16S\_DuodenalMicrobiome\_MalZam Analysis Guidance Document

# Summary

This section details the analysis using the BEECH and EEBiomarker datasets to compare the duodenal microbiome of stunted vs SAM children. Almost all children with SAM were stunted but stunted children were not servery malnourished i.e severe wasting (WLZ < -2). The raw fastQ files were analysed by the WashU team and we received an RData file ‘Zambia\_EE\_BEECH\_16s\_absQuant\_forMonica.RData’ which contained a mapping file (map) and a phyloseq object (ps2) with data from both BEECH and EEBiomarker available.

The corresponding data dictionary with variables used in this analysis can be found in Stunting\_SAM/16S\_DuodenalMicrobiome\_MalZam\_CodeBook.xlsx

## Data Cleaning:

For data cleaning, a phylogenetic tree and clinical metadata i.e morphometry, enteropathy marker and anthropometry data were added to the phyloseq object (ps2) for downstream analysis.

Data cleaning also included excluding ASVs that had no taxonomic assignment, assigned as mitochondria family or chloroplast class.

Code file: ‘Stunting\_SAM/Code/1\_Data Cleaning\_30122024.Rmd’

Inputs:

* Stunting\_SAM/Data/RData/Zambia\_EE\_BEECH\_16s\_absQuant\_forMonica.RData
* Stunting\_SAM/Data/Metadata/SamBeechMetadata22.csv
* Stunting\_SAM/Data/Metadata/Morphometry/BEECHmorphometry\_91\_edited.csv
* Stunting\_SAM/Data/Metadata/Morphometry/SAM\_morphometry.csv

Output:

* Stunting\_SAM/Data/Metadata/Final\_Metadata.csv
* Stunting\_SAM/Data/Sequences.fasta
* Stunting\_SAM/Data/RData/phyloseq\_dataset\_species.RData
* Stunting\_SAM/Data/Tree.tre

## Analysis

### Summary Table

Code file: ‘Stunting\_SAM/Code/2\_SummaryTable.R’

Inputs: Stunting\_SAM/Data/RData/phyloseq\_dataset\_species.RData

Output: Stunting\_SAM/Output/DescriptiveTable\_all.csv

### Comparison Part 1 – Alpha and Beta diversity and, relative abundance visualization

Code file:

* ‘Stunting\_SAM/Code/3\_Analysis\_SAMvsStunting1.Rmd’
* ‘Stunting\_SAM/Code/Functions.R

Inputs: Stunting\_SAM/Data/RData/phyloseq\_dataset\_species.RData

Outputs folder: Stunting\_SAM/Output

Lines 13 – 22: Input files were imported into R and necessary libraries loaded

Lines 24 – 87: Visualization of relative abundance. The output for this is in the ‘RelAbund’ folder

Lines 89 – 213: Calculation of alpha diversity measures (Shannon index and Faith’s PD) and assessing relationship with sample features using spearman correlation and Wilcoxon test.

**Note**: this analysis was initially using linear models but after review, we opted to carry out Shapiro Wilks test to formally check for normality in addition to the qqplots and residual vs fitted plots. This showed that most models did not pass the normality assumptions therefore, we opted to carry out non-parametric analysis instead.

Lines 215 – 253: Beta diversity analysis and visualization. The output for this is in the ‘BetaDiv’ folder

Lines 255 – 264: combined plot i.e Figure 1 in the manuscript stored in ‘Output’ folder.

### Comparison Part 2 – Differential abundance analysis

Code file:

* ‘Stunting\_SAM/Code/3\_Analysis\_SAMvsStunting2\_2025.Rmd’
* ‘Stunting\_SAM/Code/Functions.R

Inputs: Stunting\_SAM/Data/RData/phyloseq\_dataset\_species.RData

Outputs folder: Stunting\_SAM/Output

* SAM vs stunting:

The first section of the notebook (lines 26 -119) details linear models assessing the relationship between the log2 transformed relative abundances of taxa (phylum to genus) and malnutrition type. Tables of the model were saved, and scatter plots of significant associations were plotted.

* Genera associations with clinical data:

The second half of the notebook (lines 122 – 172) has linear models assessing the relationship between the log2 transformed relative abundances of genera with sample features i.e anthropometry, biomarkers and morphometry.

|  |  |  |  |
| --- | --- | --- | --- |
| **Current Document** | | | |
| Developed by: | Monica Mweetwa | Date: | 09/01/2025 |
| Code tested by: |  | Date: |  |
|  |  | | |