Amplicon Sequencing Project

Question: Does the female and male sex of humans house a community of bacteria that are significantly different in regard to diversity, richness, and abundance on human hands?

* Difference in sequence reads
* Ordination
* Quality score

Introduction with 5 peer reviewed articles

Methodology to estimate microbial diversity has significantly increased with an effort to understand the distribution and diversity of microbes in their natural environments. Cultivation-dependent methodology work to mimic the natural environment of microbes in a laboratory setting, with enrichment cultures being particularly important for visualizing species that favor the selected media for growth (Swanson, 2016). Different ecological metrics have been used to categorize these cultivated colonies in order to deduce community data about known bacterial species and potentially deduce information about unknown species (Birtel, 2015). Understanding the cultivation methodology to successfully grow microbial species lends information about the requirements imperative for biodiverse communities that can be beneficial in studying niche environments like the human gut which interact with pathogens and probiotics (Wu, 2012). Beyond healthcare, general dispersal of bacterial communities increases biodiversity in water sources, soils, and all microenvironments that effect the wellbeing of the larger ecosystem, with this assumption also holding true for bacterial organisms which reside on a human host or human host environment (Fuhrman, 1999).

Within the paper *Forensic identification using skin bacterial communities,* published in 2010 by Noah Fierer the diversity in human skin bacterial communities is explained to be “far higher than previously recognized, with a high degree of inter-individual variability in the composition of bacterial communities.” This discovery leads the paper to discuss that these individualized communities can be utilized as personal identifiers for criminal cases as forensic evidence. Fierer et al. specifically claims that “these skin bacteria may persist on touched surfaces for prolonged periods because many are highly resistant to environ- mental stresses, including moisture, temperature, and UV radiation,” meaning the community samples can be collected long after the host has left. Though the implications for forensic research are still under scrutiny, the data can be more broadly analyzed to deduce if there are specific differences in bacterial communities between female and male participants. By analytically evaluating these communities separately to find patterns in abundance, richness, and diversity we deduce information about the specific community composition, which can yield information about the overall biodiversity of the niche the bacteria are a part of. The inference that human hosts hold their own unique sets of bacteria helps foster the notion of unique microenvironments that endosymbiotically regulate the human body.

By explicitly looking at community differences between male and female participants information can be gathered about the overall health and well-being of each participant sex in regard to the surrounding environment. In a study published in *Science* in 2013, it was discovered that the gut microbiota is extremely relevant in determining autoimmune disease susceptibility (Markle, 2013). Markle et al. stated that when male cecal contents were transferred to female mice these mice received a higher level of protection against “pancreatic islet inflammation, autoantibody production, and the development of diabetes,” meaning this difference within the sexes microbiota was translational to inhibit the spread of disease (Markle et al). Therefore, understanding the differences in the skin microbiota could also potentially pose interesting medical questions that could yield insight on diseases current to specific, sexually biased traits.

Overall, divulging community data analysis from female and male participants, from the sequenced data sets from the Fierer et. al paper, can provide useful information about the resident skin microbiota of sexually diverse humans. By using sex as the means of a comparison further hypotheses can be made that explain differences based on hormonal differences and social differences since the presiding environment remains constant for both sample sets. We can thereby infer that because of these differences a community the female and male sex of humans house a community of bacteria that are significantly different in regard to diversity, richness, and abundance on human hands. We test this hypothesis utilizing computational amplicon sequencing analyses that parse data particular to the bacterial communities found within the Fierer et al paper.

Difference between females and males

Presiding Question and rational hypothesis

Birtel J, Walser JC, Pichon S, Bürgmann H, Matthews B (2015) Estimating Bacterial Diversity for Ecological Studies: Methods, Metrics, and Assumptions. PLOS ONE 10(4): e0125356. doi: 10.1371/journal.pone.0125356

Fuhrman, J. A. (1999). Marine viruses and their biogeochemical and ecological effects. London.

Swanson, M., Reguera, G., Schaechter, M., & Neidhardt, F. C. (2016). Microbe. Washington, DC: ASM Press.

Wu S, Wang G, Angert ER, Wang W, Li W, et al. (2012) Composition, Diversity, and Origin of the Bacterial Community in Grass Carp Intestine. PLOS ONE 7(2): e30440. doi: 10.1371/journal.pone.0030440

Methods

**Sample Origin and Sequencing**

*Sample Collection*

All analysis was done using Fierer et al. as a reference source for sequence data and sample origins. Within the paper it is stated that samples were taken from nine healthy adults (four female, five male) who worked within the same building at the University of Colorado between ages 18-40. The samples taken included computer mouse swabs touched by the owner 12h before swabbing and the palm of the individual's dominant hand used to operate the mouse. Individuals were required to maintain normal hygiene habits prior to the swab as to not increase variation. The swabs were taken using autoclave cotton-tipped swabs that were autoclave and pre-moistened with a sterile solution. The swabbing took place for 10s per sample taken with all samples being kept at -80C for less than a week before DNA extraction (Fierer et al).

*DNA Extraction and Pyrosequencing*

Fierer et al. explains that for each sample the 16s rRNA genes were amplified using the MO BIO PowerSoil DNA Isolation kit with the broken, frozen cotton swabs. These tubes were then horizontally shaken and kit procedures were followed for extraction. PCR reactions were carried out in triplicate repeats using HotMAsterMic with thermal cycling at 94C for 3min followed by 35 cycles of denaturation, annealing for 30s, and extension and final extension procedures. Replicate amplicons were pooled in agarose gel using SYBR Safe DNA gel stain from Invitrogen. Amplicon DNA concentrations were then measured with the final pool of DNA being precipitated, centrifuged, and centrifuged to create a pellet that was re-suspended in nuclease-free water. Pyrosequencing was then carried out on a 454 Life Sciences Genome Sequencer FLX instrument at the University of South Carolina (Fierer et al).

In order to analyze the sequences received we used computational programs to evaluate variables of interest. The variables we want to analyze more specifically include participant sex and commonly found bacterial community strains within the metadata set.

**Computational**

To create a meta data set with all of the vectors of interest we initially utilized the raw data sets collected from the 454 Sequencer and published by Fierer et al. This was done using a remote server and fastq-dump to download the list of files in the run table to the raw directory. QC reports were then created for each of the runs utilizing fastqc and outputted as HTML to be readable. The fastqc files were trimmed (Trimmomatic Tool) based on their quality scores to discard sequences under 150 base pairs and cut off reads when the base score drops below 25. Fastqc files were then converted to fasta files to be used as BLAST queries (Bioawk). These files were then run against NCBI database (nt), using BLAST (blastn) to output csv file which was based on the name output and included all of the sequence data, plus the bacterial communities each sequence matched to with their quality scores and additional information.

Then the Blast results were further formatted to create an output dataframe with proper column headers that denote each variable of interest from the results (qseqid). After this, the query sequence ID was split into Sample and Number components to group the files into columns (tidyr and separate) to make a vector of all the BLAST output file names, including the names of the directories. An empty matrix was created to join the data back and read in the meta data (rbind).

We than use dplyr piping syntax to select a subset of row matching each vector we want to filter (categorically) and pipe out (numerically) to create histograms. We also used ggplot syntax as a comparative feature to plot multiple categorical variables against one another using one primary numerical variable. Tables were created with kable to tabularize quantified data sets and visualize the most common bacterial communities found for singular categorical vectors (males and females).

To create a meta data set with all of the vectors of interest we initially utilized the raw data sets collected from the 454 Sequencer and published by Fierer et al. This was done using a dada pipeline within R. We ordered the samples first, then extracted the sample names from their fastq format, which was initially done using a remote server and fastq-dump to download the list of files in the run table to the raw directory. QC reports were then created for each of the runs utilizing fastqc and outputted as HTML to be readable. We then plotted the quality of each of the twenty samples of interest into a readable format in order to deduce the length to trim. It was found that quality is reduced after 200 bases, so the maximum acceptable length was made to be 225 bases. Using the dada pipeline the sequences were filtered, trimmed, into a new output folder and allowed to have 3 expected errors. A table of read counts was formatted to visualize the reads before and after filters and then again to visualize error trends. Duplicated sequences were then removed and dada was run on the reads based on 454 data recommendations. The sequences were aligned to craft a site by species matrix and a histogram representation of sequence lengths from all samples. Chimeras were removed and a singular table to give all pipeline information of the sequences trimmed and edited was created. A taxa code was initiated to yield a table with the taxonomy of each individual sequence to create a phylogeny that expresses the overall relatedness of each sequence. All of these tables were crafted in the dada pipeline through the sex of both male and female sample sets to visualize relatedness among all samples tested.

Once the data set was compiled by relatedness pyroseq organized all aspects of the data into a merged meta data set. This data set was parsed to remove any non-applicable data regarding the sex of participants (i.e. all samples which included swabs from electronic devices). The data was then melted to include all taxonomically related data sets that were coded for under the taxa file, separate from the metadata that was read in. This allowed for analysis via tables, figures, and ggplot graphs.

Dplyr and ggplot packages were used to analyze the data through representative figures and tables. The figures used look at abundance, richness, diversity metrics (applied through Shannon), a phylogeny of the entire community set, and an ordination all of which are separated by the respective sex to divulge sex based community analysis on the bacterial species present.

Results

**Set-up**

1. Quality plots: visualize quality of initial reads received from sequencing company
   1. Bad quality after 200 bases meaning acceptable sequence length = 225
2. Filter reads via filter and trim function in Dada2 package
3. Visualize reads before/after trimming (note weird discrepancies?)
4. Visual plot of errors within each sample to see trend of decreasing line
5. Remove duplicated sequences
6. Align sequences based on 454 Roche parameters (dada)
7. Site by species matrix created
8. Histogram of sequence lengths that are denoised and trimmed
9. Remove chimeras
10. Table of fixed reads
11. See assigned taxonomy; bacteria versus cyanobacteria./proteobacter/ bacteriodetes/actinobacteria,firmicutes, parcubacteria/ etc.
12. Output the denoised variants to build a phylogeny
13. Create phylogeny via a maximum likelihood tree in dada
14. 14. Phylo seq plot look at male and female components of diversity

*To-Do*

1. test abundance of bacteria on both male and female samples
   1. calculate, boxplot, table of values
2. test richness of bacteria on both male and female samples
   1. calculate boxplot, table of values
3. t.test for differences in abundance/richness based on male and female samples
4. scatter plot with regression line for abundance/richness versus diversity for female and male participants
5. ordination to order species along presumed ecological gradient to identify patterns of species distribution and abundance (male versus female)

Discussion

**Abundance**

Abundance is a measure of the total number of individuals residing within the same niche environment to statistically observe how consecrated the region is. If a high abundance occurs in one environment it can be observed as having more resources, or less environmentally de-stabling associations.

**Richness**

Richness is a measure of different morphotypes within a bacterial community. Calculations of richness within a community are imperative to showcase if only one species is present or many within the same niche environment. Biodiversity is measured through both richness and diversity factors. By observing biodiversity within communities, conservation of certain species can occur through the creation of more favorable niches. In the observation between male and female resident sample populations like richness is observed in the overlap of some varying morphotypes, however, due to the ability for potential mutation or nutritional needs of the bacteria, some morphotypes favored the males over the female’s environment and vice versa.

**Diversity**

Diversity is a measure of the relative variability among organisms from all sources or morphologies. Shannon Diversity offers a statistical means of categorizing the diversity of individuals within a community

*Abundance versus Diversity*

Male:

Female

*Richness Versus Diversity*

Male:

Female:

**Ordination**

Ordination can be used as an exploratory measure for data clustering of similar and different multivariable types

**Conclusion**