Metagenomic Analysis of the Human Oral Microbiome

Using Illumina MiSeq Sequencing

Adam Hipp

University of North Carolina at Charlotte

***Abstract:***

Illumina Sequencing, a widely used second generation sequencing method, is widely used for examining the composition of bacterial environments. Illumina sequencing is an optimal choice because it can differentiate bacteria that exist in a mixed community and compare their relative concentrations to one another. Oral microbiome metagenomic profiles need about 100,000 reads per sample for adequate analysis. To accomplish this, we use amplicon sequencing of a specific gene that will be conserved among all the bacterium in the community so that we return a consensus of bacteria present. The 16S small ribosomal subunit gene will be the focus because it has regions that are both highly conserved and highly variable, which make it an ideal choice for primer selection. There are 9 regions of high variability in 16S rDNA, but the focus will be on V3 and V4, because the entire genome is too long to complete Illumina sequencing due to limitations on read length and chemistry constraints. These variable regions are flanked with primers so they can be amplified using PCR and then compared to one another to be separated into taxa. The mixed community being analyzed is the human oral microbiome, and the 16S rDNA will be collected from saliva samples submitted by each student in the lab. Once the libraries were created, we used BioAnalyzer to examine and analyze the sequence libraries of each student. After the library size and concentration is validated with BioAnalyzer, Illumina MiSeq 600v3 sequencing is used to build the metagenomic profiles of the bacterium in our sample that exist in the oral microbiome. Illumina offers a cloud-based resource that stores the reads from these sequence libraries and provides tools to validate the data and to analyze the data on a taxonomic level. The data analysis tools and results will serve as the major component for the discussion and results section.

***Introduction:***

The goal of the Illumina sequencing labs conducted during the 2019 Fall semester were to discover what bacterial organisms existed in our oral microbiome. Through Illumina sequencing, we can distinguish numerous bacterium that exist in this mixed community that is our oral microbiome. The process from data collection to data analysis was performed over several laboratory periods and spanned from October 14 to November 18. We began by collecting saliva samples so that DNA from the oral cavity can be isolated and purified. We built our library using PCR amplification of the V3 and V4 regions of the bacterial 16S rDNA with that utilized primers with one end matching the gene region and one end the corresponds with the sequencing technology. Overall, the goal is to amplify the DNA from our saliva samples. Following this step, we cleaned the PCR products, removing unpaired amplicons from stage 1 so they do not compete with stage 2 primers. The stage 2 library is performed using primer sequences that act as a barcode so that each student’s sample can be differentiated. This allows for multiplexing samples in the same MiSeq run, which is a cost-effective strategy that makes this method more appealing. These primer sequences also contain a sequence that allows them to combine with the flow cell surface during the PCR colony, or polony, creation step during Illumina MiSeq 600v3 sequencing. The flow cell can display over 20 million reads across up to 15 lanes. The amplified reads are stored in a cloud-based resource called BaseSpace. These reads stored in BaseSpace provide the data necessary for analysis of the oral microbiome. The tools and programs offered by BaseSpace allow the storage, analysis, and sharing of genetics data that is created by the MiSeq system. We will use a pipeline called BioLockJ for extracting the data into various files. BioLockJ is a metagenomics pipeline for whole genome sequencing and 16S sequence analysis developed in the Fodor lab at University of North Carolina at Charlotte. The files generated by BaseSpace and retrieved using the BioLockJ pipeline will be used to distinguish the different bacterium in the mixed communities of the oral microbiome and determine their relative abundance.

***Materials and Methods:***

***Note*:** *This section will only include summary details and steps from each lab. It will highlight the most important parts of the protocol. A link will be provided under each section to direct you to the full lab protocol that can be accessed through Canvas, a cloud-based learning management system provided by Instructure and used by University of North Carolina at Charlotte.*

**DNA collection (Lab 6):**

* Saliva is collected using Omnigene-ORAL Saliva Collection Kit (OM-501).
* DNA is extracted and purified
  + Proteinase K used to deactivate enzymes, and degrade proteins and RNA.
  + Using isopropyl alcohol, the DNA will precipitate from the saliva.
* [**https://uncc.instructure.com/courses/106953/files/folder/Weekly%20Protocols%20for%20Labs?preview=6768712**](https://uncc.instructure.com/courses/106953/files/folder/Weekly%20Protocols%20for%20Labs?preview=6768712)

**Creating DNA Library Stage I from saliva (Lab 7):**

* Nanodrop used to estimate concentration and purity of DNA from saliva
* PCR using 16S rDNA PCR primers selected from Klindworth et al. (2013)
* Run 1.5% agarose gel electrophoresis to validate PCR reaction products
* Clean PCR 1 products using Ampure XP beads
* [**https://uncc.instructure.com/courses/106953/files/folder/Weekly%20Protocols%20for%20Labs?preview=6870259**](https://uncc.instructure.com/courses/106953/files/folder/Weekly%20Protocols%20for%20Labs?preview=6870259)

**Bar Coding Amplicons – Library Preparation Stage II (Lab 8):**

* Add index to samples using NexteraXT labeling kit
* Clean up PCR2 products with Ampure XP beads
* Quality control library with Nanodrop spectrophotometer and Qubit Fluorometer
* [**https://uncc.instructure.com/courses/106953/files/folder/Weekly%20Protocols%20for%20Labs?preview=6870259**](https://uncc.instructure.com/courses/106953/files/folder/Weekly%20Protocols%20for%20Labs?preview=6870259)
* ***NOTE****: Qubit readings measured either trace amounts of product or no product at all for all samples. Quality control was completed to determine root cause of these error. The PCR master mix was tested for varying concentrations of salt but was determined to not have caused the error. It is probable the primers were ineffective because they were 5 years old, but this could not be definitively proven. A revised protocol for PCR libraries 1 and 2 was created by Dr. Jennifer Weller to recreate PCR libraries 1 and 2.*

**Revised PCR Stage 1 & 2 protocol (Lab 9):**

* Set up 3 PCR reactions (stage 1) using a Master Mix
* Quantify product with Nanodrop
* Set up 6 PCR reactions (stage 2)
  + 3 will run for 8 cycles, 3 will run for 25 cycles
* [**https://uncc.instructure.com/courses/106953/files/folder/Weekly%20Protocols%20for%20Labs?preview=6870259**](https://uncc.instructure.com/courses/106953/files/folder/Weekly%20Protocols%20for%20Labs?preview=6870259)

**Clean Up and Validate Library Size and Concentration (Lab 10):**

* Process and clean up with Ampure XP beads
* Quantify sample with Qubit fluorometer
* Make working dilution to be used for Agilent 2100 chip-based electrophoresis
  + Tests library amplicon size and concentration
* [**https://uncc.instructure.com/courses/106953/files/folder/Weekly%20Protocols%20for%20Labs?preview=6870259**](https://uncc.instructure.com/courses/106953/files/folder/Weekly%20Protocols%20for%20Labs?preview=6870259)

**Illumina MiSeq 600v2 Sequencing (Lab 11):**

* Use Qubit fluorometer to measure concentration of sample from lab 10
  + Ideal concentration is 1.6 ng/ul, equivalent to 4nM
* Pool samples together to prepare for MiSeq system
* [**https://uncc.instructure.com/courses/106953/files/folder/Weekly%20Protocols%20for%20Labs?preview=6870259**](https://uncc.instructure.com/courses/106953/files/folder/Weekly%20Protocols%20for%20Labs?preview=6870259)

**Analysis Lab:**

* Retrieved the sequence reads from BaseSpace
* Merged reads using software called Pear (Paired End Read Merger)
* Classified organisms using BioLockJ pipeline. Computation completed through the university provided HPC cluster Mamba
* [**https://docs.google.com/document/d/1j8SdR5EyO-yDVv4\_3jnmIpAPq67hJeqb2JaVajia56M/edit#**](https://docs.google.com/document/d/1j8SdR5EyO-yDVv4_3jnmIpAPq67hJeqb2JaVajia56M/edit)

**Results**:

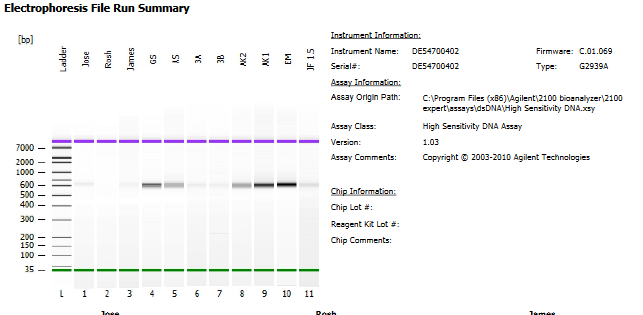


Figure 1: Bioanalyzer results -electrophoresis run

The electrophoresis run from the PCR samples submitted at the end of lab 10. The goal was to submit a 20 µl sample with a DNA concentration at 0.2ng/µl. I confirmed the concentration of my sample to be at 11.8 ng/µl with the Qubit fluorometer, so I diluted 2µl of sample with 116µl of Low TE. The reading from lane 5 shows a band that has shifted closer to 600 bp and is more faint than desired. From the Agilent reading below in Figure 2, the band recorded a concentration around 40 pg/µl, which is roughly 20% the desired concentration as seen in lanes 9 or 10. This error is most likely attributed to the low volume of sample with a high volume of Low TE used to make the dilution. Measurement error is common when such a small amount of sample is to be measured.

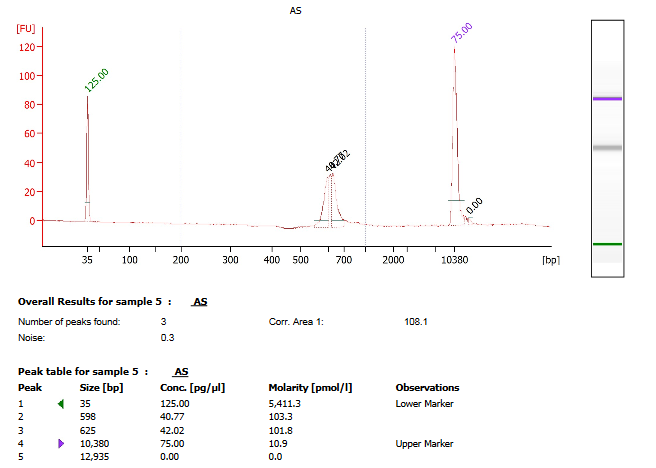


Figure 2: Lane 5 electropherogram summary

**FastQC Analysis:**

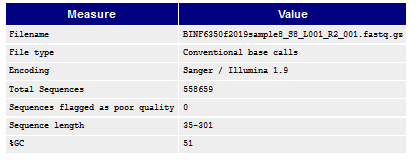
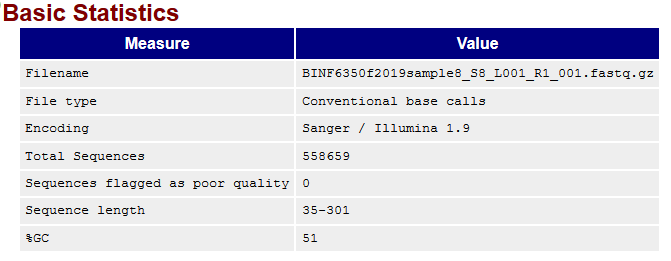


Figure 3: FastQC Statistics for forward read R1 (top) and reverse read R2 (bottom)

1. **Per base sequence quality**: Represented by a box-and-whisker plot with quality score statistics at each position. The blue line shows the mean quality score at each position. The red line contained in the yellow box is represents the median quality score at that position. The yellow box represents the range between 25th and 75th percentile scores. It is common with paired end reads for the average quality score on read 1 to be higher than read 2. Figure 4 shows the reverse read, R2. The quality scores for the first few positions are low, then increases, and then steadily drops over the length of the read as expected. Ideally the slope of the blue line, the mean quality score, would drop more gradually, and the red line, the median quality score at position, would be higher throughout the read.

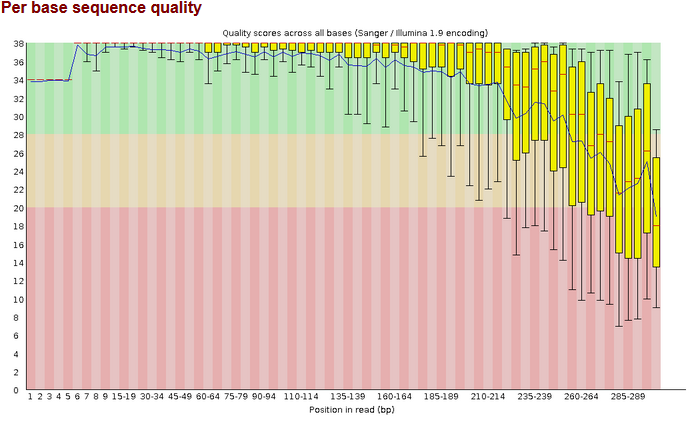


Figure 4: Reverse read, R2, showing an abnormal drop in quality score mean and at each base position

1. **Per base sequence content:** Represents the relative percentage of bases called for each of the 4 nucleotides. That is, the percentage of A calls should be equal to T calls, and percentage of G should equal C. This does not pass a FastQC test. The green and red lines, T and A nucleotides respectively, and black (G) and blue (C) should nearly overlap as these form base pairs in DNA. Figure 5 shows the forward read, R1, and the percentages of each base pair that do not coincide.

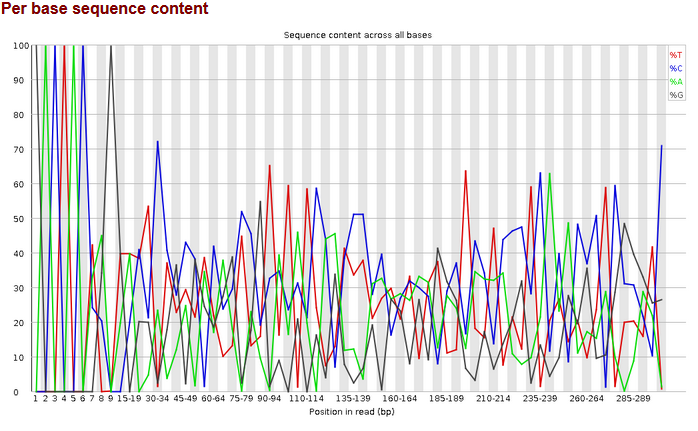


Figure 5:Per base sequence content in forward read, R1

1. **Per sequence GC content:** Plots the number of reads compared to GC%. The expectation is for the GC content to follow a normal bell curve. Reverse read R2 has a similar distribution.

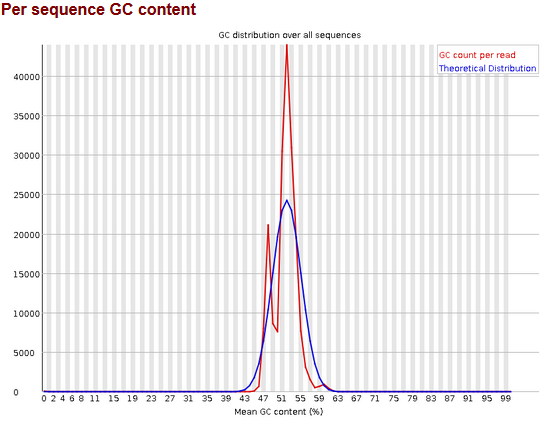


Figure 6:Per sequence GC content on reverse read R2

1. **Overrepresented Sequences:** If a sequence makes up more than 0.1% of the total reads, then it is checked against a list of common contaminants to attempt to recognize the overrepresented sequence. Figures 7 and 8 show overrepresented sequences from both the forward reads and reverse reads respectively. Each sequence is highly similar even though it is designated to a separate bin.

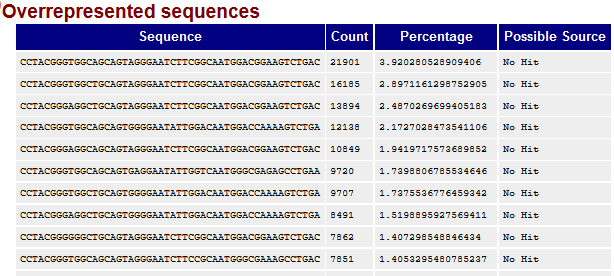


Figure 7:Overrepresented sequences from forward read R1

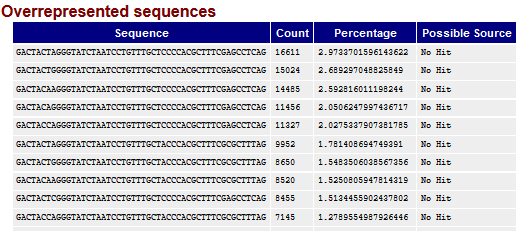


Figure 8: Overrepresented sequences from reverse read R2

**Illumina Sequencing Analysis:**

* Instrument: M03560 - Flowcell ID: 000000000-CG9M8
* Run number: 0071 - Yield – 7.81 Gbp
* Average quality score across all bases – 82% - Polony density **–** 465 +/- 45 thousand per mm
* Percentage of number of clusters that passed Chastity filter – 91% PF

Looking at the sequence statistics, we will be able to correlate the majority of our reads with a high specificity at the genus or species level. Table 1, shown below, compared the number of merged reads with the number of operational taxonomic units that match with the reads. The samples and control have a high hit ratio percentage, with only lane 5 calculated below 95%. Nearly all the merged reads from each sample will be able to be compared to an OTU and provide detail about the characteristics of that species in the oral microbiome.

Table 1:Metadata associated with reads from Illumina MiSeq

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SAMPLE\_ID | Source | Num\_Merged\_Reads | OTU\_COUNT | Hit\_Ratio |
| control | control | 369424 | 360894 | 0.97691 |
| sample1 | saliva | 874754 | 863980 | 0.987683 |
| sample10 | saliva | 1180382 | 1163615 | 0.985795 |
| sample11 | saliva | 938602 | 927128 | 0.987775 |
| sample14 | saliva | 895011 | 877002 | 0.979878 |
| sample2 | saliva | 609466 | 601758 | 0.987353 |
| sample3 | saliva | 1178829 | 1121767 | 0.951594 |
| sample4 | saliva | 980861 | 957532 | 0.976216 |
| sample5 | saliva | 809623 | 767187 | 0.947585 |
| sample7 | saliva | 826021 | 814922 | 0.986563 |
| sample8 | saliva | 556536 | 548876 | 0.986236 |
| sample9 | saliva | 704594 | 684154 | 0.97099 |

The control included 11 different species of bacteria that will be sequenced by the Illumina MiSeq. Each species contributed to 8.3% of the composition of genomic DNA mix that was amplified in the previous lab. I compared the percentage of composition of my sample for each species in the control to determine how much of my sample contributed to the overall number of reads. I also compared my sample composition to the control at the order level. These results are shown below in table 2.

Table 2:Percent composition of sample observed at genus and order compared to control

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Genomic DNA in Control Mixture** | |  |  |  |  |  |  |
| **Genus** | **Species** |  | **% Composition in Control** |  | **% Composition Genus Observed** |  | **% Composition Order Observed** |
| Bacteriodes | fragilis |  | 8.3 |  | 0.00298 |  | 5.226 |
| Bacteriodes | vulgatus |  | 8.3 |  | 0.00298 |  | 5.226 |
| Bifidobacterium | adolescentis |  | 8.3 |  | 0 |  | 0.1871 |
| Clostridiodes | difficile |  | 8.3 |  | 0 |  | 5.43 |
| Enterococcus | faecalis |  | 8.3 |  | 0.194 |  | 8.49 |
| Lactobacillus | plantarum |  | 8.3 |  | 8.61E-05 |  | 8.49 |
| Enterobacter | cloacae |  | 8.3 |  | 0 |  | 0.0278 |
| Escherichia | coli |  | 8.3 |  | 0.0956 |  | 0.0278 |
| Helicobacter | pylori |  | 8.3 |  | 0.00223 |  | 2.17 |
| Salmonella | enterica |  | 8.3 |  | 0.00108 |  | 0.0278 |
| Yersinia | enterocolitica |  | 8.3 |  | 0 |  | 0.0278 |
| Fusobacterium | nucleatum |  | 8.3 |  | 12.49 |  | 10.63 |

Table 3, shown below, shows the species for which my sample recorded a high level of composition at the genus level. The table is indicative of the genus level where my sample comprised at least 20% of the total reads.

Table 3: Abbreviated list of genus hits with highest percentage compared to total reads

|  |  |
| --- | --- |
| **Genus hits with highest percent** |  |
| **Genus** | **Percent Composition** |
| Alloscardovia | 23.932% |
| Bulleidia | 67.647% |
| Butyrivibrio | 40.622% |
| Cardiobacterium | 32.541% |
| Eikenella | 76.397% |
| Filifactor | 21.389% |
| Fretibacterium | 28.557% |
| Gemella | 22.750% |
| Johnsonella | 75.472% |
| Mogibacterium | 44.444% |

Table 4:MDS at genus level

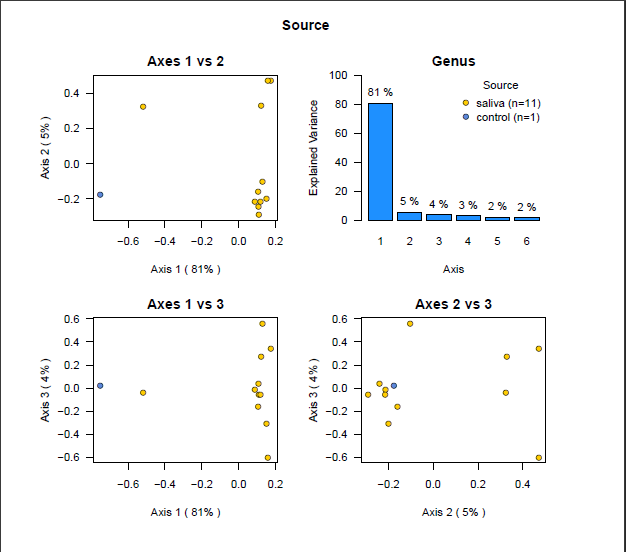
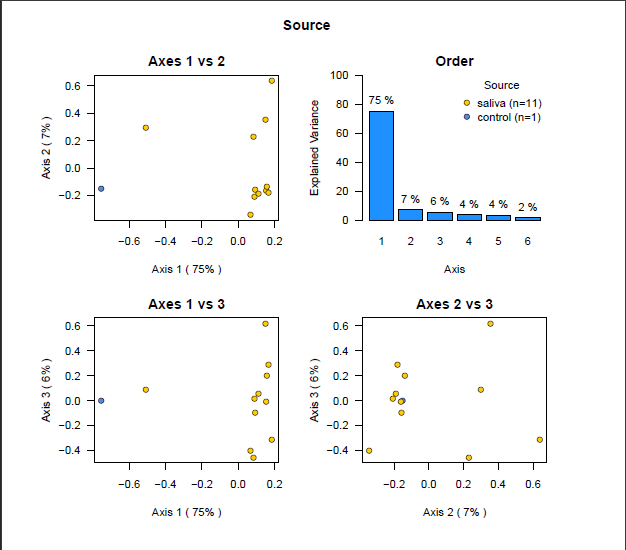


Table 5: MDS plot at order level



***Discussion:***

Illumina MiSeq sequencing is an ideal method for sequencing a mixed community of species to determine species-level identification. We used the 16S rDNA gene as a marker gene to distinguish the existence of bacterium in our oral microbiome. 16S rDNA is a preferred choice because of its highly conserved regions that flank regions of high variability. We targeted the V3 and V4 regions Illumina included notes on the analysis of the read quality of our samples. This provides a overview of the efficacy of the PCR bridge amplification on the flow cell step during sequencing. From these analyses we can determine the samples were sequenced with a high degree of accuracy and efficiency. Illumina shows a %PF of 91%. %PF is a calculation of the percentage of clusters that passed a filter responsible for detecting polony overlap. Also, the quality score calculated across all bases is at 82%, which is above the typical threshold of 70-80%. The total yield is 7.81 Gbp across 15 lanes, so the minimum reads per lane is much greater than the minimum of 100,000 base pairs required to do an effective metagenomic comparison study. Additionally, the polony density is measured at four hundred sixty-five thousand +/- forty-five thousand per millimeter squared. This is a acceptable measurement because the desired density is at five hundred thousand. Given these measurements it would be safe to assume that these samples can provide some meaningful and accurate assumptions about the species-level composition of bacterium in our oral microbiome.

Table 1 shows the number of merged reads that were created by PEAR. We can see each lane easily exceeds the one hundred thousand read minimum, with the control having the fewest reads at over three hundred thousand. Operational taxonomic units (OTU) serve to group closely related species. Table 1 shows very high percentages, between 97-98%, across all samples of OTU hits compared to the merged reads. This signifies that roughly 97% of the merged reads were able to be matched with an OTU which gives us a high sensitivity to recognize members of the mixed community at the species level. Table 2 shows

Table 3 shows the genus level members that exist in my oral microbiome that had the highest percentage composition compared to all the samples. The table below provides more detail about each of these and their significance found. I attempted to include any pathophysiology associated with that organism to draw a meaningful conclusion. Butyrivibrio is necessary for degraded plant based carbohydrates. Filifactor and Fretibacterium are associated with periodontal disease and subgingival plaque respectively. These characteristics make sure, but provides motivation to brush and floss more effectively.

|  |  |  |  |
| --- | --- | --- | --- |
| **Genus hits with highest percent** |  |  |  |
| **Genus** |  |  | **Significance** |
| Alloscardovia |  |  | Isolated from blood, the oral cavity, a tonsil specimen; actual clinical significance in these cases is not clear |
| Bulleidia |  |  |  |
| Butyrivibrio |  |  | These include fibre degradation, protein breakdown, biohydrogenation of lipids and the production of microbial inhibitorstion |
| Cardiobacterium |  |  | It may also rarely cause endocarditis, an infection of the heart valves |
| Eikenella |  |  | Sole isolate in cases of persistent empyemas and/or overwhelming pneumonias with bacteremias |
| Filifactor |  |  | Associated with periodontal disease and endodontic lesions. |
| Fretibacterium |  |  | Isolated from subgingival plaque |
| Gemella |  |  | Found in bronchial secretions and in mucus from the respiratory tract |
| Johnsonella |  |  | Cccurs in the gingival crevice of humans |
| Mogibacterium |  |  | Isolated from the Periodontal pocket of a human. |
| Moryella |  |  | Important opportunistic pathogens capable of causing a wide variety of nosocomial infections |
| Mycoplasma |  |  | Contagious respiratory infection that spreads easily through contact with respiratory fluids |
| Peptococcus |  |  | Infections of the oropharynx, lower respiratory tract, female genital tract, abdomen, skin, and soft tissues |
| Peptostreptococcaceae\_incertae\_sedis |  |  |  |
| Peptostreptococcus |  |  | Normal inhabitant of the healthy lower reproductive tract of women |
| Shuttleworthia |  |  | Shuttleworthia satelles has been isolated from the human periodontal pocket |
| Tannerella |  |  | Been implicated in periodontal diseases and is a member of the red complex of periodontal pathogens. |
| Tessaracoccus |  |  |  |
| Unclassified Anaerolineaceae Family |  |  |  |
| Unclassified Cardiobacteriaceae Family |  |  |  |
| Unclassified Spirochaetaceae Family |  |  | syphilis, Lyme disease, relapsing fever, and other illnesses |
| Unclassified Spirochaetales Order |  |  |  |

The multidimensional plot in tables 4 and 5 shows a strong correlation of 81% at the genus level and 79% at the order level in axes 1 and 2, but with two clusters of samples. Axes 1 and 3 show a show a strong correlation of 81% at the genus level and 75% at the order level. There is one obvious outlier other than the control. Axes 2 and 3 shows very little correlation at 5% for genus, and 7% for order level, which signifies that the removal of the outlier shows that the correlation is not as strong from the clustered samples.

References:

http://nextgen.mgh.harvard.edu/IlluminaChemistry.html

https://en.wikipedia.org/wiki/Illumina\_dye\_sequncing

https://github.com/msioda/BioLockJ

https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/

https://uncc.instructure.com/courses/106953/files/folder/Weekly%20Protocols%20for%20Labs?preview=6870259

Lab protocols provided by Dr. Jennifer Weller. Stored in Canvas through UNCC. Accessed December, 2019.

Mizrahi-Man, O., Davenport, E. R., & Gilad, Y. (2013). Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: evaluation of effective study designs. *PloS one*, *8*(1), e53608. doi:10.1371/journal.pone.0053608