

16S rRNA microbial biodiversity analysis using GVL-Galaxy

Part 4: workflow biodiversity analysis for BIOM data

Version 1.0

Contents

1	Introduction	3
1.1	Genomics Virtual Lab - Galaxy	3
2	Workflow	4
3	Dataset	5
4	16S biodiversity analysis converting raw count to BIOM format workflow	6
4.1	Import data.....	7
4.2	Import workflow	7
4.3	Run the workflow.....	7
4.4	Specify parameters.....	8
5	Expected output.....	12
6	Version History	16

1 Introduction

This is a step-by-step guide in performing a 16S ribosomal RNA (16S rRNA) metagenomic analysis to characterise the microbiome of samples. The aim of this tutorial is to identify the biodiversity and abundance of 16S rRNA in different samples.

The most common method at present to uncover the microbial diversity of a sample is to perform a metagenomic analysis, using the 16S rRNA sequence. The 16S rRNA gene is about 1,500 basepairs (bp) in length and is a component of a prokaryotic ribosome. Profiling 16S rRNA using sequencing technology can help researchers to identify bacterial species in given samples. 16S rRNA is a protein synthesis machinery and is highly conserved. The changes of sequence in 16S gene is used to measure the evolution of bacterial species to study phylogeny and taxonomy. However, the metagenomic study of 16S rRNA is constraint by the annotated 16S database.

All the dataset used in this tutorial is generated using Next Generation Sequencing (NGS) technology. Single-End (SE) and Paired-End (PE) refer to two types of sequencing technique commonly used in NGS. In a single-end run, the sequencer will only sequence from one end of a fragment. In a paired-end run, a fragment will be sequenced from one end and from the opposite end. If the fragments overlap each other, we refer to them as “overlap-PE” in this tutorial. Depending on the protocol used to generate the sequencing data, the workflow used for analysis contains different steps.

We have prepared two different datasets depending on the library protocol that is used for sequencing: (1) overlap-PE or (2) nonoverlap. The steps used for these two protocols are slightly different. If you do not know which protocol your sequencing data used, test it using the **Overlap detection** workflow. Details about the workflows are described in Section 2 Workflow.

1.1 Genomics Virtual Lab - Galaxy

The workflow is setup using Galaxy, which has been deployed using the [Genomics Virtual Lab \(GVL\)](#) platform.

If you are unfamiliar with the Galaxy interface, we recommend you have a look at this [Introduction to Galaxy](#) quick start guide.

2 Workflow

In total, there are 4 Galaxy workflows in the 16S rRNA suite (see Table 1):

Table 1 Summary of workflows

Workflow	Description
1. 16S_overlap_detection	To detect percentage of paired-end reads that overlap each other by 10bp. This workflow randomly selected 1000 reads from each sample to perform the detection. If over 50% of the PE reads overlap each other by at least 10bp, it is recommended to use workflow 2. If less than 50% of PE reads overlap by at least 10bp, it is recommended to use workflow 3.
2. 16S_biodiversity_for_overlapPE	For use with datasets that are sequenced using overlapping paired-end reads.
3. 16S_biodiversity_for_nonoverlap PE	For use with datasets that are sequenced using non-overlapping paired-end reads.
4. 16S_biodiversity_BIOM	This workflow performs the differential analysis and generates a number of visualisations.

This tutorial covers workflow 4 16S rRNA biodiversity BIOM for overlap PE.

3 Dataset

The dataset used for this tutorial is the same as the data from the **02_16S_overlapPE_tutorial**.

The dataset we are using for this guide is from the [16S Microbial analysis with Mothur](#) tutorial. However, we are not performing the same analysis as in the original tutorial. For the purpose of this tutorial, to demonstrate the downstream plotting steps used in the **16S_biodiversity_for_overlapPE** workflow, we have added extra dummy metadata (Table 2).

The first three columns in the metadata table below are from the original study where “*during the first 150 days post weaning (dpw), nothing was done to our mice except allow them to eat, get fat, and be merry*”.

For this tutorial, we are only making use of the *time* column and ignoring the data in the *dpw* column. To demonstrate some of the plotting steps, we have introduced two additional variables, *Food (cheese)* (column 4) indicates the type of food that was fed to the mice and *Replicate_group* (column 5) is the replicate group. Again, we stress that the last two columns have been added specifically for the **16S_biodiversity_for_overlapPE** tutorial.

WARNING: Metadata format

The header of first column in your metadata table must be named “**#SampleID**” in order to be recognised by the BIOM converter step in the workflow.

Table 2 The metadata of overlapped paired-end dataset.

#SampleID	dpw	time	Food	Replicate_Group
F3D0	0	Early	None	Group 1
F3D1	1	Early	None	Group 1
F3D2	2	Early	None	Group 1
F3D3	3	Early	Cheddar	Group 2
F3D5	5	Early	Cheddar	Group 2
F3D6	6	Early	Cheddar	Group 2
F3D7	7	Early	Swiss	Group 3
F3D8	8	Early	Swiss	Group 3
F3D9	9	Early	Swiss	Group 3
F3D141	141	Late	Cheddar	Group 4
F3D142	142	Late	Cheddar	Group 4
F3D143	143	Late	Cheddar	Group 4
F3D144	144	Late	None	Group 5
F3D145	145	Late	None	Group 5
F3D146	146	Late	None	Group 5
F3D147	147	Late	Swiss	Group 6
F3D148	148	Late	Swiss	Group 6
F3D149	149	Late	Swiss	Group 6

4 16S biodiversity analysis converting raw count to BIOM format workflow

The workflow shown in Figure 1 is designed to convert an OTU raw count table to a BIOM file, which is then used for further analysis using the phyloseq R package.

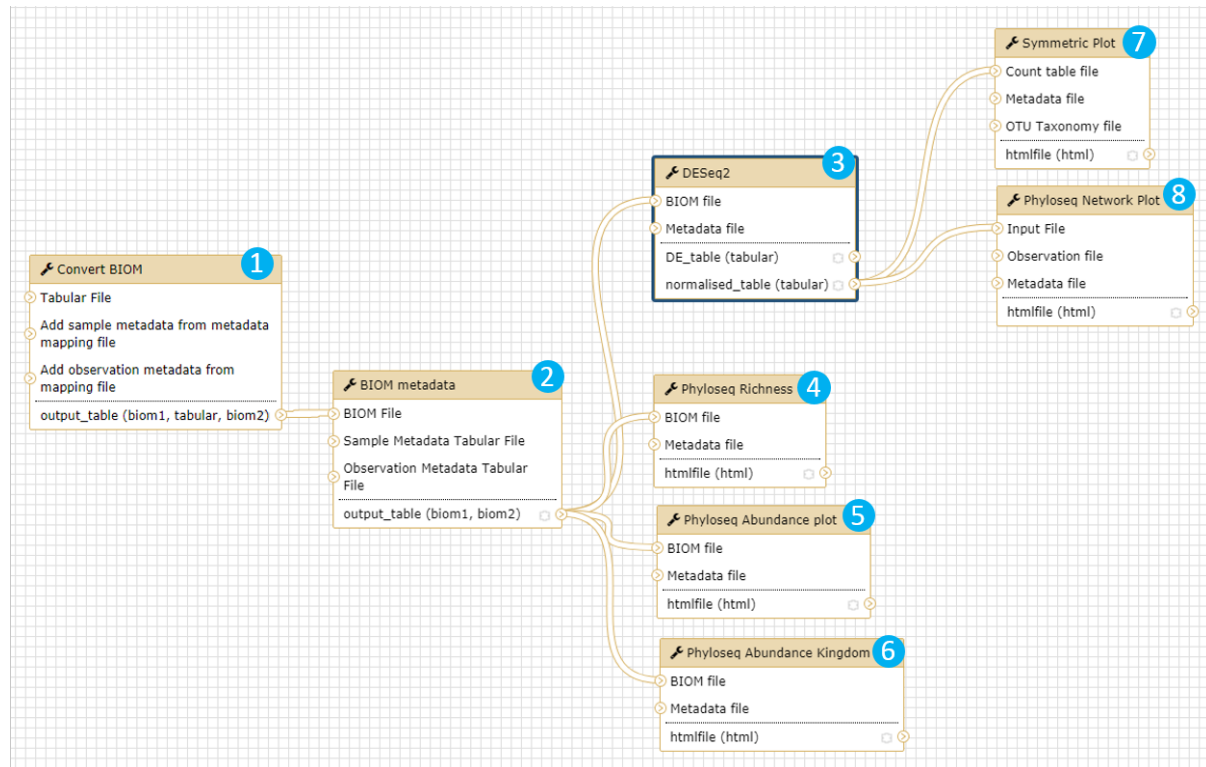


Figure 1 16S biodiversity BIOM workflow

Feature	Description
1. BIOM convert	Convert tabular data (e.g., raw count) to BIOM file format
2. BIOM add metadata	Add metadata to BIOM file
3. DESeq2	A customised feature to normalise raw count table in BIOM file
4. Phyloseq Richness plot	Diversity plot
5. Phyloseq Abundance plot	Abundance plot by factor (e.g, Food)
6. Phyloseq Abundance taxonomy	Abundance plot by taxonomy (e.g, phylum)
7. Symmetric Plot	Display the expression data in barplot between samples
8. Phyloseq Network plot	Network plot for samples similarity

4.1 Import data

PRE-REQUISITE:

The raw count table in this tutorial is available in the **“Shared Data”** in the Galaxy instance.

The raw count table is a tabular formatted file. The rows contain the OTU IDs and the column contains the library name. This raw count table was generated using the workflow **16S_biodiversity_for_overlap_PE**. The details of the workflow can be found in the **02_16S_overlapPE_tutorial** guide.

1. Click **Shared Data** from the top menu
2. Click **Data Libraries** from the dropdown menu
3. Click **Tutorial data: BIOM**
4. Check the checkbox next to **name**



5. Click **to History** from the top menu
6. Type a history name e.g., **16S_biodiversity_BIOM**
7. Repeat step 1 to step 5 to import GreenGenes taxonomy file (**gg_13_5_taxonomy.txt** from the GreenGenes folder) into the same history that was created in step 6 (**16S_biodiversity_BIOM**).
8. Repeat Step 1 to 5 again, but this time click on the folder metadata to import **metadata_OverlappedPE.txt**
9. Click **Analyze Data** from the top menu

4.2 Import workflow

If you are a new user, follow this section to import the shared Galaxy workflow named **“16S_biodiversity_BIOM”** into your workspace. You can skip this section if you have already imported the workflow before.

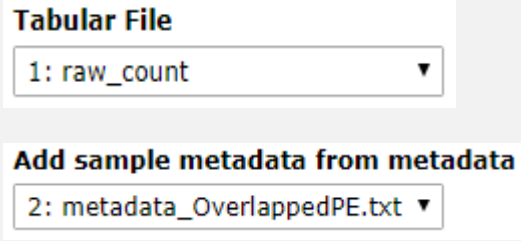
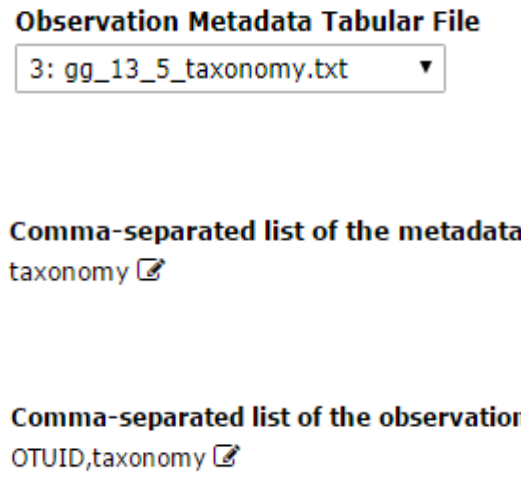


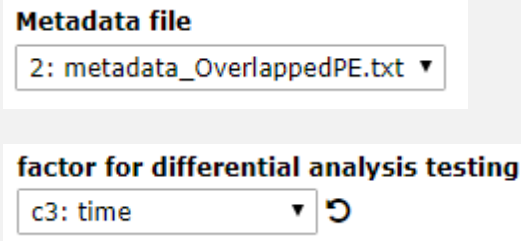

1. Click **Shared Data** from the top menu
2. Click **Workflows** on the dropdown menu
3. Click **16S_biodiversity_BIOM**
4. Click **Import**

4.3 Run the workflow






1. Click on **Workflow** from the top menu
2. Click on **16S_biodiversity_BIOM** workflow
3. Click **Run** from the dropdown menu

4.4 Specify parameters

While the entire workflow consists of 8 total steps, you do not need to specify the parameters for all steps. Follow the following step-by-step guide which highlights the parameters that need modification.

Step	Actions	Screenshot
Step 1: Convert BIOM	<ul style="list-style-type: none"> Select expression value table (e.g., raw_count) from the dropdown menu Select metadata file (e.g., metadata_OverlappedPE.txt) from the "Add sample metadata from metadata mapping file" dropdown menu 	 <p>Tabular File 1: raw_count ▼</p> <p>Add sample metadata from metadata 2: metadata_OverlappedPE.txt ▼</p>
Step 2: BIOM metadata	<ul style="list-style-type: none"> Select green genes annotation file "gg_13_5_taxonomy" from the "Observation Metadata Tabular File" dropdown menu Type "taxonomy" in the textbox under "Comma-separated list of the metadata fields to split on semicolons" Type "OTUID" and "taxonomy" separated by comma in the textbox under "Comma-separated list of the observation metadata field names" 	 <p>Observation Metadata Tabular File 3: gg_13_5_taxonomy.txt ▼</p> <p>Comma-separated list of the metadata taxonomy </p> <p>Comma-separated list of the observation OTUID,taxonomy </p>
Step 3: DESeq2	<ul style="list-style-type: none"> Select metadata file (e.g., metadata_OverlappedPE.txt) from the dropdown menu Select "c3: time" from the "factor for differential analysis testing" 	 <p>Metadata file 2: metadata_OverlappedPE.txt ▼</p> <p>factor for differential analysis testing c3: time ▼ </p>

Step	Actions	Screenshot
Step 4 : Phyloseq Richness	<ul style="list-style-type: none"> Select metadata file (e.g., metadata_OverlappedPE.txt) from “Metadata file” the dropdown menu Select “c5: Replicate_Group” from the “Column used for X-axis” Select “c4: Food” from the “Column used as legend” dropdown menu 	<p>Metadata file</p> <p>2: metadata_OverlappedPE.txt ▼</p> <p>Column used for X-axis</p> <p>c5: Replicate_Group ▼ ↺</p> <p>Column used as legend</p> <p>c4: Food ▼ ↺</p>
Step 5: Phyloseq Abundance plot	<ul style="list-style-type: none"> Select metadata file (e.g., metadata_OverlappedPE.txt) from “Metadata file” the dropdown menu Select “c5: Replicate_Group” from the “Column used for X-axis” Select “c3: time” from the “Column used as legend” dropdown menu Select “c4: Food” from the “Column used as factor 1” dropdown menu 	<p>Metadata file</p> <p>2: metadata_OverlappedPE.txt ▼</p> <p>Column used for X-axis</p> <p>c5: Replicate_Group ▼ ↺</p> <p>Column used as legend</p> <p>c3: time ▼ ↺</p> <p>Column used as legend</p> <p>c4: Food ▼ ↺</p>
Step 6: Phyloseq Abundance taxonomy	<ul style="list-style-type: none"> Select metadata file (e.g., metadata_OverlappedPE.txt) from “Metadata file” the dropdown menu Select “c5: Replicate_Group” from the “Column used for X-axis” dropdown menu Select “c4: Food” from the “Column used as legend” dropdown menu Click on “Phylum” in the “select a taxonomy rank” 	<p>Metadata file</p> <p>2: metadata_OverlappedPE.txt ▼</p> <p>Column used for X-axis</p> <p>c5: Replicate_Group ▼ ↺</p> <p>Column used as legend</p> <p>c4: Food ▼ ↺</p>

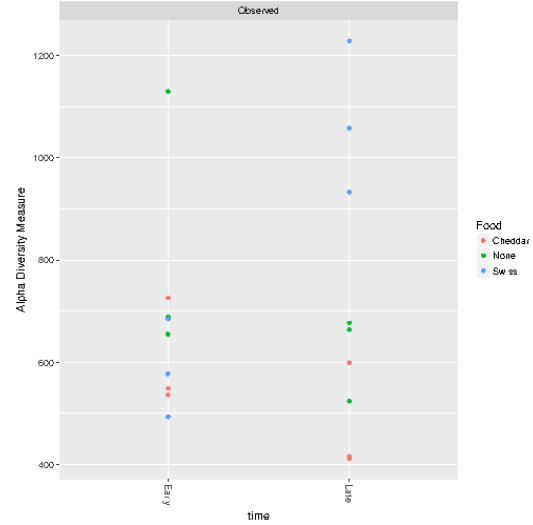
Step	Actions	Screenshot
		<p>select a taxonomy rank</p> <p> <input type="radio"/> Kingdom <input checked="" type="radio"/> Phylum <input type="radio"/> Class <input type="radio"/> Order <input type="radio"/> Family <input type="radio"/> Genus <input type="radio"/> Species </p>
Step 7: Symmetric Plot	<ul style="list-style-type: none"> Select metadata file (e.g., metadata_OverlappedPE.txt) from "Metadata file" the dropdown menu Select green genes annotation file "gg_13_5_taxonomy" from the "OUT taxonomy file" dropdown menu Select "Phylum" under "Select a taxonomy rank" Check the checkbox under "is the data normalised?" Select "c3: time" from the "Variable to compare" dropdown menu Type "Early, Late" in the textbox under "Fill in two comparable group separated by comma" 	<p>Metadata file</p> <p>2: metadata_OverlappedPE.txt ▼</p> <p>Observation Metadata Tabular File</p> <p>3: gg_13_5_taxonomy.txt ▼</p> <p>Select a taxonomy rank</p> <p>Phylum </p> <p>is the data normalised?</p> <p><input checked="" type="checkbox"/> </p> <p>Variable to compare</p> <p>c3: time ▼ </p> <p>Fill in two comparable group separated by comma</p> <p>Early, Late </p>
Step 8: Phyloseq Network Plot	<ul style="list-style-type: none"> Select green genes annotation file "gg_13_5_taxonomy" from the "Observation file" dropdown menu Check the checkbox under "is the data normalised?" Select metadata file (e.g., metadata_OverlappedPE.txt) 	<p>Observation Metadata Tabular File</p> <p>3: gg_13_5_taxonomy.txt ▼</p> <p>is the data normalised?</p> <p><input checked="" type="checkbox"/> </p>

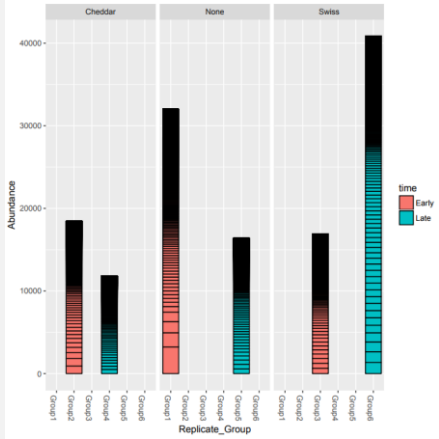
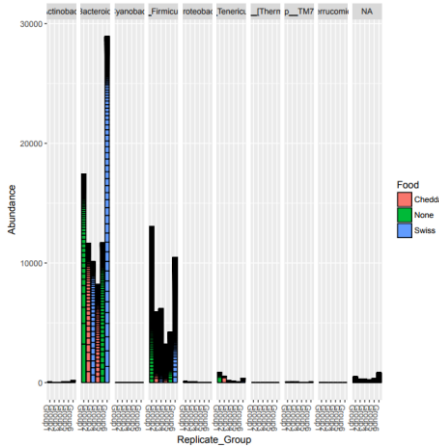
Step	Actions	Screenshot
	<p>from the “Metadata file” dropdown menu</p> <ul style="list-style-type: none"> • Select “c5: Replicate_Group” from the “Select a group for correlation calculation” dropdown menu • Select “c4: Food” from the “Column used as legend” dropdown menu 	<p>The screenshot shows the GVL-Galaxy interface with three dropdown menus. The first menu, labeled 'Metadata file', has '2: metadata_OverlappedPE.txt' selected. The second menu, labeled 'Select a group for correlation calculation', has 'c5: Replicate_Group' selected. The third menu, labeled 'Column used as legend', has 'c4: Food' selected and a refresh icon to its right.</p>

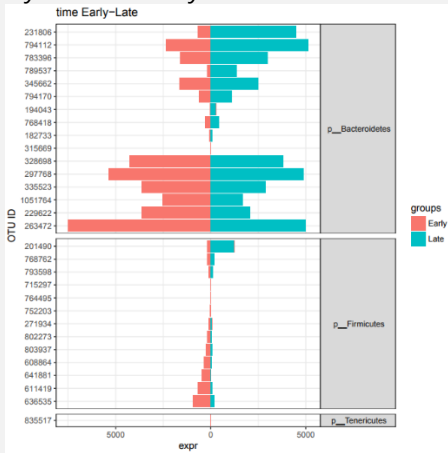
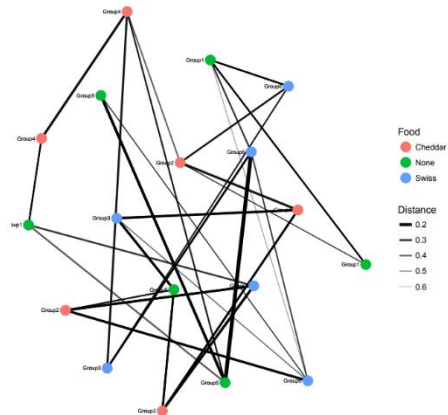
5 Expected output

The following table provides a list of the outputs from each step. Most outputs are intermediate files required for the following step in the workflow. Most of the time you will only be interested in the output from some steps e.g. after quality checking, after filtering steps, after plotting.

Step	Brief Description	Output files
Step 1: convert BIOM	This step converts the raw count input into a BIOM file following the naming structure shown on the right.	<i>"Convert BIOM on data XXXX and XXXX"</i>
Step 2: BIOM metadata	This step takes in two inputs: 1) BIOM file from step 1 and 2) Annotation file from GreenGenes that we imported from the Shared data library It generates an annotated BIOM file with information about the study (metadata) and the taxonomy details for the counts.	<i>"BIOM metadata on data XXXX and XXXX"</i>
Step 3: DESeq2	This step takes the annotated BIOM file from step 2 and generates two output files: 1) Normalised count table and 2) A table listing the significant OTU results between the groups selected for comparison (in this tutorial <i>"time"</i> was selected)	1) <i>"DESeq2 Normalised Table.txt"</i> 2) <i>"DESeq2 DE.txt"</i>

Step	Brief Description	Output files
Step 4: Phyloseq Richness	<p>This step creates a biodiversity abundance plot using the R phyloseq package.</p> <p>The plot generated depends on the settings for the Column used for X-axis and the categories used for the legends. The vertical (y) axis is the abundance values.</p>	<p><i>"Phyloseq Richness.html"</i></p>  <p>In this tutorial, we see the overall abundance (count) for each sample as represented by a dot. The X-axis shows the replicate group.</p>

Step	Brief Description	Output files
<p>Step 5: Phyloseq Abundance plot</p>	<p>This is an abundance plot of all samples in different time (early versus late). The horizontal (x) axis are the samples and the vertical (y) axis is the abundance.</p> <p>The stacked bar shows the abundance values of each OTU from greatest to least separated by a horizontal line.</p>	<p><i>"Phyloseq Abundance plot.html"</i></p> 
<p>Step 6: Phyloseq Abundance taxonomy plot</p>	<p>This is an abundance plot of all samples in different food group under the taxonomy "phylum" as selected in Step 6 of the workflow.</p> <p>The horizontal (x) axis are the samples and the vertical (y) axis is the abundance. The stacked bar shows the abundance values of each OTU from greatest to least separated by a horizontal line.</p>	<p><i>"Phyloseq Abundance taxonomy.html"</i></p> 

Step	Brief Description	Output files
Step 7: Symmetric plot	This symmetric plot shows the normalised counts abundance between the two time points (early vs late). The results shown is only for the dataset under taxonomy “phylum” as selected in step 7 of the workflow.	<p>“Symmetric Plot SymmetricPlot.html”</p>  <p>The figure is a symmetric plot titled "time Early-Late". The y-axis is labeled "OTU ID" and lists various OTU numbers. The x-axis is labeled "expr" and ranges from -5000 to 5000. The plot shows horizontal bars for each OTU, with red bars representing the "Early" group and blue bars representing the "Late" group. The bars are symmetric around the zero line. The plot is divided into two main sections: the top section is labeled "p_Bacteroidetes" and the bottom section is labeled "p_Firmicutes". A legend on the right indicates "groups" with "Early" in red and "Late" in blue.</p>
Step 8: Phyloseq Network plot	This plot shows a correlation network of samples based on the microbiome profiles using the normalised dataset.	<p>“Phyloseq Network Plot.html”</p>  <p>The figure is a network plot showing correlations between samples. The nodes are colored based on the "Food" category: red for "Checkdair", green for "None", and blue for "Swiss". The edges represent the "Distance" between samples, with line thickness corresponding to distance values: 0.2 (thickest), 0.3, 0.4, 0.5, and 0.6 (thinnest). The plot shows a complex network of connections between samples, with some nodes having many connections and others having fewer.</p>

6 Version History

Version	Date	Modified by	Description
1.0	2017-09-20	QFAB (Mike, Xin-Yi)	<ul style="list-style-type: none">Initial version