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 and ASV assignment

These FASTQ files underwent cleaning and filtering and ASV calling using a pipeline like 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 (for SEEM and BEED only)
* 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\_Apr.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’

Inputs: FASTQ files retrieved from SRA (not provided in GitHub)

Output: FASTQ files without primer sequences

#### Filtering and trimming FASTQ files

Code file: ‘Meta-Analysis/Code/SEEM/2\_TrimAndFilter\_DADA2\_Seem.sh’ is used to execute the R script: ‘Meta-Analysis/Code/SEEM/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/SEEM/3\_call\_ASVs\_DADA2\_seem.sh’ is used to execute the R script: ‘Meta-Analysis/Code/SEEM/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\_Apr.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/BEED/1\_RemovePrimersBeed.sh’

Inputs: 62 FASTQ files retrieved from SRA (not provided in GitHub)

Output: FASTQ files without primer sequences

#### Filtering and trimming FASTQ files

Code file: ‘Meta-Analysis/Code/BEED/2\_TrimAndFilter\_DADA2\_Beed.sh’ is used to execute the R script: ‘Meta-Analysis/Code/BEED/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/BEED/3\_call\_ASVs\_DADA2\_beed.sh’ is used to execute the R script: ‘Meta-Analysis/Code/BEED/3\_call\_ASVs\_DADA2\_beed.R’

Inputs: filtered FASTQ files from section 1.1.2.2

Output: seqtab.agg\_beed.RDS – ASV counts 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 (not provided in GitHub)

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

#### Filtering, trimming and calling ASVs

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

Input files: FASTQ files retrieved from SRA archive (not provided in GitHub)

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

## Taxonomy assignment for all datasets

Taxonomic assignment was done for all studies combined and was based on the silva v138.1 (**silva\_nr99\_v138.1\_train\_set.fa.gz** and **silva\_species\_assignment\_v138.1.fa.gz**).

Code file: ‘Meta-Analysis/Code/4\_TaxonomyAssignment\_All\_2025.R’

Input files are

* RData/seqtab.agg\_beed.RDS – BEED dataset (not provided in GitHub)
* RData/ExtDat\_Afr\_CountTab.RData – AFRIBIOTA dataset (not provided in GitHub)
* RData/seqtab.agg\_Zam.RDS – Zambian dataset (not provided in GitHub)
* RData/seqtab.agg\_seem.RDS – SEEM dataset (not provided in GitHub)

Output:

* All\_CountTab2025.RData – merged ASV count table for all studies (not provided in GitHub)

All\_TaxTab2025.RData – taxonomy table for all studies (not provided in GitHub)

## Data cleanup

Code file: ‘Meta-Analysis/Code/5\_DataCleaning\_Jul23.Rmd’

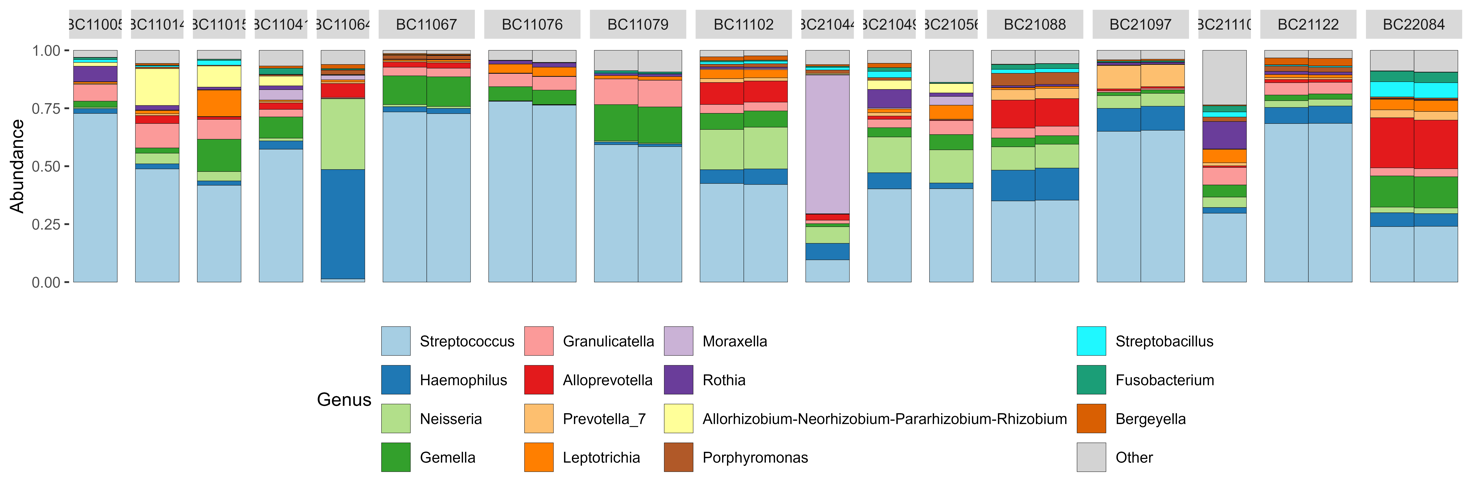
Input files:

* All\_CountTab2025.RData – merged ASV count table for all studies (not provided in GitHub)
* All\_TaxTab2025.RData – taxonomy table for all studies (not provided in GitHub)
* ExtDataMetadata\_Jul14.xlsx – Metadata for AFRIBIOTA, BEED, and SEEM datasets (not provided in GitHub)
* SamBeechMetadata\_2.csv – Metadata for Zambian dataset (not provided in GitHub)

Lines 4 – 38: Setting up working directory and importing data

Lines 40 – 84: Creating phyloseq object, removing of spike-in (Alicyclobacillus) data and ambiguous assignment as class (Chloroplast) and family level (Mitochondria).

Lines 86 – 118: 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.**

Lines 125 – 273: calculation of absolute abundance for the BEED, BECH ad SEEM studies. To get these absolute values, the following steps were taken:

* Calculate bacterial load factor:
* Determine number of spike-in cells added to the sample in millilitres when the number of cells added to the samples in microliters is known.
  + 1.1 x 106 was added to 100uL of sample for BEED so multiply this by 10 to get cells/ mL
  + 9.9 x 105 was added to 200uL of sample for BEECH so multiply this by 5 to get cells/ mL
  + 1.1 x 106 was added to 50uL of sample for SEEM so multiply this by 20 to get cells/ mL
* The number of cells was multiplied by the bacterial load factor to get the actual bacterial load

A phyloseq object with these normalized counts was created for downstream analysis.

Final Output files for downstream analysis:

* ‘Meta-Analysis/RData/AllDatasets\_RelAbundNoSpikeNoDuplicates.RData‘ – data for all studies (AFRIBIOTA, BEED, SEEM and BEECH)
* Meta-Analysis/RData/BEED\_SEEM\_BEECH\_AbsASV.RData – absolute abundance data for BEED, BEECH and SEEM.

## Statistical analysis

### Summary Stats

Summary statistics of anthropometry, sex, and age by study was generated in tabular and graphical format.

Code file: ‘Meta-Analysis/Code/6\_DescriptiveStats\_Aug.R’

Input: ‘Meta-Analysis/RData/AllDatasets\_RelAbundNoSpikeNoDuplicates.RData’

### AFRIBIOTA comparison

Since the AFRIBIOTA cohort was different from the BEECH, BEED and SEEM datasets in terms of age, a comparison analysis was done to see the microbial drivers of difference between these groups. This was done with the acknowledgment that there were also differences in the DNA extraction and sequencing method between these groups. To do this, alpha and beta diversity was compared.

Code file: ‘Meta-Analysis/Code/7\_AfribiotaComparison.R’

Input file: Meta-Analysis/RData/AllDatasets\_RelAbundNoSpikeNoDuplicates.RData

Lines 20 – 39: Relative abundance plot of the top 15 genera (total sum of abundance across all samples) was visualized using bar plots.

Lines 41 – 67: Alpha diversity was visualised using boxplots and was correlated with anthropometry measures using Pearson correlation analysis.

Lines 69 – 107: The bray Curtis distances were plotted in a PCoA plot and the relative abundance of genera were exported for use as input for PERMANOVA analysis in the PRIMER7 software. This used the Anderson’s correction which accounts for imbalanced sample size between groups.

Lines 109 – 114: Combined figures of all analysis were generated.

### Alpha diversity analysis (BEED, BEECH and SEEM)

This code was used to estimate the alpha diversity (Shannon index) and regression analysis with clinical features.

Code file: ‘Meta-Analysis/Code/8\_AlphaDiversity\_Meta.R’

Input file: ‘Meta-Analysis/RData/BEED\_SEEM\_BEECH\_AbsASV.RData’

### Core taxa identification (BEED, BEECH and SEEM)

This code was used to determine the genera that were present in at least 80% of samples in each study. Their relative abundance was also calculated and exported.

Code file: ‘Meta-Analysis/Code/8\_CoreTaxa\_Met\_Aug.R’

Input file: ‘Meta-Analysis/RData/BEED\_SEEM\_BEECH\_AbsASV.RData’

### Differential abundance analysis (BEED, BEECH and SEEM)

This code was used to determine the associations between the absolute abundance of the core genera and anthropometry and age.

Code file: ‘Meta-Analysis/Code/9\_DifferentialAbundance\_Analysis\_AugGit.R’

Input file: ‘Meta-Analysis/RData/BEED\_SEEM\_BEECH\_AbsASV.RData’

Lines 3 – 13: Importing data and necessary packages

Lines 15 – 49: Pre-processing the data to fit format for analysis and creating plot colors

Lines 50 – 148: Regression analysis with anthropometry and age for datasets combined and individual study analysis.

Lines 151 – 221: Pearson correlation analysis for BEED and SEEM datasets for both the full dataset and subsets.

|  |  |  |  |
| --- | --- | --- | --- |
| **Current Document** | | | |
| Developed by: | Monica Mweetwa | Date: | 02/01/2025 |
| Code reviewed by: | Nate McNulty | Date: | 01/04/2025 |
| Code revised by | Monica Mweetwa | Date: | 14/08/2025 |