# MGTST-Outline

*Nate Olson*

*2016-07-08*

**Objectives**

- Provide a detailed description of the dataset and qa/qc methods used to validate the use of the dataset for evaluating 16S metagenomic pipelines and differential abundance detection methods

- Demonstrate how the dataset is used to evaluate the performance of different pipelines and differential abundance methods

- Provide an R package to facilitate using the dataset for evaluating pipelines and differential abundance detection methods

## Abstract

## Background

## Methods

### Experimental design

### Sample selection

### Sequencing

### Sequence processing

### Data analysis

## Results

### Sample Selection

### wetlab QC

### seq data QA and sequence processing

focus is on characterizing and validating the data, highlight the quality of the data and study

- Sample selection

- Wetlab QC
  - sample concentration results summary
  - qPCR

- ERCC
- bacterial quant (http://www.zymoresearch.com/dna/dna-analysis/femto-bacterial-dna-quantification-kit)

- Seq data QA
  - number of reads
  - read length distributions
  - PhiX error rate analysis
  - base quality summary
- Sequence processing
  - Table - pipeline sequence budget
    - number of reads filtered due to low quality
    - number of reads merged
    - number of chimeras
- OTU table
  - section objective - identify/ highlight OTUs used in the following sections
  - Figure OTU abundance distribution by pipeline
  - Summary of Pre vs. Post specific OTUs
    - abundance
    - taxonomy
- Count Variance
  - section objective - characterize count variance between PCR replicates
    - is the variance correlated with experimental values e.g. biological sample, PCR plate, well, sequencing depth, or observed count value
  - Figure - relationship between count and PCR replicate variance
- Normalization
  - section objective - used PCR replicate variance values to validate normalization methods
  - Compare variance distributions for different normalization methods
  - Test-train or cross-validation based approach????
    - split set of replicates based on the distribution/ range of sequences in a dataset
- Response linearity
  - section objective - demonstrate how the dataset is used to characterize relative abundance estimates and identify potential sources of bias
  - Figure observed vs expected plots
  - Figure representative OTUs showing different types of response linearity
  - Differentiating between high and low linearity OTUs
- Differential Abundance
  - section objective - demonstrate how the dataset can be used to evaluate the limit of differential abundance
  - Figure - MA plot
  - Pre and Post unique OTUs
  - OTUs in both pre and post
  - Differential abundance dectection between unmixed and tritrated samples

# Discussion

# Acknowladgements

# References