Workflow for Final Project:

-only comparing 1 metro station and 1 hospital

1. Download raw sequencing data from SRP101374 link ­-> <https://www.ncbi.nlm.nih.gov/sra/?term=SRP101374>

a. NCBI

b. Hospital 1 and Station 1 files

c. How to download?

Justin

2. Sequenced reads were processed with a quality control step to remove the adapter regions and low quality reads, and mapped to human genome to filter out contaminations by following the previously described steps ([Li X. et al., 2017](https://www.frontiersin.org/articles/10.3389/fmicb.2017.00632/full#B38)).

Joel, Justin weekend of 11/17/17

Start with fastq files, fastq files and use Mothur

3. The high-quality reads were mapped to nr database by DIAMOND ([Buchfink et al., 2015](https://www.frontiersin.org/articles/10.3389/fmicb.2017.00632/full" \l "B10)) with the default setting.

4. Lowest common ancestor (LCA) algorithm was implemented with LCA mapper from mtools of MEGAN5 ([Huson et al., 2011](https://www.frontiersin.org/articles/10.3389/fmicb.2017.00632/full" \l "B26)) for taxonomy profiling of each read.

5. The reads mapped to eukaryotes were removed for the further analysis

6. The relative abundance for each species was distilled from the LCA results using an *in-house* script.

-will need to figure out a script for determining relative abundance

- http://deneflab.github.io/MicrobeMiseq/demos/mothur\_2\_phyloseq.html (might be a good option for this)

7. The bacterial species were queried against PATRIC ([Wattam et al., 2014](https://www.frontiersin.org/articles/10.3389/fmicb.2017.00632/full" \l "B63)) for categorization. For potential pathogenic species annotation, three lists of potential pathogens ([Kembel et al., 2012](https://www.frontiersin.org/articles/10.3389/fmicb.2017.00632/full" \l "B30); [Forsberg et al., 2014](https://www.frontiersin.org/articles/10.3389/fmicb.2017.00632/full#B20); [Wattam et al., 2014](https://www.frontiersin.org/articles/10.3389/fmicb.2017.00632/full#B63)) were combined. Maybe skipable, because it depends if we want to do figures that have potential pathogenic species.

8. For 16S rRNA gene extraction, the filtered reads were mapped against ribosomal RNA SILVA (database where they compared 16s’s) reference sequences (SSURef\_NR99\_115; [Quast et al., 2013](https://www.frontiersin.org/articles/10.3389/fmicb.2017.00632/full#B54)) using assign\_taxonomy.py (python) from QIIME ([Caporaso et al., 2010](https://www.frontiersin.org/articles/10.3389/fmicb.2017.00632/full" \l "B12)) with blast as the assignment method.

9. Shannon alpha-diversity of each shotgun-sequencing sample was calculated with VEGAN (R) ([Dixon, 2009](https://www.frontiersin.org/articles/10.3389/fmicb.2017.00632/full#B17)) based on the relative abundance of each species from 1 M subsampled reads.

10. Weighted UniFrac ® distance was calculated PhylosEq ® ([McMurdie and Holmes, 2013](https://www.frontiersin.org/articles/10.3389/fmicb.2017.00632/full" \l "B44)) based on the species-level taxonomic profile.

Only if we can get data from other studies “pre-analyzed OTUs”

11. To evaluate the community dissimilarities between 16S rRNA samples, Bray-Curtis dissimilarity ® (type of similarity index, gives you a value how similar is: can be done with package ® phylosEq) was calculated based on the relative abundance of each genus

12. Wilcoxon rank sum test on R

13. To detect significant items such as genera or species, analysis of similarities (ANOSIM) was implemented using QIIME ([Caporaso et al., 2010](https://www.frontiersin.org/articles/10.3389/fmicb.2017.00632/full" \l "B12)). Also, adonis (QIIME) was employed to evaluate the significance of a variable in determining variation of distances.

Our analyses:

1. Rarefaction curve YES

2. Heatmap with relative abundances YES

3. What else????

Project due:

Meeting Dates:

Monday 13th, NOV, 1-2pm Hopefully with Tracy (by then we should know our jobs?? Plan the rest)

Friday 17th, NOV, 5:30pm, Justin’s place

Sunday 26th, NOV, 5pm Justin’s place