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 is available in the Read.me document.

## Data cleaning, ASV assignment and taxonomic identification

These fastqQ files underwent cleaning, filtering and taxonomic assignment using a similar pipeline that produced 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 (sequchttps://gitlab.com/Gordon\_Lab/amplicon\_sequencing\_pipeline) implemented for the Zamina datasets involved 3 main steps:

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

The retrieved datasets did not have primer sequences therefore only filtering and trimming and ASV calling using DADA2 was implemented.

### SEEM dataset

This dataset consisted of 58 fastq files from 36 samples. Some samples were run in duplicate but treated as separate samples for this stage.

#### Checking for Primer sequences and filter them out

The fastq files were visually inspected for the primer sequences at the start if the reads. This was supported by using DADA2 to determine how many sequences had the 515f and 506r primer sequences in the fastQ files. These primers were those described in the methods section of the manuscript.

Code file: ‘Meta-Analysis/Code/1\_CheckForPrimers.Rmd’

Inputs: 62 fastq files retrieved from SRA

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: 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\_Seem.sh’ & ‘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& ‘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 run in duplicate but treated as separate samples for this stage.

#### Checking for Primer sequences

The fastq files were visually inspected for the primer sequences at the start if the reads. This was supported by using DADA2 to determine how many sequences had the 515f and 506r primer sequences in the fastQ files. These primers were those described in the methods section of the manuscript.

Code file: ‘Meta-Analysis/Code/1\_CheckForPrimers.Rmd’

Inputs: 62 fastq files retrieved from SRA

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’ & ‘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& ‘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’

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 samples were run in duplicate but treated as separate samples for this stage.

#### 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.

Code file: ‘Meta-Analysis/Code/1\_CheckForPrimers.Rmd’

Inputs: fastq files retrieved from SRA archive

There was no evidence of primer sequences, so primer filtering was not done 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 has single 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/DADA2\_Afribiota.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 Zambian dataset and make compatible with Afribiota, BEED, and SEEM dataset. The Zambian dataset was received with the spike-in (Alicyclobacillus) removed and absolute abundances calculated already.

Lines 241 – 255: Merging all datasets together and saving for downstream analysis

Final Output files:

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

### Comparison of datasets

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

Inputs:

* Meta-Analysis/Data/RData/AllMAlStudies.RData

Output Location: ‘Meta-Analysis/Outputs’

#### Summary Stats (Lines 10 - 82)

The first section of the notebook shows the distribution of the anthropometry in table and graphical format.

* Some samples had LAZ > -2. These were filtered out which brought the number of samples down to 158 from 207.
* After looking at the metadata, it was realised that some samples from the BEED were sequenced twice. These were averaged. This left 149 samples for analysis.

#### Relative Abundance (Lines 84 -117)

The relative abundance of the top 15 genera was visualized using barplots

#### Alpha diversity (Lines 120 – 194)

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

#### Beta diversity (Lines 196 – 276)

PERMANOVA analysis was carried out on the available metadata, LAZ, WAZ, WLZ, Age, Sex, Country and Study.

The Zambian dataset had 2 studies while the Afribiota study contained 2 countries.

Country and study accounted for the largest variation on the 1st and 2nd principal components and so these were plotted.

#### Combined plots (Lines 279 – 294)

The last part of the Rmd file created plots to go with the manuscript.

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