Giheon Koh
B01239526, University of Central Arkansas
FALL 2018
Employed in University of Arkansas for Medical Science,
4301 W Markham St, Little Rock, AR 72205
Dr. Se-Ran Jun
sjun@uams.edu

Abstract

Humans are theoretically identical in their genetic compositions but the small differences in the DNA and RNA exist to the phenotypic diversities across the human population. For this internship, I examined the 16S rRNA survey data sampled from healthy humans at different body sites. After filtering contaminated sequences (from mitochondria and chloroplast), uninformative and rare OTUs, and sample normalization based on alpha rarefaction, I compared samples based on alpha diversity and then, performed Principle Coordinate Analysis to explore relationships between samples based on beta diversity. Additionally, I determined the significance of groups of samples by body sites based on Kruskal-Wallis and permutation-based statistical test with adjusted p-values. Finally, I identified the healthy human core microbiome by the parts at the OTUs as well as the taxonomy level for given body sites.

Introduction

Inspiration

The studies of the human microbiome[1] have uncovered that the healthy person notably differ in the microbiome[1] diversities in the human body parts such as Stool, Skin, and Oral environments. Much of these diversities remain unexplained, although diet, host genetics and early microbial exposure have all been implicated. The studies report that the instabilities in the microbiomes[1] have been associated with numerous diseases, including inflammatory bowel disease, multiple sclerosis, diabetes, allergies, asthma, autism and cancer. Understanding this variability in the healthy microbiome[1] has thus been a major challenge in microbiome[1] research, dating back at least to 1960s, continuing through the Human Microbiome Project (HMP) and beyond. Cataloguing the necessary and sufficient sets of microbiome[1] features that support correcting microbial configurations that are implicated in disease. Thus, finding features that broadly distinguish the healthy from unhealthy microbiomes[1] will aid in the diagnosis of microbiome[1] related diseases and could potentially provide new means to prevent disease onset or to improve prognosis.

Software

For this internship, all the works went through the MacOSX operating system. I explored QIIME 2 (which is a next generation microbiome bioinformatics platform) for quantitative microbiome[1] analysis and programmed in Python for filtering and cleaning the 16S survey data, and treating metadata[4] and taxonomy[3]. R was also essential to generate the heatmaps indicating the diversities of core microbiomes[1]. Thus, I used Mac OSX, Python and R for the internship research.

Microbiome Characterization

Microbiome can be characterized by two ways: Shotgun metagenomic sequencing and 16S amplicon sequencing technologies. The shotgun metagenomic technology surveys the entire genomes of all the organisms present in the sample. It can sequence all DNA materials from all the organisms (including bacterial, archaea, virus, fungi) so that it can provide information of the presence or absence of specific functional pathways in the sample.

Unlike Shotgun metagenomic sequencing technology, the 16S amplicon sequencing technology amplifies and sequences only small hypervariable regions of 16S rRNA gene which is a universal gene for bacteria and Archaea and can be used to identify all the species present in a huge number of samples very quickly. However, this method is limited to the identifications of bacteria and archaea and cannot directly provide functional information of microbial community. After demultiplexing and sequence quality control (these steps were not performed in this project because the data (feature table) which we

downloaded from the Human Microbiome Project (https://hmpdass.org) were already taken care of with these steps), the standard pipeline for 16S amplicon sequencing data analysis starts by clustering sequences within a percent sequence similarity threshold (typically 97%) into operational taxonomic units (OTUs[2]) against the reference database of 16S rRNA sequences. We employed Greengenes database (https://greengens.secondgenome.com/downlods) as the reference database in this study. To answer which organisms are present in the sample and how microbial community structure looks like, it explores OTUs' taxonomic information, OTU[2] phylogeny (phylogenetic tree), and relative abundances of OTUs[2] in the community, which comes from the diversity analysis.

16S amplicon sequencing technology

There are two ways to read or sequence target hypervariable regions: Paired-end and Single-read sequencing technologies. The paired-end sequencing technology is the reading technology literally starting from the both edges whereas the single-read sequencing technology is reading from one-side edge *(see the examples a and b in Figure 1)*. The given OTU[2] sample from HMP is sequenced in using of the Paired-end technology for the more precise information.

In the research, we used data from 16S amplicon sequencing method because our goal is to investigate the structures of healthy human microbiomes[1] which will allow us to compare the healthy human microbiomes[1] structures to the structures of diseases human microbiomes[1], especially for the gut, equally considered with stool, sample in the research.

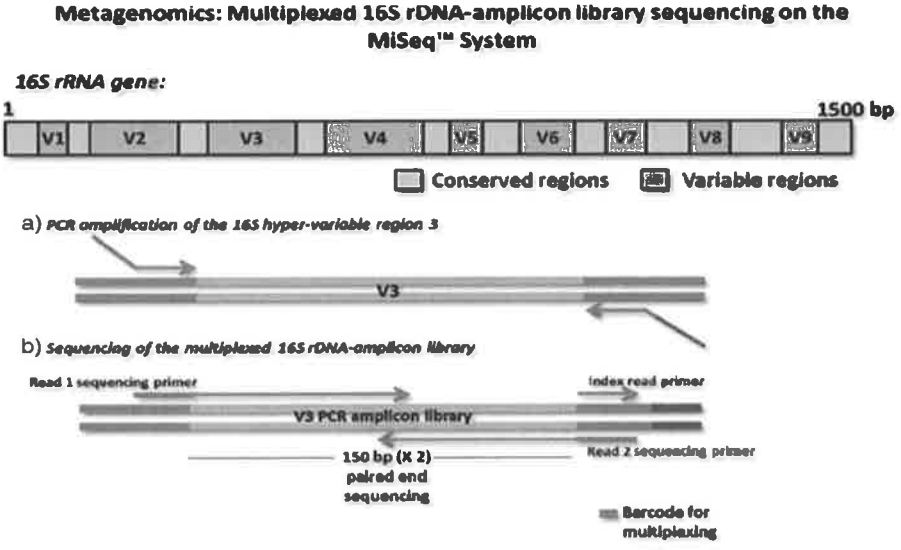

Figure 1 | The 16S amplicon sequencing method finds some variable regions against conserved regions. It means that the variable regions will be captured for each sample, thus to know what kind of organisms are present in the sample. The hypervariable region 4(V4) is the most popular target for 16S amplicon sequencing for accuracy of community structure recovery. However, for this project, the samples we used were sequenced from targeting the regions V1, V3 and V3, V5.

Data Collection

Three datasets were selected and one dataset was made for the research from HMP website, *hmpdacc.org*: OTU[2] sample table, taxonomy[3] table, metadata[4], and the data from 16S sequencing regions of V13 and V35, which are explained V13 as the data from V1 and V3, and V35 as the data from V3 and V5 of 16S sequences. Additionally, the data of V13 and V35 involve the OTU[2] table with metadata[4] and taxon[3−1]. Thus, a total of 43208 features with 3841 samples for V13 and 45412 features with 4857 samples for V35 were available for this study, representing most of the target Human Microbiome Project (HMP) cohort of the individuals. In this report, we will only look over the samples of V13 for the same processes with V35.

Thesis Statement

Therefore, through the internship, we investigated the 16S rRNA survey data associated with the taxonomy[3] to identify the healthy human microbiomes[1] in terms of the core OTUs[2] for the samples to compare with disease microbiome[1] using the samples collected at different body sites focused on Stool, Skin, and Oral.

Procedures

Data Cleaning and Filtering

Filtering and cleaning data is essential to delete uninformative data to reduce a possibility of error occurrence as well as to give more power to the analytical result. In this stage, I wrote and run the Python codes and QIIME2 code.

To filter and clean data, setting for the valid data frame to input into the analytical process was placed at the first step. Since the data have too much numbers of features and samples, we run BLAST to find the regions of similarity between biological sequences and merge. The BLAST, as an acronym of Basic Local Alignment Search Tool from NIH, compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. With the 97% of statistical significance of features and samples from metadata[4] in this research, it generates more detailed and significant samples by merging owing to their similarities based on the Greengene database, therefore, it results the table of representative OTUs[2], the powerful representatives.

As the table of representative OTUs[2] was generated through the BLAST, the next step was to compare and find the common OTU[2] IDs in the representative OTUs[2] table, OTU[2] samples, taxonomy[3]and metadata[4] using Python to generate the closed OTU[2] table for the research. Because the closed OTU[2] table should have only identified OTUs[2] and samples, we used the representative

OTUs[2] table, whose OTUs[2] have been identified in BLAST step in association with taxonomic information, and the metadata[4], whose resources are known, to compare with OTU[2] samples in terms of certain OTUs[2] and samples, thus, we filtered for commonly existing data across the all data tables with comparing information to process analysis and generated the table dimension of 5051 OTUs of 2899 samples from 43208 by 3841 data table.

With the closed OTU[2] table against samples, I went over to filter the data against taxonomy[3] for excluding all the mitochondria and chloroplast and let the OTUs[2] be involved only in the kingdom of bacteria to prevent a possibility of unreasonable result. Also, we considered a sequencing machine error, so we removed rare OTUs[2] and taxonomy[3] with 0.0005% of detection[15]as well as the samples having no closed OTUs[2], then the table whose dimension of 5007 by 2898 was generated.

Phylogenetic Tree

A Phylogenetic tree is needed to know the evolutionary path of OTUs[2], and bind them according to their common ancestors. A phylogenetic tree is a diagram that illustrates the evolutionary relationships among OTUs[2] *(See the 1st picture in Figure 2)*. It is consisted of branches (lineages), whose length represents the evolutionary time, nodes (OTUs[2]), root of tree (the common ancestor of all OTUs[2]). For the phylogenetic tree, the tree topology is the important concept with its branching pattern. In the tree topology, there are two types of trees: rooted and unrooted tree *(See the 2nd and 3rd picture in Figure 2)*. Rooted tree shows directed to a unique node (root), known as common ancestor while the unrooted tree shows the relatedness of OTUs[2] without assuming ancestry at all that there is no evolutionary path. To infer evolutionary relationships between the taxonomies[3], it is required to convert unrooted tree to rooted tree. There are two major ways to convert: by outgroup and by midpoint which is based on distance measure. The outgrouping way uses taxonomies[3] that are known to fall outside of the group of interest whereas the midpoint way is grabbing out the midway point between the two most distant taxonomy[3] in the tree. In the practice, the phylogenetic tree is generated through the midpoint way.

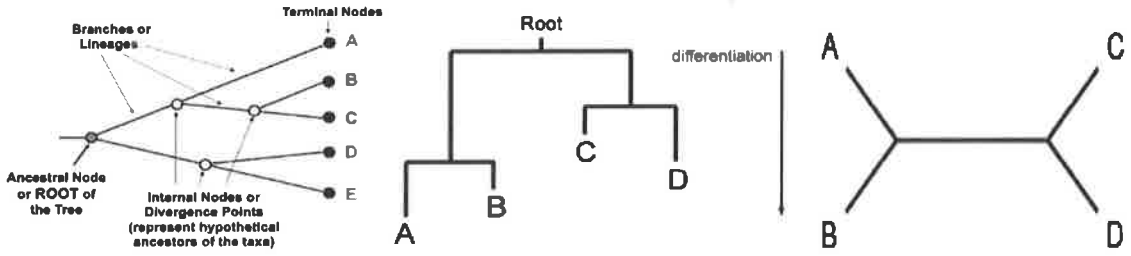

Figure 2 | left: phylogenetic tree(a), mid: rooted tree(b), right: unrooted tree(c)
a. tree says that B and C are more closely related to each other than either is to A and then A, B, and C form a clade that is a sister group of the clade composed of D and E.

There are steps to generate a phylogenetic tree and all the processes went over through QIIME 2 in Linux. First, with rRNA sequences, we align the multiple sequences. Aligning the multiple sequences is the process that involves an attempt to place residues in columns that derive from a common ancestral residue by substitutions, then, becomes a representation of a set of sequences in which equivalent characters are aligned in columns *(See the figure 3).* While doing this, there should be some empty spots, gap penalties, so we have one more step to mask these empty spots in sequences. Then, to go over the next step, we use one of the phylogenetic inferences: Distance methods, Maximum Likelihood, Bayesian inference, or else. In the practice, we used the maximum likelihood inferences to estimate the tree reliability, then, converted to rooted tree using midpoint method.

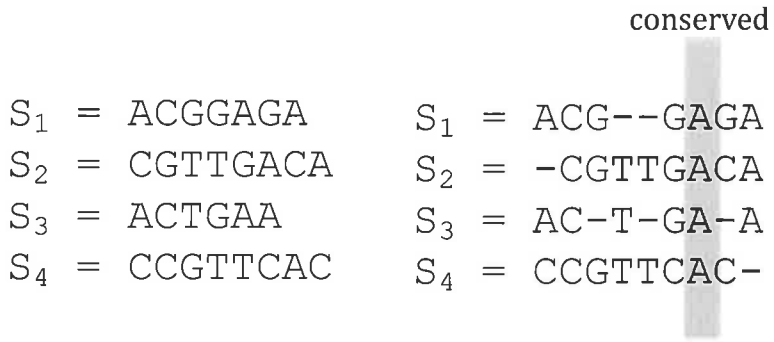

Figure 3 | S_1, S_2, ..., S_k a set of sequences over the same alphabet. The goal is to find alignment that maximizes some score function.

Alpha Rarefaction

To compare the large numbers of OTUs[2] with the same conditions, the normalization was necessary. Since every samples have different rRNA sequencing depth which came through the paired-end reading method, the alpha rarefaction was conducted with the depth of 5000 and 10000 to control the sequencing depths to start with the same standards. For the definition, the alpha rarefaction is a technique to assess species richness from the results of sampling, and it allows the calculation of species richness for a given number of individual samples based on the construction of rarefaction curves. The reason why we set the depths in two ways is to consider the possibility to miss any information after the depth of 5000, but as there is no big difference between the two depths, we determined to set and use the data with the depth of 5000. The *Figure 4* is the table of sample IDs with their sequencing depths from closed OTU[2] table. As seen in the figure, each of sample ID has different depths. Deleting all the samples which have their depths below 5000, the QIIME 2 algorithm controlled the rest samples' sequencing depth to 5000, thus, we could see the standard species richness and finally available to compare.

Sample ID	Sequence Count
700013549	3,565
700014386	14,608
700014403	9,793
700014409	8,828
700014412	5,355
700014415	8,950
700014418	5,355
700014421	6,444
700014424	4,084
700014427	9,564
700014430	6,099
700014445	3,639
700014488	15,989
700014497	9,163
700014501	8,178
700014515	8,312

Figure 4

The rarefaction curves are necessary for estimating species richness and created by randomly re-sampling the pool of N samples multiple times and then plotting the average number of species found in each sample. Thus, the rarefaction algorithm generates the expected number of species in a small collection of n samples drawn at random from the large pool of N samples. In the Figure 5, the rarefaction curves generally grow rapidly at first, as the most common species are found, but the curves plateau as only the rarest species remain to be sampled.

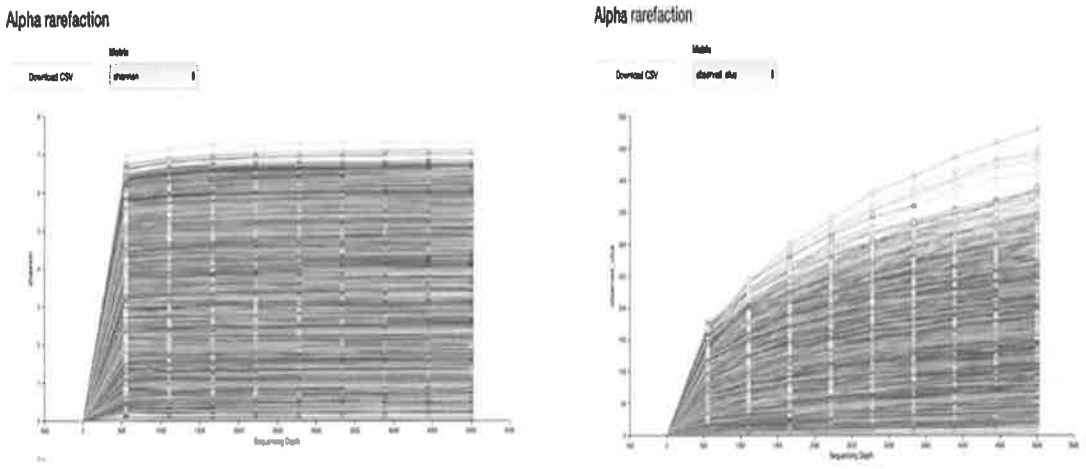

Figure 5 | Alpha rarefaction curve (left: Shannon diversity index[7], right: Observed OTUs [6]) Alpha rarefaction curves as a method to compare the shape of a curve rather than absolute numbers of species

Diversity Tests

Diversity tests are the major part in microbiome[1] research. Many classic statistical tests are available to analyze microbiome[1], and a hypothesis testing in microbial taxa[3–1] can be conducted by comparing alpha diversity[12] and beta diversity[13] indices. In the diversity tests, there are four kinds of alpha diversity[12] indices which are the Pielou's Evenness[5], Observed OTUs[2],[7], Faith phylogenetic diversity (PD)[6], and Shannon diversity index[7] as well as the four kinds of indices for the beta diversity[13] which are the Bray Curtis dissimilarity[8], Jaccard[9], Unweighted UniFrac*[10], and Weighted UniFrac*[11]

In this research, we visualized the alpha diversity[12] with boxplot, bar plot, and tested in Kruskal-Wallis method, and beta diversity[13] was tested with permutation MANOVA and Principle Coordinate Analysis (Classical Multidimensional Scaling).

Alpha Diversity Test

In the *Figure 6*, the boxplot was generated with the Faith PD[6] index for alpha diversity[12]. The x-axis represents the each of body sites, and y-axis represents the diversity in Faith PD[6]. As seen in the figure, the diversities are all different by the body parts. Thus, we can say that there are different alpha diversities[12] by the body parts with Faith PD[6] index.

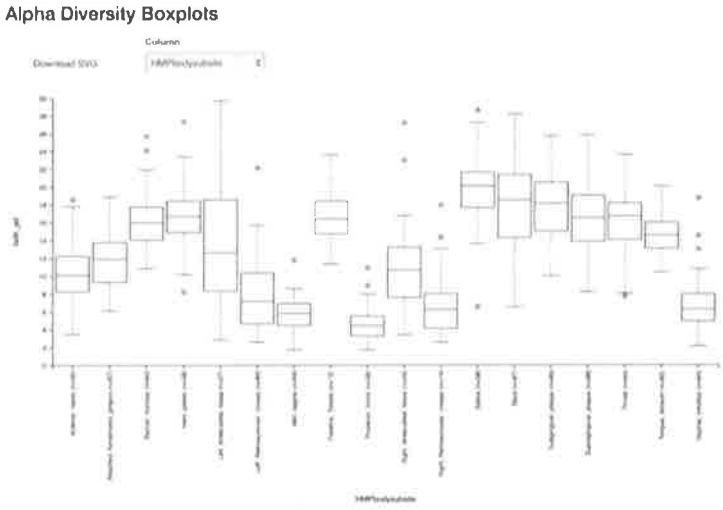

Figure 6, Alpha diversity boxplot | Alpha diversity within subjects by body habitats (x-axis), grouped by area, as measured using the Faith PD of 16S rRNA gene OTUs matched to reference genomes (y-axis).

For testing its significance, the Kruskal-Wallis test was conducted in both ways of all groups and pairwise. The resultant p-value of Kruskal Wallis test for all groups is 1.4567e-12 *(See the 1st in Figure 7)*, which rejects the null hypothesis, the alpha diversity[12] of microbiomes[1] across the body sites is

same, and concludes that the alpha diversity[12] of microbiomes[1] is different with the p-value, 1.4567e-12. In the pairwise Kruskal-Wallis test, there were many of p-values and q-values for each of group 2 members under a member of group 1 *(See the 2nd in Figure 7)*. In the second table in *Figure 7*, the group 1 indicates the compared OTU[2] and the group 2 indicates the comparing OTUs[2]. The p-value in pairwise Kruskal-Wallis tests for the alpha diversities between group 1 and group 2. The q-values in the Figure are important because it determines how much significance the p-value has for going over too much comparisons, thus, the q-values test for the significance of p-values with the null hypothesis, p-value is significant. For the result of both Kruskal Wallis test, we can conclude that the alpha diversities across the body parts are different with the p-value(s).

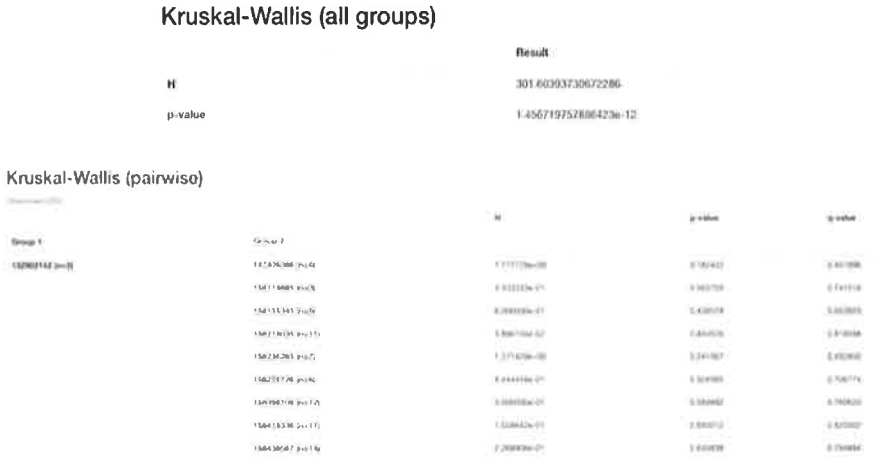

Figure 7, Kruskal-Wallis Tests | Kruskal-Wallis Test was conducted for the reason why the variables are not normally distributed. The Kruskal-Wallis test was reviewed to analyze taxonomic diversity data.

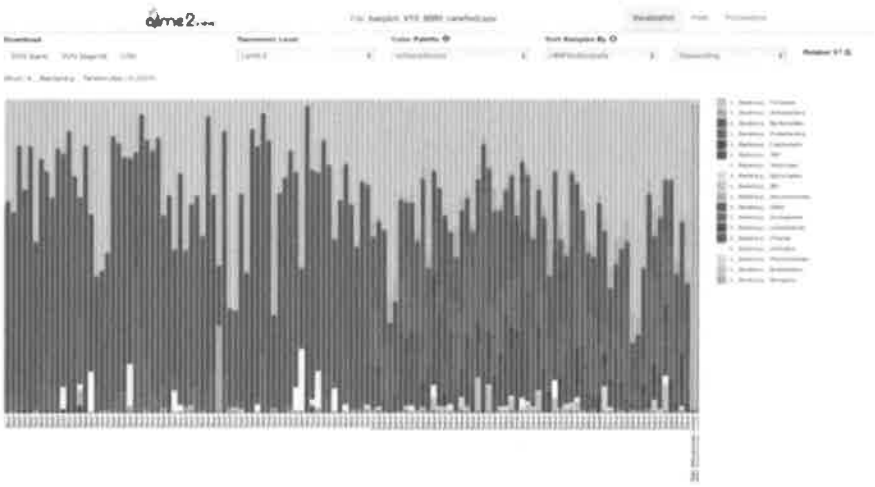

Figure 8, Alpha diversity bar graph | Metagenomic reads from the samples to determine relative abundances for each species at level 2, phylum level.

Beta diversity Test

The data are tested in beta diversity[13] with the hypothesis, the average rank similarity between samples within a group is the same as the average rank similarity between samples belonging to different groups, are same. The *Figure 8* shows that the analysis used the pseudo-F test statistics and results 55.0414 with the p-value of 0.001 for permutation MANOVA which rejects the null hypothesis and conclude that the beta diversities[13] of microbiomes[1] are different with the p-value of 0.001. The PERMANOVA is generally used with one of distance measure methods. In this research, PERMANOVA using Bray-Curtis dissimilarity distance measure was conducted to show the composition of the microbiome[1], to assess the association with beta diversity[13] measures, and to test for microbial divergence among population.

Overview

	PERMANOVA results
method name	PERMANOVA
test statistic name	pseudo-F
sample size	1160
number of groups	18
test statistic	55.0414
p-value	0.001
number of permutations	999

Figure 8 | Permutation MANOVA test and the resultant p-value for the beta dversity

Principle Coordinate Graph

In the practice, we focused on the diversities and run the Principle Coordinate Analysis (Classical MDS) to examine the similarity and dissimilarity of microbial structure by body parts and visualize as much variability as possible, which means the three-dimensional graph.

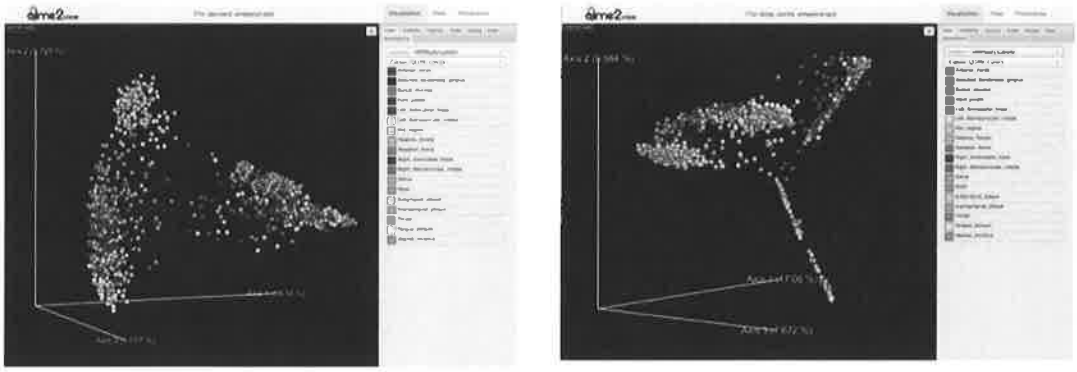

Figure 9, Beta diversity Principle Coordinate plots (left: in Jaccard[9], right: in Bray Curtis dissimilarity[8]) | Principle Coordinates plot showing variation among samples demonstrates that primary clustering is by body area, broadly with the oral, skin, and Stool.

The Principle Coordinate graph shows that the samples are grouped based on body sites which means that the samples from the same body sites have more similar microbial community structures than samples from different body sites *(See the Figure 9)*.

Core Microbiome

With the rarefied data, we combined taxa[3-1] to easily identify which body parts the taxa[3-1] are involved. In the data, there are 16 specified body parts, but as our objective is to identify the microbiomes[1] in the major body parts which are the Stool, Skin, and Oral, we categorized each taxon[3-1] to the major body parts using Python.

First, to know and assign new taxonomic information, we filtered the data against "Attached Keratinized gingiva", "Buccal mucosa", "Hard palate", "Palatine Tonsil", "Saliva", "Supragingival plaque", "Throat", and "Tongue dorsum" for oral part, "Left Retroarticular crease", "Right Retroarticular crease", "Left Antecubital fossa", "Right Antecubital fossa" for skin part, and "Stool" for stool part. Then, I changed the taxonomic information to the corresponding major body parts either Stool, Skin, or Oral. So that the OTU[2] data frame with consisting of only 3 body parts are made.

Identification for the core healthy microbiomes

For the identification for the core healthy microbiomes, we used the R with the libraries, "Phyloseq", and "Microbiome". Before we started, because the "Phyloseq" and "Microbiome" libraries treat the proportions of OTUs[2] of samples, not the frequencies, we had to scale the table for how much proportions the OTUs[2] taking in each sample. Making a function in R, we calculated how many OTUs[2] each sample has, then divided each OTUs[2] of samples by the total frequency of OTUs[2]in the samples, thus, the proportions of each OTU[2] in each sample were calculated and assigned. Referring from exemplary R script from GitHub, the first step was to make phyloseq object, which is the data type for core microbiome analysis in R, with reading the OTU[2] mapping table, metadata, and taxonomy data. With the normalized data, we examined at each detection[15] of 0, 0.01, 0.05, 3, 5, 10, 15, 25, 30 % with each prevalence[14] of 50, 60, 70, 80, 90% and visualized to the heatmaps.

In the analysis part, we investigated the relative population frequencies at different compositional abundance threshold which are the range of detection, then retrieved associated taxa names. By selecting taxonomy of only OTUs[2]that are core members based on thresholds that were used, we finally found the total core abundance in each sample which is sum of abundances of the core members by tested at each detection and prevalence level.

For example of V13 Stool data, after the normalization of the Stool OTU mapping table and making the phyloseq object with the mapping table, metadata, and taxonomy data, we examined the core microbiome at the maximum detection and prevalence to find the core microbiome. As the result of

investigating the relative population frequencies at different compositional abundance threshold in range of detection, the OTU ID, 194909, has the most relative population frequency and when assigning the taxonomic information to the OTU ID, we could know that the OTU is involved in the genus of Bacteroides in the kingdom of bacteria. Therefore, we could conclude that the most core microbiome in 16S amplicon regions of V1 and V3 at the body part, Stool, is most specifically the genus of Bacteroides in the kingdom of Bacteria.

Visualization

In *Figure 10*, we have the heatmaps for the body parts, Stool, Oral, and Skin. These show that the relative patterns of high-abundance OTUs[2]s against a background of taxa[3−1] that area mostly low-abundance or absent in a sparse matrix. We could also see that as the detection[15] level went higher, the prevalence[14] of each taxon[3−1] went lower, so that we could identify the core microbiomes[1] at different detection[15] level with the prevalence[14] of each taxon[3−1]. Thus, the results determine the core microbiome[1]across various abundance/ prevalence[14] thresholds with composition abundances.

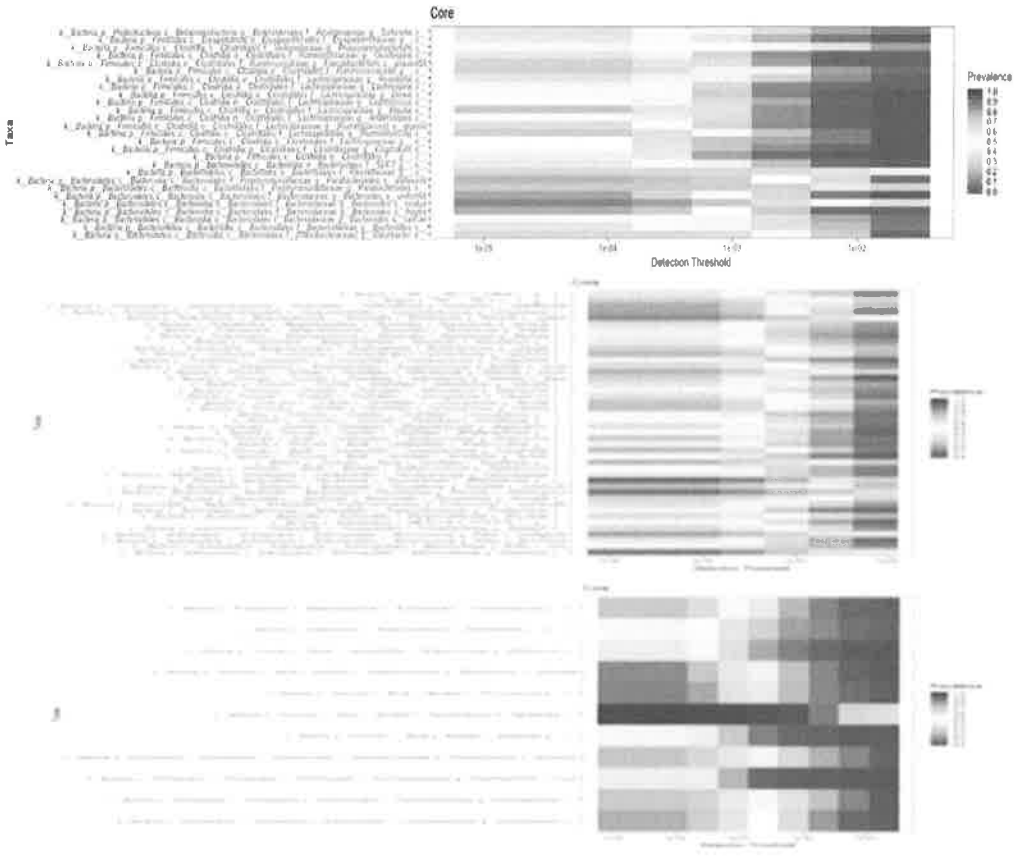

Figure 10, Healthy Microbiome Heatmap (top: Stool, middle: Oral, bottom: Skin) | An ecologically-organized heatmap using ordination methods (classical Multidimensional Scaling and Principle Coordinate Analysis) to organize the rows and columns instead of hierarchical cluster analysis.

Conclusion

Through the internship, I examined the data, which refer to the 16S rRNA survey data sampled from healthy humans at different body sites associated with the taxonomic information. After filtering the mitochondria and chloroplast, uninformative and rare OTUs[2], and alpha rarefaction, I compared samples based on alpha diversity[12] and performed ordination analysis using classical multidimensional scaling and principle coordinate analysis to explore relationships between samples from different body sites based on beta diversity[13]. Furthermore, I determined whether groups of samples by body sites are significantly different from one another based on Kruskal-Wallis and permutation MANOVA test with adjusted p-values. Then, I lastly identified the healthy human core microbiomes[1] by the parts at the OTUs[2] as well as the taxonomy[3] level.

This internship has been a great chance to let me realize how important the statistics and the analysis of data are. Once, I did mistake on interpretation and omit any of procedures, it brought me a totally different and ridiculous results. Through the experiences, I was aware of the importance of statistics and deriving the valid analysis.

References

"QIIME 2 User Documentation¶." *QIIME 2*, docs.qiime2.org/2018.11/.

GitHub, Phyloseq: Analyze microbiome census data using R. Retrieved from https://joey711.github.io/phyloseq/

BLAST: Basic Local Alignment Search Tool. (n.d.). Retrieved from https://blast.ncbi.nlm.nih.gov/Blast.cgi

Rarefaction (ecology). (2017, July 27). Retrieved from https://en.wikipedia.org/wiki/Rarefaction

Hypothesis testing and statistical analysis of microbiome. (2017, June 23). Retrieved from https://www.sciencedirect.com/science/article/pii/S2352304217300351

Terminology

Microbiome[1]: the full collection of genes of all the microbes[1-1] in a community;

Microbe[1-1]: simply a microorganism, especially a bacterium causing disease;

Microbiota[1-2]: a collection or community of microbes;

Operational Taxonomic Unit (OTU)[2]: the unit used to classify groups of closely related individuals;

Taxonomy[3]: commonly known as the study of defining and naming groups of biological organisms based on shard characteristics, and this would be applied to identify which taxonomy the OTU is involved;

Taxon[3-1]: a taxonomic group of any rank, such as a species, family, or class;

Metadata[4]: uniquely defined in the field as the OTU description for the types, formats, and how the data was generated;

Pielou's Evenness[5]: how close in numbers each species in an environment is;

Faith phylogenetic diversity (PD)[6]: a qualitative measure of community richness based on phylogenetic analogue of taxon richness and the sum of the branch lengths of the phylogenetic tree connecting all species found in samples;

Observed OTUs [6] : number of OTUs detected;

Shannon diversity index[7]: a quantitative measure of community richness;

Bray Curtis dissimilarity[8]: non-phylogenetic based method that takes abundance into account;

Jaccard[9]: the number of objects in common/ total number of objects;

Unweighted UniFrac*[10]: sum of branch length that is unique to one environment meaning that it does not take abundance into account;

Weighted UniFrac*[11]: branch lengths are weighted by the relative abundance of sequences.

UniFrac*: a distance metric used for comparing biological communities.

Alpha Diversity[12]: the general diversity of each local species pool

Beta Diversity[13]: the differences in species composition among the sites.

Prevalence[14]: a statistical concept referring to the number of presence in a particular population.

Detection[15]: The limit of detection for any analytical procedure, the point at which analysis is just feasible, may be determined by a statistical approach based on measuring replicate blank samples.