16S\_DuodenalMicrobiome\_Meta-Analysis Analysis Guidance Document

# Summary

FASTQ files were retrieved from the SRA archive using the SRA linux toolkit and compared with our Zambian dataset. Details of these samples are available in the README.md document.

## Data cleaning, ASV assignment and taxonomic identification

These FASTQ files underwent cleaning, filtering and taxonomic assignment using a pipeline similar to the one used to process the Zambian dataset. This step was done on the datasets separately, as they were different sequencing runs with different error rates.

The amplicon\_sequencing\_pipeline (https://gitlab.com/Gordon\_Lab/amplicon\_sequencing\_pipeline) implemented for the Zambia datasets involved 3 main steps:

* Removing primer sequences using BBTools
* Filtering and trimming with DADA2
* Calling ASVs with DADA2

### SEEM dataset

#### Identifying and removing primer sequences

The FASTQ files were inspected for the primer sequences (515f and 806r) using the primerHits function adapted from the ITS DADA2 pipeline (Meta-Analysis/Code/CheckPrimers.Rmd).

This analysis showed presence of primer sequences at the start of most reads therefore, bbtools ([**bbmap@38.63**](mailto:bbmap@38.63)) was used to trim these out before filtering.

Code file: ‘Meta-Analysis/Code/RemovePrimersSeem.sh’

Output: FASTQ files without primer sequences

#### Filtering and trimming FASTQ files

Code file: ‘Meta-Analysis/Code/2\_TrimAndFilter\_DADA2\_Seem.sh’ is used to execute the R script: ‘2\_trim\_and\_filter\_DADA2\_seem.R’

Inputs: FASTQ files without primer sequences from section 1.1.1.1

Output: filtered FASTQ files

#### Calling ASVs with DADA2

Code file: ‘Meta-Analysis/Code/3\_call\_ASVs\_DADA2\_seem.sh’ is used to execute the R script: ‘3\_call\_ASVs\_DADA2\_seem.R’

Inputs: filtered FASTQ files from section 1.1.1.2

Output: seqtab.agg\_seem.RDS– ASV counts table

### BEED dataset

This dataset consisted of 58 FASTQ files from 36 samples. Some samples were sequenced twice and treated as separate samples for this stage.

#### Identifying and removing primer sequences

The FASTQ files were inspected for the primer sequences (515f and 806r) using the primerHits function adapted from the ITS DADA2 pipeline (Meta-Analysis/Code/CheckPrimers.Rmd).

This analysis showed presence of primer sequences at the start of most reads therefore, bbtools ([**bbmap@38.63**](mailto:bbmap@38.63)) was used to trim these out before filtering.

Code file: ‘Meta-Analysis/Code/1\_RemovePrimersSeem.sh’

Inputs: 62 FASTQ files retrieved from SRA

Output: FASTQ files without primer sequences

#### Filtering and trimming FASTQ files

Code file: ‘Meta-Analysis/Code/2\_TrimAndFilter\_DADA2\_Beed.sh’ is used to execute the R script: ‘2\_trim\_and\_filter\_DADA2\_beed.R’

Inputs: FASTQ files without primer sequences from section 1.1.2.1

Output: filtered FASTQ files

#### Calling ASVs with DADA2

Code file: ‘Meta-Analysis/Code/3\_call\_ASVs\_DADA2\_beed.sh is used to execute the R script: ‘3\_call\_ASVs\_DADA2\_beed.R’

Inputs: filtered FASTQ files from section 1.1.2.2

Output: seqtab.agg\_beed.RDS– ASV counts table

### Taxonomy assignment for BEED and SEEM datasets

Code file: ‘Meta-Analysis/Code/ 4\_TaxonomyAssignment.R’

Note: Taxonomic assignment was done using the **silva\_nr99\_v138.1\_train\_set.fa.gz** and **silva\_species\_assignment\_v138.1.fa.gz** databases.

Inputs: Sequence tables from sections 1.1.1.3 & 1.1.2.3 i.e seqtab.agg\_seem.RDS and seqtab.agg\_beed.RDS

Output:

* ExtDat\_Bang\_CountTab.RData – BEED ASV count table
* EXtDat\_Bang\_TaxTab.RData – BEED Taxonomy table
* ExtDat\_SEEM\_CountTab.RData – SEEM ASV count table
* EXtDat\_SEEM\_TaxTab.RData – SEEM Taxonomy table

### AFRIBIOTA dataset

This dataset consisted of 31 FASTQ files from 31 samples. This dataset contained single-ended reads.

#### Checking for primer sequences

DADA2 was used to check if primer sequences were present in the FASTQ files. The sequences used were those described in the methods section of the manuscript which had subtle differences from those used in the BEED and SEEM dataset.

Code file: ‘Meta-Analysis/Code/CheckForPrimers.Rmd’

Inputs: FASTQ files retrieved from SRA archive

There was no evidence of primer sequences, so primer filtering was not performed on this dataset.

#### Filtering, trimming, calling ASVs with DADA2 and taxonomic assignment

Code file: ‘Meta-Analysis/Code/DADA2\_Afribiota.Rmd.

Input files: FASTQ files retrieved from SRA archive

Lines 15 – 57: Filtering and trimming. Because this dataset only had forward reads, the parameters used were that of the forward reads from the beed and seem analysis.

Lines 59 – 101: ASV calling and removing chimera reads

Lines 114 – 133: Taxonomic assignment using the **silva\_nr99\_v138.1\_train\_set.fa.gz** and **silva\_species\_assignment\_v138.1.fa.gz** databases.

Lines 135 – 141: Saving output files for downstream analysis.

Final Output files:

* Meta-Analysis/Data/RData/ExtDat\_Afr\_CountTab.RData – ASV Count table
* Meta-Analysis /Data/RData/EXtDat\_Afr\_TaxTab.RData – Taxonomy table

### Merging data and cleanup

Code file: ‘Meta-Analysis/Code/5DataCleaning.Rmd.

Input files:

* ExtDat\_Afr\_CountTab.RData – Afribiota ASV Count table
* EXtDat\_Afr\_TaxTab.RData – Afribiota Taxonomy table
* ExtDat\_Bang\_CountTab.RData – BEED ASV count table
* EXtDat\_Bang\_TaxTab.RData – BEED Taxonomy table
* ExtDat\_SEEM\_CountTab.RData – SEEM ASV count table
* EXtDat\_SEEM\_TaxTab.RData – SEEM Taxonomy table
* ExtDataMetadata.xlsx – Metadata for Afribiota, BEED, and SEEM datasets
* Zambia\_EE\_BEECH\_16s\_absQuant\_forMonica.RData – Zambia Count and Taxonomy tables
* SamBeechMetadata\_2.csv – Metadata for Zambian dataset

Lines 4-49: Setting up working directory and importing data

Lines 52 – 65: merging Afribiota, BEED, and SEEM count tables

Lines 67 – 107: merging Afribiota, BEED, and SEEM taxonomy tables

Lines 110 – 124: Creating phyloseq object for Afribiota, BEED, and SEEM data.

Lines 126 – 205: removal of spike-in (Alicyclobacillus) data and calculation of absolute abundance.

**Note**: lines 52 – 206 were adapted from the code used to determine absolute abundance in the Zambian dataset.

Lines 209 – 238: Importing the Zambian dataset and merging it compatible with Afribiota, BEED, and SEEM dataset. The Zambian dataset was received with the spike-in (Alicyclobacillus) already removed and absolute abundances calculated.

Lines 241 – 287: Merging all datasets together and excluding samples that met either of the following criteria:

* Weight-for-length z-score < -3 (had severe acute malnutrition)
* Length-for-age z-score > -2 (was not stunted)

After looking at the metadata, it was realised that some samples from the BEED were sequenced twice (Figure 1). These were averaged leaving 128 samples in total for analysis.

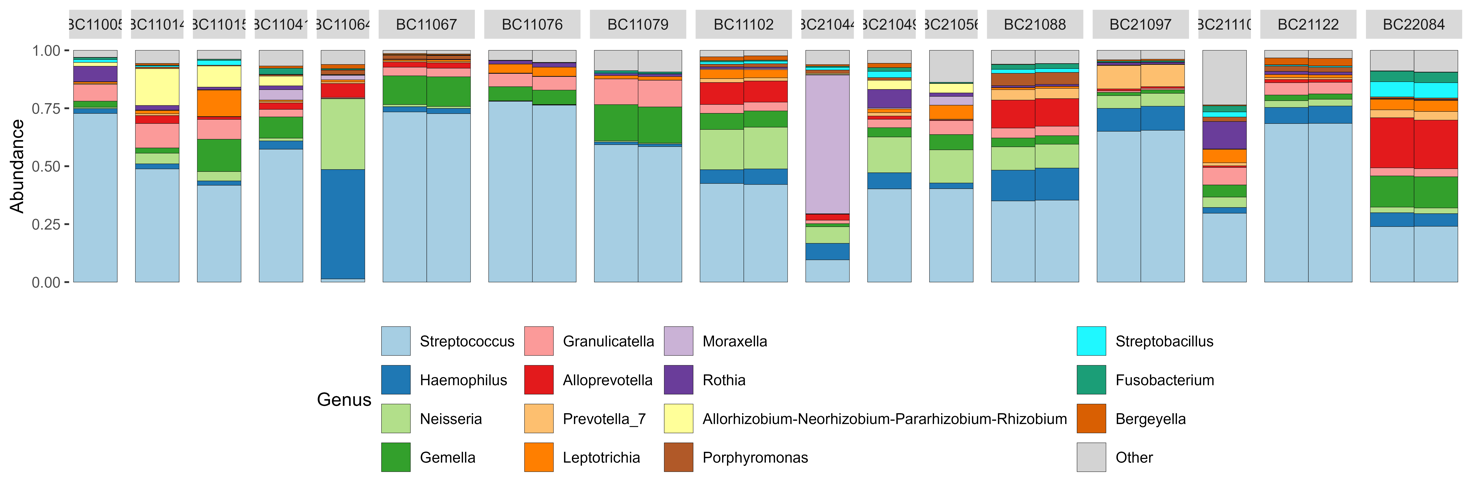


Figure 1: Relative abundance of genera from Bangladeshi children.

Final Output files:

* Meta-Analysis/Data/RData/AllMAlStudies.RData
* Meta-Analysis/Data/RData/ StuntingCombined\_unique.RData

### Comparison of datasets

Code file: ‘Meta-Analysis/Code/6\_STCombinedAnalysis\_Apr.R’

Inputs:

* Meta-Analysis/Data/RData/StuntingCombined\_unique.RData

Output Location: ‘Meta-Analysis/Outputs’

#### Summary Stats (Lines 20 - 71)

The first part of the R script calculates the median (IQR) of age and anthropometry and saves them in table and graphical format.

#### Relative Abundance (Lines 72 - 105)

The relative abundance of the top 15 genera (total sum of abundance across all samples) was visualized using barplots.

This section also contains a code that produces a Venn diagram of the core taxa found in the age-matched children (lines 107 – 159)

#### Alpha diversity (Lines 160 – 213)

Alpha diversity was visualised using boxplots and was correlated with anthropometry measures using Pearson correlation analysis.

#### Beta diversity (Lines 214 – 283)

Bray Curtis distances were calculated and used as input for pairwise PERMANOVA analysis of the different countries/ studies. The bray Curtis distances were also plotted in a PCoA plot.

**Note**: The AFRIBIOTA study contained 2 countries.

#### Log Linear models (Lines 285 – 428)

This last part of the R script details the linear modelling carried out to assess differences between children that were older versus younger than 2.5 years as well as between the age-matched cohorts (BEED, SEEM, BEECH).

#### Combined plots (Lines 430 – 447)

This section of the part of the R script created plots to go with the manuscript.

|  |  |  |  |
| --- | --- | --- | --- |
| **Current Document** | | | |
| Developed by: | Monica Mweetwa | Date: | 02/01/2025 |
| Code tested by: |  | Date: |  |
|  |  | | |