**Early Life Bacteria Fungi Manuscript Analysis: README**

**Author: Nelly Amenyogbe**

**22-July-2020**

Table of Contents

[Introduction 2](#_Toc46953514)

[Data file descriptors 2](#_Toc46953515)

[Metadata 2](#_Toc46953516)

[Amplicon Sequencing data 2](#_Toc46953517)

[**Rdata/raw\_data/MOTHUR\_16S** 2](#_Toc46953518)

[**Rdata/raw\_data/MOTHUR\_ITS2** 3](#_Toc46953519)

[Analytical script descriptors 3](#_Toc46953520)

[Data preparation 3](#_Toc46953521)

[Prepare phyloseq objects 3](#_Toc46953522)

[**scripts/data\_preparation/ms\_16S\_make\_phyloseq.R** 3](#_Toc46953523)

[**scripts/data\_preparation/ms\_ITS2\_make\_phyloseq.R** 3](#_Toc46953524)

[Remove Contaminating OTUs 4](#_Toc46953525)

[**scripts/data\_preparation/ms\_QC\_16S\_data.R AND ms\_QC\_ITS2\_data.R** 4](#_Toc46953526)

[Remove low depth samples 4](#_Toc46953527)

[**scripts/data\_preparation/ms\_16S\_depth.R AND ms\_ITS2\_depth.R** 4](#_Toc46953528)

[Data Analysis 4](#_Toc46953529)

[Alpha Diversity 4](#_Toc46953530)

[Beta Diversity 4](#_Toc46953531)

[**scripts/data\_analysis/ms\_16S\_beta\_diversity.R AND ms\_ITS2\_beta\_diversity.R** 4](#_Toc46953532)

[**scripts/data\_analysis/ms\_ITS2\_prepare\_bray\_curtis.R AND ms\_16S\_prepare\_bray\_curtis.R** 5](#_Toc46953533)

[**scripts/data\_analysis/ms\_16S\_ITS2\_bray\_curtis\_child\_distance.R** 5](#_Toc46953534)

[Taxonomic composition 5](#_Toc46953535)

[**scripts/data\_analysis/ms\_16S\_taxa\_barplot.R** 5](#_Toc46953536)

[**scripts/data\_analysis/ms\_ITS2\_taxa\_barplot.R** 5](#_Toc46953537)

[**scripts/data\_analysis/ms\_mothers\_deseq.R** 6](#_Toc46953538)

[**scripts/data\_analysis/ms\_16S\_shared\_otus** 6](#_Toc46953539)

# **Introduction**

This folder contains all raw numerical data presented in the manuscript:

**Bacterial and fungal gut community dynamics over the first five years of life in predominantly rural communities in Ghana**

The **ms\_bacteria\_fungi\_analysis** directory contains all input data, scripts, and figures associated with the study manuscript. The data is organized into three main folders:

* **Rdata:** Contains all data utilized for manuscript analysis, in assay-type specific folders. Intermediate files generated by the raw input data, that are then passed into later analyses, are saved in the **R\_export**directory.
* **scripts:** This directory contains all R scripts associated with analyses performed for this manuscript. Each script can be run independently of others. Where intermediate files are required that do not include the raw input data, these have been saved in Rdata/R\_export and are appropriately refereced in each script
* **figures:** This directory contains all figures used to generate figures for the manuscript, in data-type specific folders

# **Data file descriptors**

## **Metadata**

* **Rdata/sample\_metadata.csv:** contains all metadata for study particiants utilized for data analysis
* **Rdata/ms\_bacteria\_fungi\_data\_dictionary.xlsx:** excel spreadsheet containing explanations of all variables included in metadata or sample data sheets for amplicon sequencing

## **Amplicon Sequencing data**

### **Rdata/raw\_data/MOTHUR\_16S**

* **sample.names.csv:** sample metadata for all samples included in the 16S amplicon sequencing dataset, including blanks and controls
* **\*.shared**: OTU table generated with the MOTHUR pipeline used to parse raw fastq files
* **\*.taxonomy**: Taxonomic assignment of all OTUs included in the OTU table
* **physeq\_16S\_raw.rds**: Microbiome data is supplied as a phyloseq object. Access the data using the readRDS() command in R statistical software. The latest version can be found here:

<https://cran.r-project.org/bin/windows/base/>

R studio is a user-friendly tool to facilitate the use of R, and can be downloaded here:

<https://rstudio.com/products/rstudio/download/>

The Phyloseq R package must be installed to load the phyloseq object. Instructions to download phyloseq can be found here:

<https://bioconductor.org/packages/release/bioc/html/phyloseq.html>

### **Rdata/raw\_data/MOTHUR\_ITS2**

* **its2\_sample.names.csv:** sample metadata for all samples included in the 16S amplicon sequencing dataset, including blanks and controls
* **its2.\*.shared**: OTU table generated with the MOTHUR pipeline used to parse raw fastq files for fungi
* **its2.\*.taxonomy**: Taxonomic assignment of all OTUs included in the OTU table for fungi
* **physeq\_ITS2\_raw.rds**: Microbiome data is supplied as a phyloseq object. Load according to instructions outlined for 16S data.

# **Analytical script descriptors**

**NOTE**: scripts/session\_info.docx and sesson\_info.Rmd contain the package versions used to perform all analyses within the scripts described below.

## **Data preparation**

## **Prepare phyloseq objects**

### **scripts/data\_preparation/ms\_16S\_make\_phyloseq.R**

In this script, we used the output OTU table (.shared) and taxonomy file (.taxonomy) from MOTHUR processing of the fastq files, and combine these with sample metadata to produce a phyloseq object. This data-friendly format will ease the process of performing any analyses, including QA, on these data.

* **Output figures:** none
* **Output files:** Phyloseq object for raw data: Rdata/raw\_data/MOTHUR\_16S/physeq\_16S\_raw.rds

### **scripts/data\_preparation/ms\_ITS2\_make\_phyloseq.R**

In this script, we used the output OTU table (.shared) and taxonomy file (.taxonomy) from MOTHUR processing of the fastq files, and combine these with sample metadata to produce a phyloseq object. This data-friendly format will ease the process of performing any analyses, including QA, on these data.

* **Output figures:** none
* **Output files:** Phyloseq object for raw data: Rdata/raw\_data/MOTHUR\_ITS2/physeq\_ITS2\_raw.rds

## **Remove Contaminating OTUs**

### **scripts/data\_preparation/ms\_QC\_16S\_data.R AND ms\_QC\_ITS2\_data.R**

In this script, we prepared the data for biological analysis by flagging contaimating OTUs and removing them from analysis and trimming the blank samples from the dataset. Samples with low counts will be removed after comparison with the fungal data

* **Output figures:** none
* **Output files:** Phyloseq object of QC’d data, LOW COUNT SAMPLES REMAIN
  + 16S: Rdata/R\_export/prev\_16S\_physeq\_filtered.rds
  + ITS2: Rdata/R\_export/prev\_ITS2\_physeq\_filtered.rds

## **Remove low depth samples**

### **scripts/data\_preparation/ms\_16S\_depth.R AND ms\_ITS2\_depth.R**

In this script we plot the sequencing depth of all samples, and then remove samples with depth below 1000 reads, together with OTUs with counts less than 3 across the dataset

* **Output figures:** Boxplots for sequencing depth:
  + **Supl. Figure 1A**: sequencing\_depth/prev\_16S\_depth.pdf
  + **Supl. Figure 1B:** sequencing\_depth/prev\_ITS2\_depth.pdf
  + **Supl. Figure 1C** Samples with fungi detected: sequencing\_depth/prev\_ITS2\_detected.pdf
* **Output files:** Phyloseq objects ready for biological analysis:
  + 16S: Rdata/R\_export/ps\_16S\_for\_analysis.rds

## **Data Analysis**

## **Alpha Diversity**

In this script, we determined Shannon Diversity and Observed Richness and performed the Wilcoxon test to compare each age stratification to maternal richness.

* **Output figures:** Boxplots for alpha diversity:
  + **Figure 1A:** alpha\_diversity/stool\_16S\_alphadiv.pdf
  + **Figure 1B:** alpha\_diversity/stool\_ITS2\_alphadiv.pdf
* **Output files:** none

## **Beta Diversity**

### **scripts/data\_analysis/ms\_16S\_beta\_diversity.R AND ms\_ITS2\_beta\_diversity.R**

Ordinate all samples and use PERMANOVA to determine variance explained by age

* **Output figures:** Ordination of samples:
  + **Figure 2A:** beta\_diversity/ord\_16S\_nmds.pdf AND **Figure 2B:** beta\_diversity/ord\_ITS2\_nmds.pdf
* **Output files:** none

### **scripts/data\_analysis/ms\_ITS2\_prepare\_bray\_curtis.R AND ms\_16S\_prepare\_bray\_curtis.R**

In this script, we create a Bray-Curtis distance matrix, and transform this into a data frame of every pairwise combination. Metadata is added for the first and second pair. This data will be exported for the next analysis. Stepping stone scripts for ms\_16S\_ITS2\_bray\_curtis\_child\_distance.R.

* **Output figures:** **Figure 4A:** distances of mothers stool to baby stool: beta\_diversity/prev\_16s\_nb\_st\_distance.pdf
* **Output files:** data frames of pairwise Bray-Curtis distances:
  + Rdata/R\_export/its\_bray\_curtis\_dist.csv
  + Rdata/R\_export/16s\_bray\_curtis\_dist.csv

### **scripts/data\_analysis/ms\_16S\_ITS2\_bray\_curtis\_child\_distance.R**

In this script, we compare the bray-curtis distance between all age stratifications with the oldest children for bacteria and fungi

* **Output figures:** **Figure 2C:** distances children to oldest children for bacteria and fungi: beta\_diversity/its\_s16\_child\_dist.pdf
* **Output files:** none

## **Taxonomic composition**

### **scripts/data\_analysis/ms\_16S\_taxa\_barplot.R**

In this script we (1) transform count to relative abundance data, (2) Generate barplots for top-10 community composition for breastmilk and stool samples separately and (3) Generate boxplots to visualize temporal trends for associated taxa over age.

* **Output figures:**
  + **Figure 3A:** Barplots of stool taxonomic composition: taxonomy/stool\_16S\_taxabars.pdf
  + **Supl. Figure 2:**Barplot of bacteroides/prevotella: taxonomy/bar\_prev\_bac.pdf
  + **Figure 5A:** Barplot of breastmilk taxonomic composition: taxonomy/bm\_16S\_taxabars.pdf
  + **Figure 4 C-F:** Boxplots of selected genera
    - taxonomy/prev\_select.pdf
    - taxonomy/ecoli\_select.pdf
    - taxonomy/faec\_select.pdf
    - taxonomy/blaut\_select.pdf
* **Output files:** none

### **scripts/data\_analysis/ms\_ITS2\_taxa\_barplot.R**

In this script we (1) transform count to relative abundance data, (2) Generate barplots for top-10 community composition for breastmilk and stool samples separately.

* **Output figures:**
  + **Figure 3B:** Barplots of stool taxonomic composition: taxonomy/stool\_ITS2\_taxabars.pdf
  + **Figure 5B:** Barplot of breastmilk taxonomic composition: taxonomy/bm\_ITS2\_taxabars.pdf
* **Output files:** none

### **scripts/data\_analysis/ms\_mothers\_deseq.R**

In this script, we used DESeq2 to identify differentially abundant OTUs between mothers one week (Mother\_0-5 days) and one month (Mothers\_26-35 days) post-partum for stool and breastmilk.

* **Output figures:** Differentially-abundant OTUs identified by deseq
  + **Figure 4B:** Stool 16S: taxonomy/deseq\_mot\_16s\_stool.pdf
  + **Figure 5C:** Breastmilk 16S: taxonomy/deseq\_mot\_16s\_bm.pdf
* **Output files:** none

### **scripts/data\_analysis/ms\_16S\_shared\_otus**

In this script, we identify number of shared OTUs between mother-infant related and unrelated pairs, to determine if related pairs share more OTUs in common. We plot the most shared OTUS.

* **Output figures:** 
  + Shared OTUs between mother breastmilk and infant stool
    - **Figure 5D:** shared\_otus/shared\_16s\_bm\_otus.pdf
    - **Figure 5E:**shared\_otus/shared\_ITS2\_bm\_otus.pdf
    - **Supl. Figure 4A:** shared\_otus/bm\_16S\_shared\_linegraph.pdf
    - **Supl. Figure 4B:** shared\_otus/bm\_ITS2\_shared\_linegraph.pdf
  + Shared OTUs between mother stool and infant stool
    - **Supl. Figure 3A:** shared\_otus/shared\_16s\_ms\_otus.pdf
    - **Supl. Figure 3B:** shared\_otus/shared\_ITS2\_ms\_otus.pdf
    - **Supl. Figure 3C:** shared\_otus/prev\_16s\_st\_shared\_linegraph.pdf
    - **Supl. Figure 3D:** shared\_otus/prev\_its2\_st\_shared\_linegraph.pdf
* **Output files:** none