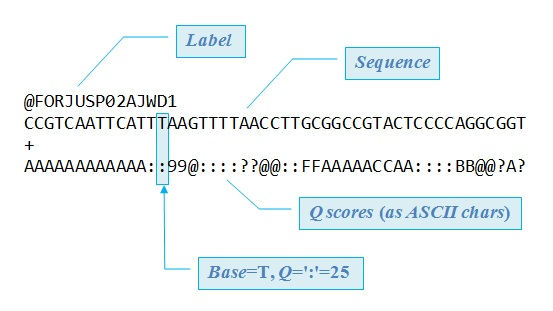
Brief overview of bioinformatic analyses

Disclaimer: The following guide is meant as a general overview of at what we do to analyze metagenomic sequencing samples. We attempt to keep up with the latest in “best practices”, but these analytic tools continue to improve and the correct combination of tools and statistical methods is in perpetual debate. Please let us know if you have any questions, require clarification, or want to help us improve our methods!

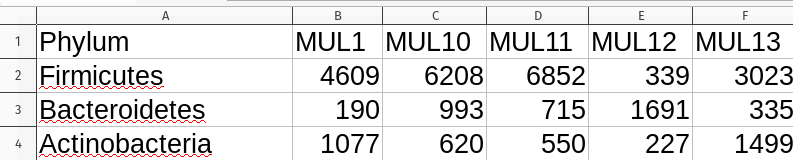
Email: [e](mailto:edoster@colostate.edu)nriquedoster@gmail.com

1. Starting with sample collection
   1. DNA extraction
   2. Library preparation
   3. Metagenomic sequencing
      1. DNA samples are placed on the sequencer and the output is in “fastq” format
         1. These samples can be compressed to decrease their file size. Sample names can be unzipped “.fastq” or zipped “.fastq.gz”
         2. Below is an example of a sequence (read) in a sample consisting of 4 lines:
            1. Sample read
            2. Sample sequence
            3. “+” left over from previous fastq versions
            4. Quality (Q) score for each base in “Phred33 format”
         3. Depending on the sequencing method, samples can consist of two separate files for “forward” and “reverse” reads. Forward reads can end with “.R1.fastq.gz” and reverse reads with “R2.fastq.gz”.
         4. Typically, the sequencing laboratory includes a file with sequencing results
            1. Check this file to summarize average read quality (column: Mean Quality Score (PF)” and total raw reads (column: # Reads) for each sample.

“190513\_M01247\_0347\_000000000-CCMKD\_Morley\_demux.csv”

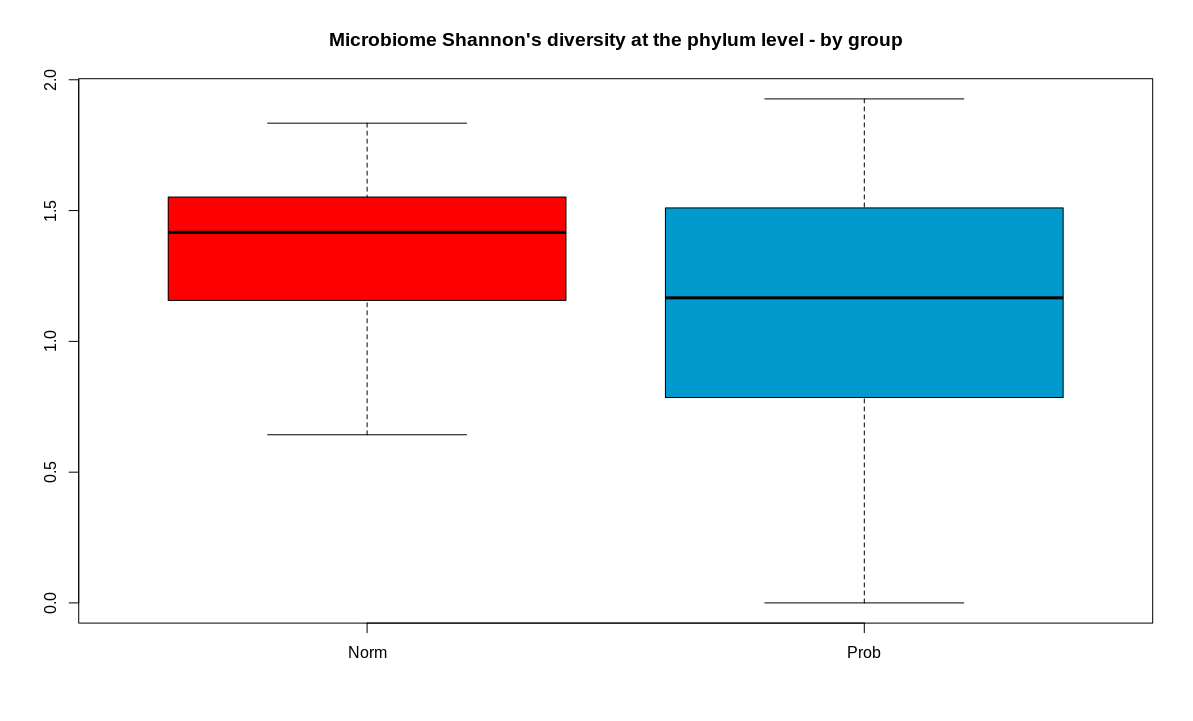
* + - * 1. **These values for raw reads and mean phred scores can be reported using summary statistics (e.g. totals, mean, median, and standard deviation). Statistical comparisons can be made between sample groups using the wilcoxon test or generalized linear models (glm function in R).**

1. Bioinformatic analyses
   1. The goal with bioinformatic analysis is to process sample reads computationally and identify (classify) which important genetic features are present in our samples such as bacterial taxa, antimicrobial resistance genes, or other genes of interest. There are many ways achieve this goal and we currently use 2 different combination of bioinformatic tools (pipelines). The final output file of interest is a count table containing the number of reads classified to all the different features (rows) in each sample (columns).
   2. For 16S sequencing and microbiome characterization:
      1. We utilize Qiime2 (<https://qiime2.org/>) up to the taxonomic classification of all sample reads and then export the results for statistical analysis using the R programming software.
         1. Check the file named “microbiome\_Phylum\_Raw.csv” in the microbiome\_matrices folder.

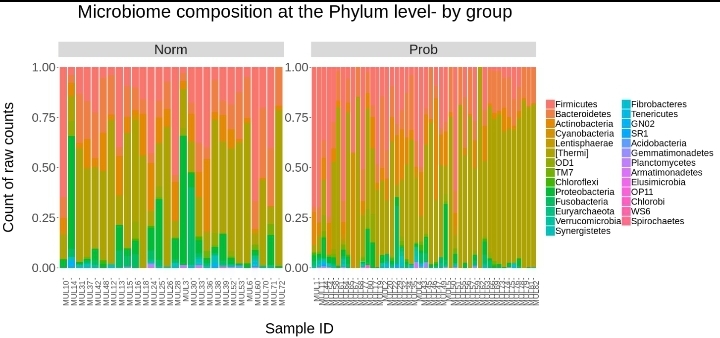


* + - 1. **These counts can be summarized by comparing the total count of classified reads between sample groups. We cannot use these values as quantitative measures of bacterial abundance for specific taxa, but you can use wilcoxon tests or generalized linear models to compare total counts of classified reads.**
  1. For shotgun metagenomic sequencing we characterize the microbiome and resistome using:
     1. A pipeline developed by our research group, Microbial Ecology Group, called AMRplusplus <http://megares.meglab.org/amrplusplus/latest/html/>. Final output files are statistically analyzed using the R programming software.

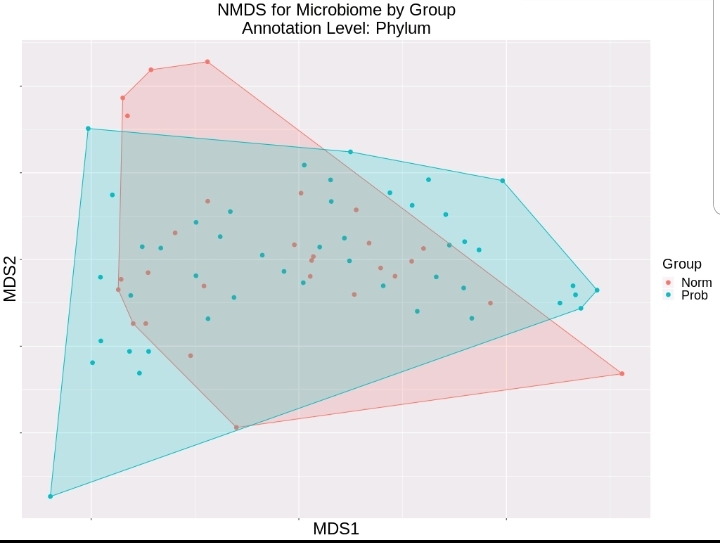
1. Statistical analysis
   1. Statistical analysis is undoubtedly the most complex component required for characterizing metagenomic sequencing samples. We adopt a lot of the techniques developed for ecology to analyze the multivariate data representing the microbiome and resistome. We use the R programming software with various R software packages and you can see the main code we use at this repository (<https://github.com/EnriqueDoster/MEG_R_metagenomic_analysis>).
   2. The count of classified reads will first need to normalized to account for differences in sequencing performance between samples. There are many ways to normalize counts and we use “Cumulative sum scaling”(CSS) developed by Paulson et al 2013 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4010126/>).
   3. Then, the CSS-normalized counts are used for all the following methods of summarizing metagenomic samples (and check out this helpful website <https://mb3is.megx.net/gustame/home>):
      1. Diversity indices like “Richness” and “Shannon’s index” to summarize the microbiome/resistome with a single value that represents the unique number of features and how evenly distributed the counts were, respectively.
         1. Check out the file named MUL\_metadata\_plus\_diversity.csv to see richness and shannons’s diversity values for each taxonomic level. For the purposes of this analysis, we suggest to stay to the higher taxonomic levels like Phylum, Class, and Order.
         2. **These diversity values can be reported using summary statistics (e.g. totals, mean, median, and standard deviation). Statistical comparisons can be made between sample groups using the wilcoxon test or generalized linear models (glm function in R).**
         3. **Check out the figures named “Microbiome\_richness\_phylum\_bygroup.png” and “Microbiome\_shannon\_phylum\_bygroup.png”. Here’s an example of a barplot:**



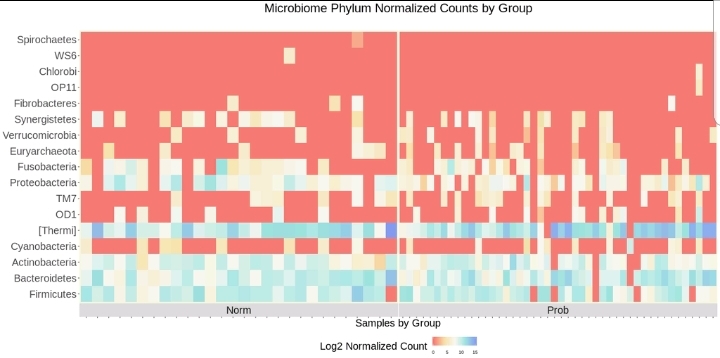
* + 1. We can also look at the composition of the microbiome and resistome by looking at the relative abundance of each feature per sample.
       1. Take a look at the file, “microbiome\_composition\_phylum.png” to see the relative abundance of each Phylum in each sample. You can make your own figures if needed by looking at the “normalized” count table.
       2. **The relative abundance of features can be reported using summary statistics to get a broad overview of the most abundant features. For example, we might report which features make up the majority of reads found across all samples and specify what percentage they represent.**



* + 1. However, comparing relative abundances is troublesome because of issues inherent to compositional data (ie. the proportion of a feature is directly affected by changes in proportion of other features) and because count tables contain many zeroes (sparse).
       1. Therefore, we use a “Zero-inflated Gaussian model” (ZIG) that allows us to combine two different distributions and fit a model that better represents metagenomic count data. Log fold changes in abundance are estimated and the Benjamini-Hochberg method is employed to adjust p-values for multiple testing.
          1. Take a look at the “stats” directory and look for the file named “Group\_Microbiome\_Phylum\_GroupNorm-GroupProb\_Model\_Contrasts.csv”.
          2. **These results have to be interpreted carefully because low abundance features will often be significantly different between sample groups. In combination with your knowledge of which features are most abundant in your samples, pick a threshold value for “Avg. Expression” and report everything above that threshold. For example, you might say something like out of 27 compared, X phyla had average expressions > 1 and X were significantly different between sample groups (p-value < 0.05).**
  1. While the ZIG model allows for the comparison of specific features, we use a different method (ordination) to reduce the dimensionality of our metagenomic samples and compare the microbiome/resistome composition between sample groups. We recommend a non-parametric ordination method called non-metric multidimensional scaling (NMDS). In the figure below, each sample is represented by a single point and the distance between points signifies how similar or different samples are from one another. We then make a polygon that covers all samples in a sample group and allows us to visualize the amount of overlap between sample groups.
     1. Take a look at the file “NMDS\_Group\_Phylum.png” for the example shown below.
     2. **We can test the probability that samples are more different between treatment groups than within. We use a permutation taste named Analysis of Similarity (ANOSIM) to calculate a p-value.**



1. Heatmaps allow the visual comparison of log-normalized counts for features and can include a lot of information in a single figure. Heatmaps serve as exploratory figures and cannot be used for statistical comparison. Each sample is a column of the heatmap, features are rows and the color of each cell represents the log-normalized abundance of each feature.
   1. Take a look at the file named “Phylum\_Group\_Heatmap.png” for an example of a heatmap for phylum-level log-normalized counts compared between sample groups.
   2. **Overall, we haven’t really used heatmaps in our papers as a main figure, but we have included them as supplementary files as they allow a broad overview of which features are abundant across all samples.**

****