0.001
Collection_month	0.001
Collection_time	0.001
Collection_timestamp	0.004
Contraceptive	0.001
Cosmetics_frequency	0.023
Csection	0.006
Depression_bipolar_schizophrenia	0.035
Diet_type	0.006
Drinking_water_source	0.002
Economic_region	0.001
Elevation	0.001
Env_material	0.001
Env_package	0.001
Exercise_frequency	0.016
Exercise_location	0.004
Flossing_frequency	0.007
Flu_vaccine_date	0.02
Height_cm	0.041
Ibd	0.02
Latitude	0.001
Livingwith	0.004
Longitude	0.001
Milk_cheese_frequency	0.040
Other_supplement_frequency	0.005
Race	0.001
Roommates	0.042

<i>Metadata Variable</i>	<i>P-value</i>
Sample_type	0.001
Seafood_frequency	0.050
Seasonal_allergies	0.005
Sex	0.008
Sleep_duration	0.047
Softener	0.045
State	0.001
Subset_age	0.002
Subset_antibiotic_history	0.001
Subset_bmi	0.006
Subset_healthy	0.001
Subset_ibd	0.013
Sugary_sweets_frequency	0.045
Teethbrushing_frequency	0.029
Tonsils_removed	0.047
Types_of_plants	0.007
Vegetable_frequency	0.035
Weight_change	0.023
Weight_kg	0.001

TABLE III. BETA DIVERSITY P-VALUES

Table showing the metadata variables with a beta diversity p-value less than 0.05, indicative of a statistically significant correlation of microbiome diversity between samples for each metadata variable.

great variety of interactions between microbiota and the human body, possibly leading to disease or physiological imbalance.

This idea is reflected in this study, as there was a significant association between alcohol consumption and frequency with both alpha and beta diversity meaning that alcohol has a noticeable impact on microbiome diversity, which can in turn, possibly affect human physiology in other ways. Additionally, microbiome composition after an appendectomy is significantly altered, thought to be due to the appendix's role in regulation of intestinal microbiota, corroborating the correlation seen in this study [11]. Further, age, method of delivery, diet, antibiotic use, and probiotic use have been found to alter one's microbiome composition, supporting the statistically significant correlation between both alpha and beta diversity and metadata variables representing age, c-section delivery method, diet regimens, and antibiotic use seen in this study [12].

Additionally, the plethora of variables relating to the environment from which patient samples were collected, including both the geographic location and variables like whether the participant has a cat, had a significant impact on the alpha and beta diversity of the microbiome data. This has been explained to be due to the increased urbanization of many cities, leading to decreased aggregation of necessary microbiota as well as increased abundance of inflammatory diseases due to the dense human population [13]. Finally, daily habits such as exercise and sleep were found to produce a significant correlation with both alpha and beta diversity, though these

lifestyle variables seem to affect the beta diversity more than the alpha diversity based on the increased number of lifestyle variables that have a beta diversity p-value less than 0.05 such as flossing frequency, if the participant uses fabric softener, or how often the participant brushes their teeth. As shown by both statistically significant correlations between many of the metadata variables tested in the American Gut Project with both alpha and beta diversity, this project supports the idea that multiple lifestyle, health, and diet variables discussed here affect the composition of the human microbiome.

C. Downfalls

The American Gut Project currently includes 39,017 participant samples, which were initially anticipated to all be included in this project. Due to storage constraints when opening the imported sequence artifact (.qza), a small subset containing 998 participant samples was used for analysis. Because this is a small subset, there are only a small number of participants that have instances such as gluten intolerance or cardiovascular disease, which makes statistical inferences much less reliable. Therefore, future work is aimed to implement all possible samples to increase the accuracy of statistical correlations between metadata and microbiota diversity.

D. Implications

Due to the continual research investigating the effects of microbiome modulation on human health outcomes, there is need for understanding how human lifestyle, health, and diet variables change the makeup of the microbiome. Based on the results of this experiment, correlation between microbiota presence and biological bias towards health conditions, diseases, or lifestyle habits can be implemented into healthcare screening and diagnosis. One's microbiome sampling may be able to propose a way to provide a new method of diagnosis and medical management for patients based on their respective microbiome composition.

ACKNOWLEDGMENT

I would like to express gratitude to my thesis advisor, Dr. Kate Cooper, for all the meetings throughout the past year and especially for her insight and assistance through each stage of this project.

REFERENCES

- [1] A. Shreiner, J. Kao, and V. Young, "The gut microbiome in health and in disease," vol. 31, no. 1, pp. 69–75, Jan. 2015, doi: 10.1097/MOG.000000000000139.[Online]. Available: <https://www.ncbi.nlm.nih.gov/pubmed/25394236>
- [2] M. Hair and J. Sharpe, "Fast Facts About The Human Microbiome," 2014. [Online]. Available: https://depts.washington.edu/ceeh/downloads/FF_Microbiome.pdf
- [3] J. M. JANDA and S. L. ABBOTT, "16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls," vol. 45, no. 9, pp. 2761–2764, 2007, doi: 10.1128/JCM.01228-07.[Online]. Available: <https://www.ncbi.nlm.nih.gov/pubmed/17626177>
- [4] Y. Feng et al., "An examination of data from the American Gut Project reveals that the dominance of the genus Bifidobacterium is associated with the diversity and robustness of the gut microbiota," vol. 8, no. 12, pp. e939–n/a, Dec. 2019, doi: 10.1002/mbo3.939. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/mbo3.939>

- [5] D. McDonald et al., "American Gut: an Open Platform for Citizen Science Microbiome Research," vol. 3, no. 3, May 2018, doi: 10.1101/277970. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pubmed/29795809>
- [6] M. Daniel et al., "American Gut: an Open Platform for Citizen Science Microbiome Research," vol. 3, no. 3, pp. e00031-18, doi: 10.1128/mSystems.00031-18. [Online]. Available: <https://doi.org/10.1128/mSystems.00031-18>
- [7] E. Bolyen et al., "Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2," vol. 37, no. 8, pp. 852–857, 2019, doi:10.1038/s41587-019-0209-9. [Online]. Available: <https://doi.org/10.1038/s41587-019-0209-9>
- [8] R. C. Edgar, B. J. Haas, J. C. Clemente, C. Quince, and R. Knight, "UCHIME improves sensitivity and speed of chimera detection," vol. 27, no. 16, pp. 2194–2200, 2011, doi: 10.1093/bioinformatics/btr381.
- [9] Y. Feng et al., "An examination of data from the American Gut Project reveals that the dominance of the genus *Bifidobacterium* is associated with the diversity and robustness of the gut microbiota," vol. 8, no. 12, p. e939, 2019, doi:10.1002/mbo3.939. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/31568677>
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6925156/>
- [10] Hagerty, S. L., Hutchison, K. E., Lowry, C. A., & Bryan, A. D. (2020). An empirically derived method for measuring human gut microbiome alpha diversity: Demonstrated utility in predicting health-related outcomes among a human clinical sample. *PLoS ONE*, 15(3). <https://doi.org/10.1371/journal.pone.0229204>
- [11] Cai, S., Fan, Y., Zhang, B., Lin, J., Yang, X., Liu, Y., Liu, J., Ren, J., & Xu, H. (2021). Appendectomy is associated with alteration of human gut bacterial and fungal communities. *Frontiers in Microbiology*, 12. <https://doi.org/10.3389/fmicb.2021.724980>
- [12] Hasan, N., & Yang, H. (2019). Factors affecting the composition of the gut microbiota, and its modulation. *PeerJ*, 7. <https://doi.org/10.7717/peerj.7502>
- [13] Tasnim, N., Abulizi, N., Pither, J., Hart, M. M., & Gibson, D. L. (2017). Linking the gut microbial ecosystem with the environment: Does Gut Health depend on where we live? *Frontiers in Microbiology*, 8. <https://doi.org/10.3389/fmicb.2017>