16S rRNA microbial biodiversity analysis using GVL-Galaxy

Part 2: workflow for analysing overlapping paired-end reads

Version 1.2

Implemented by QFAB Bioinformatics | QCIF

Contents

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, a protein synthesis machinery that is highly conserved. The changes in the 16S rRNA sequence is commonly used as an indication of bacterial evolution and to study phylogeny. This information is leveraged when profiling the 16S rRNA to help researchers identify bacterial species in given samples. Known bacterial species or strains are matched against an annotated 16S database, while those that do not match are considered novel sequences.

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 techniques commonly used in NGS. In a single-end protocol, the sequencer will only sequence from one end of a fragment. In a paired-end protocol, a fragment will be sequenced from both ends. If the sequenced ends 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 protocol that is used for sequencing: (1) overlap-PE and (2) nonoverlap-PE. 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.

1.1 Genomics Virtual Lab - Galaxy

The workflow is setup using Galaxy, which has been deployed using the Genomics Virtual Lab (GVL) platform (version 4.1).

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_biodivesity_BIOM	This workflow performs the differential analysis and generates a number of visualisations.

This tutorial covers workflow 2 16S_biodiversity_for_overlapPE.

CAUTION:

This metagenomic 16S workflow implemented in Galaxy expects paired-end fastq files with following specified filename format. All the input FASTQ files must be in the format:

FILENAME_R1.fastq and FILENAME_R2.fastq

Where *FILENAME* is the name of the library (e.g. lung, brain etc); *R1* is the forward end and *R2* is the reverse end.

3 Dataset

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 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

The sequences was generated using Illumina MiSeq sequencer using PE-sequencing with reads of 2 x 250bp. There are 18 pairs of fastq files, which is a subset of the original dataset[1]. We have already included the required input files as part of the Galaxy Data Libraries so you do not need to download it separately.

Table 3 Overlapping paired-end data filename.

Library	Filename(Forward)	Filename(Reverse)
F3D0	F3D0_R1.fastq	F3D0_R2.fastq
F3D1	F3D1_R1.fastq	F3D1_R2.fastq
F3D2	F3D2_R1.fastq	F3D2_R2.fastq
F3D3	F3D3_R1.fastq	F3D3_R2.fastq
F3D5	F3D5_R1.fastq	F3D5_R2.fastq
F3D6	F3D6_R1.fastq	F3D6_R2.fastq
F3D7	F3D7_R1.fastq	F3D7_R2.fastq
F3D8	F3D8_R1.fastq	F3D8_R2.fastq
F3D9	F3D9_R1.fastq	F3D9_R2.fastq
F3D141	F3D141_R1.fastq	F3D141_R2.fastq
F3D142	F3D142_R1.fastq	F3D142_R2.fastq
F3D143	F3D143_R1.fastq	F3D143_R2.fastq
F3D144	F3D144_R1.fastq	F3D144_R2.fastq
F3D145	F3D145_R1.fastq	F3D145_R2.fastq
F3D146	F3D146_R1.fastq	F3D146_R2.fastq
F3D147	F3D147_R1.fastq	F3D147_R2.fastq
F3D148	F3D148_R1.fastq	F3D148_R2.fastq
F3D149	F3D149_R1.fastq	F3D149_R2.fastq

[1] The original dataset can be downloaded from https://zenodo.org/record/165147#.WYvbjfmqpBc

4 16S biodiversity analysis for overlapping paired-end data workflow

PRE-REQUISITE:

Before using this workflow the dataset needs to be a *List of Dataset Pairs***.** See the **01_16S_OverlapDetection_tutorial** for a step-by-step guide on how to prepare the dataset ready for analysis.

This workflow shown in Figure 1 is designed to process overlapping paired-end datasets. If you do not know whether your dataset contains overlaps or not, see the **01_16S_overlapDetection_tutorial** guide to evaluate the overlapping status of your dataset first. As a guideline, if at least 50% of your reads are overlapping, you can run this tutorial. If less than 50% of the reads are overlapping then we suggest you run workflow 3 (see the **03_16S_nonOverlapPE_tutorial** guide).

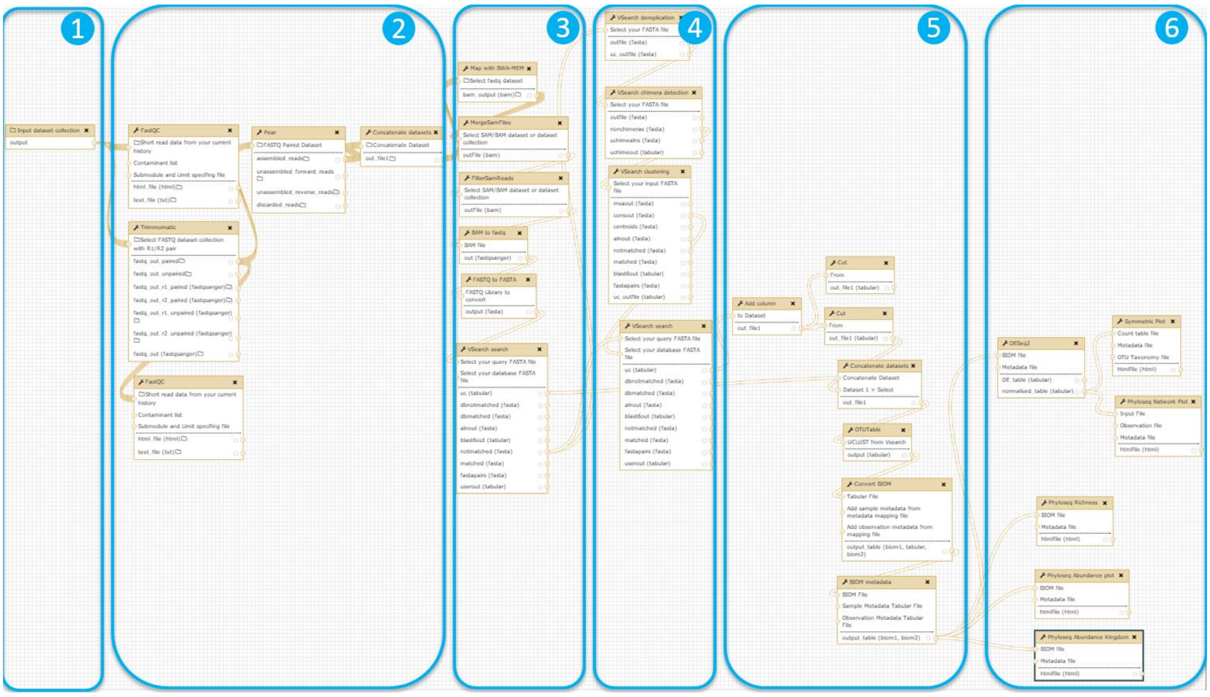


Figure 1 16S workflow for analysing the biodiversity in overlapping paired-end data.

The following table provides a brief overview of the components in the workflow above.

Box	Description
1. Input dataset collection	Dataset collection type
2. Sequence assessment, cleaning and merging	This section of the workflow performs quality checking of the raw data, trimming and removing adapters, and merge overlapping PE reads.
3. Remove host genome and search against known data (eg., Greengenes)	This section of the workflow filters out reads that mapped to the human genome (hg19). The remaining reads are then mapped against a database of known sequences (e.g. Greengenes). The remaining unknown sequences are passed to the next section.
4. Identify potential novel OTUs	This section of the workflow takes the unknown reads and performs collapsing of duplicate sequences, chimera detection and cluster sequences to build novel Operational Taxonomic Unit (OTU) references. The reads that are not mapped to the Greengenes database are used to map against these newly created novel OTUs.
5. BIOM File conversion	This section of the workflow combines the output from component 3 and 4 to create a table in the UCLUST format. Essentially this combined output returns the results whether a match (hit) exists for a query read against a database. Further information about the UCLUST format is available from the site[2]. The UCLUST table is transformed to a count table, and then, converted into a BIOM format.
6. Run differential analysis testing and plots creation	This last section of the workflow generates the following: • Normalised table • Diversity plot • Abundance plots • Symmetric plot • Network plot

[2] http://www.drive5.com/uclust/uclust_userguide_1_1_579.html#_Toc257997686

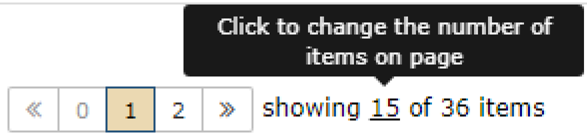
4.1 Workflow and data preparation

We have prepared the input data already and is shared as a *Data Library* in Galaxy. Follow section 4.1.1 below to import the data to your account.

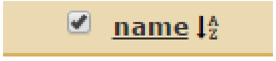
A reminder, if you are unfamiliar with the Galaxy interface, we recommend you have a look at this **Introduction to Galaxy** quick start guide.

4.1.1 Import data

1. Click **Shared Data** from the top menu
2. Click **Data Libraries** on the dropdown menu
3. Click **Tutorial data: Overlapped PE dataset**
4. As the page can only show 15 items at a time, this will make importing data slow. You can change the number of items that is displayed on one page by clicking on the **15**. A window will appear, enter **36** and press enter. The page should update to show all 36 items on one page.



5. Check the checkbox next to **name**

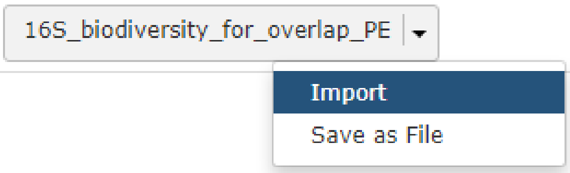


6. Click **to History** from the top menu
7. Type a history name e.g., **16S_biodiversity_for_overlap_PE**
8. Repeat steps 1 to 5 to import the Greengenes taxonomy file (**gg_13_5_taxonomy.txt** from the **GreenGenes** folder) into the same history that was created in step 6 (**16S_biodiversity_for_overlap_PE**).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

4.1.2 Import workflow

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

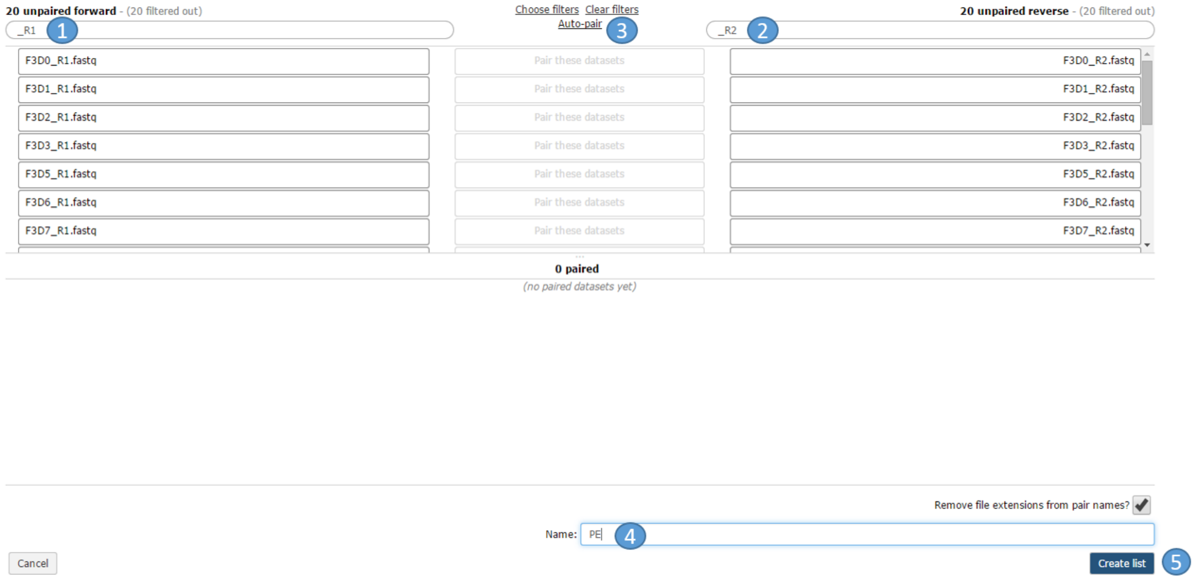
1. Click **Shared Data** from the top menu
2. Click **Workflows** on the dropdown menu
3. Click **16S_biodiversity_for_overlap_PE**
4. Click **Import**



4.2 Creating a List of Dataset Pairs

Note that in this dataset all forward reads have the "_R1.fastq" suffix on the filename and all reverse reads have the "_R2.fastq" suffix. This will become apparent in the step 5.

1. Make sure the correct history is selected.
2. Click on the ☑ icon near the top of the history panel, just under the title.
3. Click on **All**
4. Click on **For all selected ...** > **Build List of Dataset Pairs**
5. Follow the steps in Figure 2
   - 1 = type "**_R1**" (as determined by your file naming format)
   - 2 = type "**_R2**" (as determined by your file naming format)
   - 3 = click on **Auto pair**
   - 4 = type "**PE**"
   - 5 = click on **Create list**



20 unpaired forward - (20 filtered out)

20 unpaired reverse - (20 filtered out)

_R1

_R2

F3D0_R1.fastq	Pair these datasets	F3D0_R2.fastq
F3D1_R1.fastq	Pair these datasets	F3D1_R2.fastq
F3D2_R1.fastq	Pair these datasets	F3D2_R2.fastq
F3D3_R1.fastq	Pair these datasets	F3D3_R2.fastq
F3D5_R1.fastq	Pair these datasets	F3D5_R2.fastq
F3D6_R1.fastq	Pair these datasets	F3D6_R2.fastq
F3D7_R1.fastq	Pair these datasets	F3D7_R2.fastq

0 paired

(no paired datasets yet)

Remove file extensions from pair names? ☑

Name: PE

Figure 2 Build a list of dataset pairs

4.3 Rename Sequence header

This workflow is designed to work with a collection of FASTQ files from a study. Throughout its analysis, it will concatenate all reads into a master file, at which point we loose the information of which library the reads originate from. In order to address this issue, we have developed a tool called **reheader** to append the library filename to the end of the FASTQ identifier. This allows us to track which reads belong to which library and perform the quantification step.

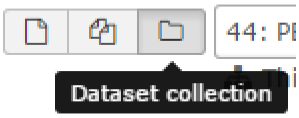
A FASTQ header before applying the **reheader** tool will look like this:

@M00967:43:000000000-A3JHG:1:1101:18327:1699 1:N:0:188

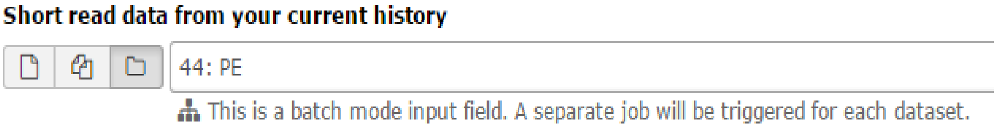
A FASTQ header after applying the **reheader** tool will look like this:

@M00967:43:000000000-A3JHG:1:1101:18327:1699_**F3D0/1**

1. Click on **reheader** from the *Tools: Data processing* panel on the left-hand panel
2. Select the **Data collection** icon



3. Select the newly created *Dataset Pairs* (e.g., **44:PE**). Yours may be named differently.



Short read data from your current history

44: PE

This is a batch mode input field. A separate job will be triggered for each dataset.

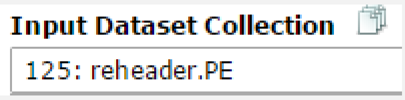
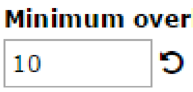
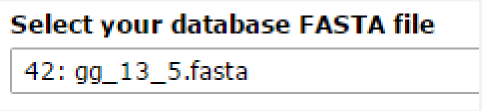
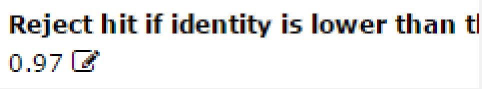
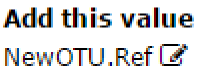
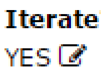
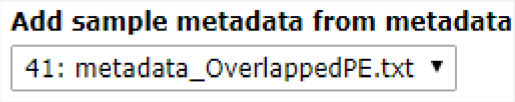
4. Click **Execute**

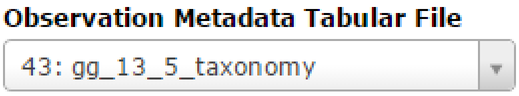
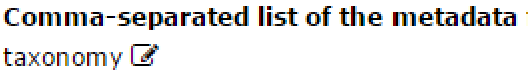
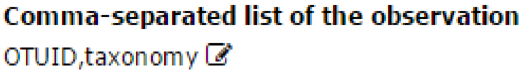
4.4 Run the workflow

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

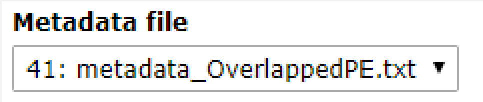
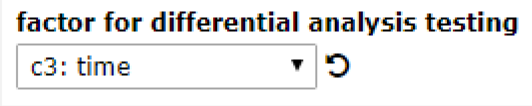
4.5 Specify parameters

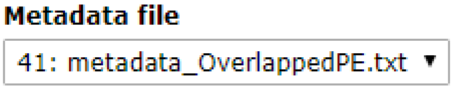
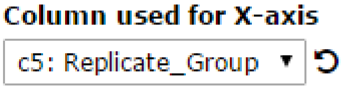
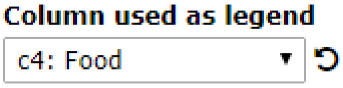
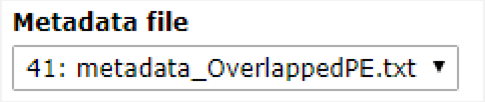
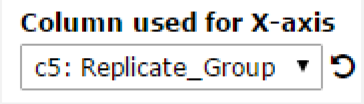
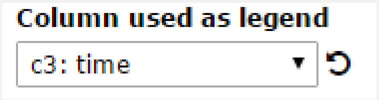
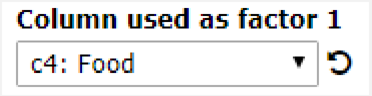
While the entire workflow consists of 29 total steps, you do not need to specify the parameters for all steps. Follow the table below, which highlights the parameters that need modification.

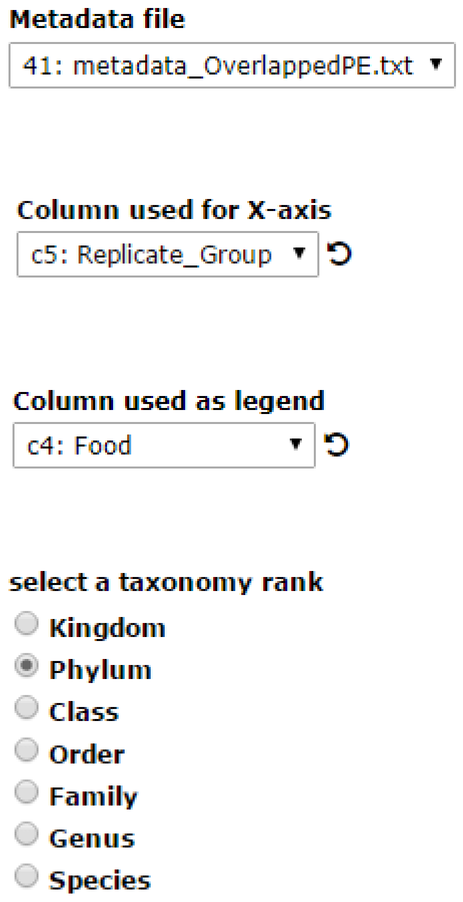
Step	Actions	Screenshot
Step 1: Input dataset collection	Select **reheader.PE** from the "*Input Dataset Collection*"	 Input Dataset Collection 125: reheader.PE
Step 4: Pear	Type "**10**" in the textbox under "*Minimum overlap size*" You can change this value if 10 base pairs is too strict or relaxed.	 Minimum overlap size 10
Step 12: Vsearch search	- Select **gg_13_5.fasta** from the "*Select your datasbase FASTA file*" dropdown menu. *Note:* If you have another database of FASTA sequences, you can use that here instead. - Type "**0.97**" in the textbox under "*Reject hit if identity is lower than this value*" - Make sure "*UCLUST-like output*" is "**Yes**"	 Select your database FASTA file 42: gg_13_5.fasta  Reject hit if identity is lower than tl 0.97
Step 17: Add column	- Type "**NewOTU.Ref**" in the textbox under *"Add this value"* - Select "**YES**" from the "*Iterate?*" dropdown menu	 Add this value NewOTU.Ref  Iterate? YES
Step 22: Convert BIOM	- Select metadata file (e.g., **metadata_OverlappedPE.txt**) from the *"Add sample metadata from metadata mapping file"* dropdown menu	 Add sample metadata from metadata 41: metadata_OverlappedPE.txt

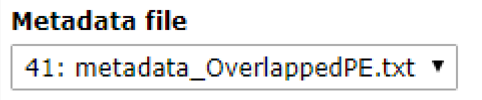
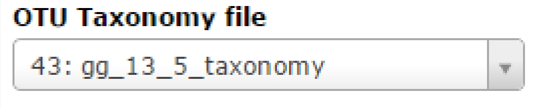
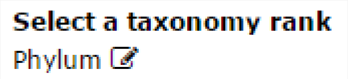
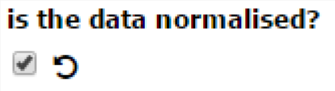
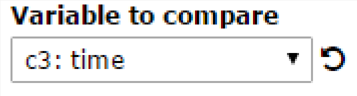
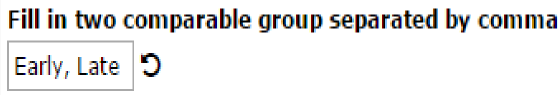
Step	Actions	Screenshot
Step 23: BIOM metadata	• Select green *genes* annotation file "**gg_13_5_taxonomy**" from the "*Observation Metadata Tabular File*" dropdown menu *Note:* If you used another database in step 12, you should also have its corresponding taxonomy annotation file which you use ere. This is a two column file with the first column being the OTU id and the second column is the taxonomy name. • 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*"	 Observation Metadata Tabular File 43: gg_13_5_taxonomy  Comma-separated list of the metadata taxonomy  Comma-separated list of the observation OTUID,taxonomy

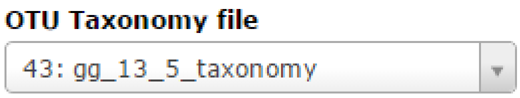
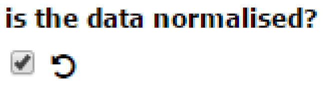
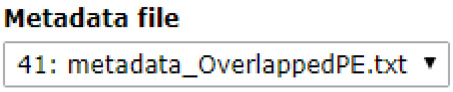
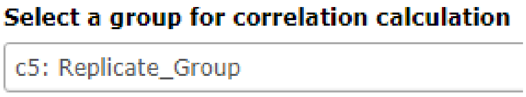
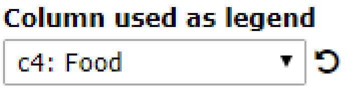
From Step 24 onwards, you can change the parameters for different comparisons depending on your study. For this tutorial we are interested in looking at the differences between the microbiome of the mice in the *early* versus *late* time.

Step	Actions	Screenshots
Step 24: DESeq2	• Select metadata file (e.g., metadata_OverlappedPE.txt) from the dropdown menu • Select "**c3: time**" from the "*factor for differential analysis testing*"	 Metadata file 41: metadata_OverlappedPE.txt  factor for differential analysis testing c3: time

Step	Actions	Screenshots
Step 25 : Phyloseq Richness	- 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	 Metadata file 41: metadata_OverlappedPE.txt ▼  Column used for X-axis c5: Replicate_Group ▼  Column used as legend c4: Food ▼
Step 26 : Phyloseq Abundance plot	- 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	 Metadata file 41: metadata_OverlappedPE.txt ▼  Column used for X-axis c5: Replicate_Group ▼  Column used as legend c3: time ▼  Column used as factor 1 c4: Food ▼

Step	Actions	Screenshots
Step 27 : Phyloseq Abundance Kingdom	- 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"	 **Metadata file** 41: metadata_OverlappedPE.txt **Column used for X-axis** c5: Replicate_Group **Column used as legend** c4: Food **select a taxonomy rank** ○ Kingdom ● Phylum ○ Class ○ Order ○ Family ○ Genus ○ Species

Step	Actions	Screenshots
Step 28 : Symmetric Plot	- Select metadata file (e.g., metadata_OverlappedPE.txt) from "*Metadata file*" the dropdown menu	 **Metadata file** 41: metadata_OverlappedPE.txt ▾
	- Select green genes annotation file "**gg_13_5_taxonomy**" from the "*OUT taxonomy file*" dropdown menu	 **OTU Taxonomy file** 43: gg_13_5_taxonomy ▾
	- Select "**Phylum**" under "*Select a taxonomy rank*"	 **Select a taxonomy rank** Phylum
	- Check the checkbox under "**is the data normalised?**"	 **is the data normalised?** ☑
	- Select "**c3: time**" from the "*Variable to compare" dropdown menu*"	 **Variable to compare** c3: time ▾
	- Type "**Early, Late**" in the textbox under "*Fill in two comparable group separated by comma*"	 **Fill in two comparable group separated by comma** Early, Late

Step	Actions	Screenshots
Step 29 : Phyloseq Network Plot	- 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) from the “*Metadata file*” dropdown menu - 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	 OTU Taxonomy file 43: gg_13_5_taxonomy  is the data normalised? ☑  Metadata file 41: metadata_OverlappedPE.txt  Select a group for correlation calculation c5: Replicate_Group  Column used as legend c4: Food

5 Expected output

The following table provides a list of the outputs from each step when using the **16S_biodiversiy_for_overlapPE** workflow. The majority of outputs are intermediate files required for the following step in the workflow and most of the time, you will only be interested in the output from some steps, for example, after quality checking, after filtering steps, after plotting.

Step	Brief description	Example output files
n/a, before the workflow we need to run **reheader**	The reheader tool generates one FASTQ file with the sequence header renamed and one log file for each input. When the input is a data collection of N files, the output is 2 data collections: 1) reheader.PE (holds FASTQ output) 2) reheader.PE.log (holds log outputs) Think of a data collection like a folder, with a nested structure like the one shown on the right. Only the "*reheader.PE*" collection is used in subsequent steps.	• reheader.PE ○ Sample F3D0 ▪ Forward ▪ Reverse ○ Sample F3D1 ▪ Forward ▪ Reverse ○ Etc… • reheader.PE.log ○ Sample F3D0 ▪ Forward ▪ Reverse ○ Sample F3D1 ▪ Forward ▪ Reverse ○ Etc… In this tutorial there were 18 x 2 = 36 total input files, so there will be 36 x 2 = 72 output files.

Step	Brief description	Example output files
Step 1: Input dataset collection	This component in the workflow is designed to take in a data collection as an input (e.g., the output of reheader.PE above)	n/a
Step 2: FastQC	Quality checks for each library in a HTML or text format. Like the reheader tool, the FastQC tool also generates a 2 data collection as output: 1) *"FastQC on collection:RawData"* 2) *"FastQC on collection:Webpage"* The collections have the following hierarchical structure shown on the right.	• FASTQC on collection Webpage (html) ○ Sample F3D0 ▪ Forward ▪ Reverse ○ Sample F3D1 ▪ Forward ▪ Reverse ○ Etc… • FASTQC on collection RawData (text) ○ Sample F3D0 (text file) ▪ Forward ▪ Reverse ○ Sample F3D1 ▪ Forward ▪ Reverse ○ Etc…

Step	**Brief description**	**Example output files**
Step 3: Trimmomatic	Output after removing adapters and low quality reads. The dataset is separated into *paired* and *unpaired* data. The output is two data collections: 1) *Trimmomatic across collection XX* 2) *Trimmomatic across collection XX* **NOTE:** both data collections will have the same label but the first one is for paired data and the second collection is for unpaired data.	The collections have the following hierarchical structure: • *Trimmomatic across collection XXXX* (paired) o Sample F3D0 ▪ Forward ▪ Reverse o Sample F3D1 ▪ Forward ▪ Reverse o Etc… • *Trimmomatic across collection XXXX* (unpaired) o Sample F3D0 ▪ Forward ▪ Reverse o Sample F3D1 ▪ Forward ▪ Reverse o Etc…
Step 4: Pear	One FASTA output file is generated per PE input, after merging the overlapping ends and one log file is also generated per input. The output is two data collections: 1) *"Pear on collection XXXX:Assembled reads"* *2) "Pear on collection XXXX.log"*	The collections have the following hierarchical structure: • *Pear on collection XXXX:Assembled reads* o Sample F3D0 o Sample F3D1 o etc • *Pear on collection XXXX.log* o Sample F3D0 o Sample F3D1 o etc

Step	Brief description	Example output files
Step 5: FASTQC	**Note:** This step happens in parallel with Step 4 and takes the paired-end output from Step 3 output 1. The outputs are the same as Step 2:FastQC above	Same as Step 2
Step 6: Concatenate datasets	This component is used to concatenate FASTA from PEAR's output and will generate one data collection with the naming structure shown on the right	• *Concatenate datasets on collection XXXX* ○ Sample F3D0 ○ Sample F3D1 ○ etc
Step 7: BWA-MEM	The reads are mapped to the host genome and the output is a data collection of BAM files with the naming structure shown on the right. The number of outputs depends on the number of inputs.	• *Map with BWA-MEM on collection XXXX (mapped reads in BAM format)* ○ Sample F3D0 ○ Sample F3D1 ○ etc
Step 8: MergeSamFiles	All BAMs in step 7 are merged into one. The output has the naming structure shown on the right.	*MergeSamFiles on data XXX and others: Merged BAM dataset*
Step 9: FilterSamReads	Reads that mapped to the host genome in Step 7 are removed, generated one BAM file with the naming structure shown on the right.	*"FilterSamReads on data XXX: filtered BAM"*
Step 10: BAM to fastq	The BAM file (step 9) is converted to a FASTQ file with the naming structure shown on the right.	*"BAM to fastq on data XXX"*
Step 11: FASTQ to FASTA	The FASTQ file (step 10) is converted to a FASTA file with the naming structure shown on the right.	*"FASTQ to FASTA on data XXX"*

Step	**Brief description**	**Example output files**
Step 12: Vsearch search	This step takes in an input query and a database file. It looks for a match for each query sequence in the database and generates three output files with the naming structure shown on the right. 1) Is a tabular file following the UCLUST format 2) Is a FASTA file of the query sequences with a match in the database 3) Is a FASTA file of the query sequences without a match in the database Output 3 is used in the next step.	1) *"VSearch search on data XXXX and XXXX: UCLUST like output"* 2) *"VSearch search on data XXXX and XXXX: Matching query sequences*" 3) *"VSearch search on data XXXX and XXXX: Non-matching query sequence"*
Step 13: Vsearch dereplication	This steps remove all duplicate sequences and generates one FASTA output listing the unique sequences.	*"Vsearch dereplication on data XXXX"*
Step 14: Vsearch chimera detection	The tool searches for possible chimera sequences and generates two FASTA files with the naming structure shown on the right. Only output 2 (non chimera) is used for the next step.	1) *"Vsearch chimera detection on data XXX"* *2) "Vsearch chimera detection on data XXX: Non chimera"*
Step 15: Vsearch clustering	This tool clusters the sequences into groups and generates a FASTA output that contains a list of consensus sequences with the naming structure shown on the right	"*Vsearch clustering on data XXXX: Consensus Sequences*"

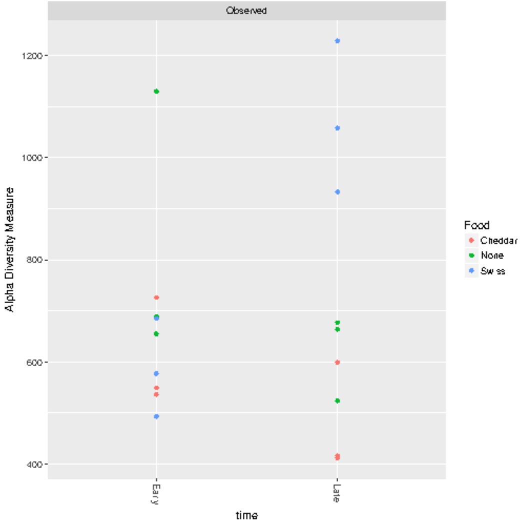
Step	**Brief description**	**Example output files**
Step 16: Vsearch search	Similar to step 12: Vsearch search This time we are searching the remaining reads from Step 12, output 3, against a new database created from Step 15.	1) *“VSearch search on data XXXX and XXXX: UCLUST like output”* 2) *“VSearch search on data XXXX and XXXX: Matching query sequences*“ 3) *“VSearch search on data XXXX and XXXX: Non-matching query sequence”*
Step 17: Add column	This step adds a column (NewOTU.Ref) to the UCLUST output (output 1) from step 16 and generates a new tabular output with the naming structure shown on the right.	“*Add column on data XXXX*”
Step 18: Cut	This step extracts columns 10 (original ref sequence ID) and 11 (new ref ID) from the output from step 17 and generates a new file with the two columns. **Note:** The output of this step can be used as a mapping reference to map the original sequence ID to the new sequence ID. This output is not used in the workflow.	*“Cut on data XXXX”*
Step 19: Cut	This step extracts columns 1-9 and 11 to a new file. Essentially we are using the new reference ID instead of the old reference ID. This output is used in the next step.	*“Cut on data XXXX”*

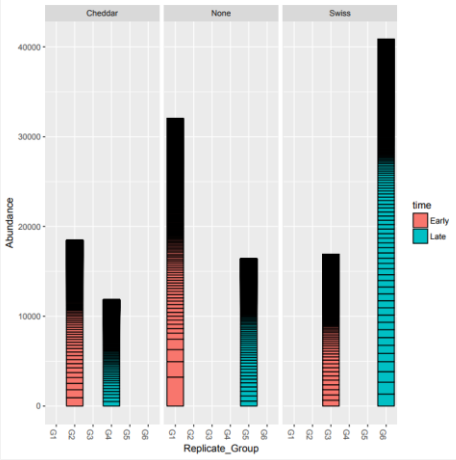
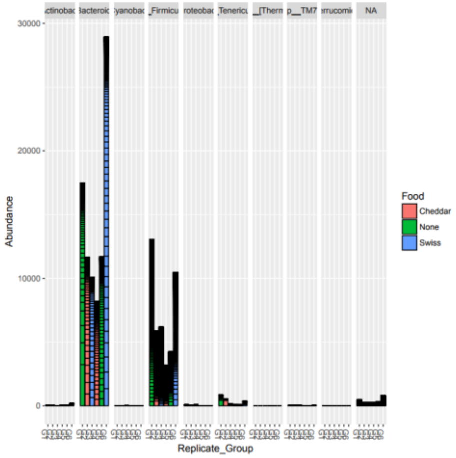
Step	**Brief description**	**Example output files**
Step 20: Concatenate datasets	This step merges the two outputs from step 12 with step 19 and generates a combined tabular file with the naming structure shown on the right. Reminder: Step 12 – is performing a search against Greengenes (known sequences) Step 19 – is the output after performing a search against novel sequences	*"Concatenated datasets on data XXXX and XXXX"*
Step 21: OTUTable	This step converts the tabular output from step 20 into a count table where the rows are OTUId and the columns are SampleIDs. Each cell then indicates the number of times OTU-i appears in Sample-j for example.	*"OTU TABLE Concantenate datasets on data XXXX and XXXX"*

Note: Repeating differential analysis without post-processing dataset

Now the dataset is transformed into an OTU Table in step 21, you can use this output in the fourth workflow (**04_16S_biodiversity_BIOM**) directly to perform other comparisons, without re-running all the data processing steps.

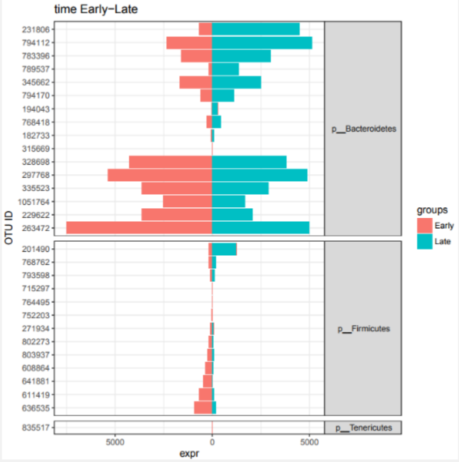
Step	**Brief description**	**Example output files**
Step 22: convert BIOM	This step converts the output from step 21 into a BIOM file following the naming structure shown on the right.	*“Convert BIOM on data XXXX and XXXX”*
Step 23: BIOM metadata	This step takes in two inputs: 1) BIOM file from step 22 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.	*“BIOM metadata on data XXXX and XXXX”*
Step 24: DESeq2	This step takes the annotated BIOM file from step 23 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 “*time”* was selected)	*1) “DESeq2 Normalised Table.txt”* 2) *“DESeq2 DE.txt”*
Step 25: Phyloseq Richness	This step creates a biodiversity abundance plot using the R phyloseq package. 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.	“*Phyloseq Richnness.html*”

Step	Brief description	Example output files
		 In this tutorial, we see the overall abundance (count) for each sample as represented by a dot. The X-axis shows the replicate group.

Step	Brief description	Example output files
Step 26: Phyloseq Abundance plot	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. The stacked bar shows the abundance values of each OTU from greatest to least separated by a horizontal line.	*"Phyloseq Abundance plot.html"* 
Step 27: Phyloseq Abundance taxonomy plot	This is an abundance plot of all samples in different food group under the taxonomy "phylum" as selected in Step 27 of the workflow. 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.	*"Phyloseq Abundance kingdom.html"* 

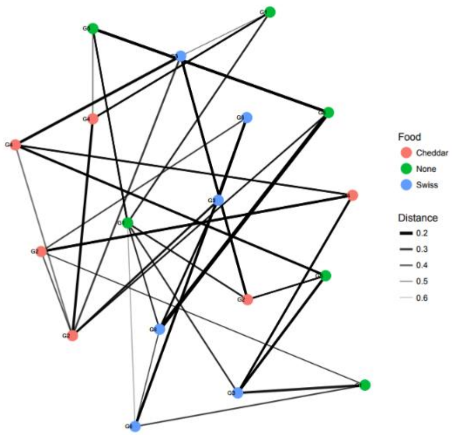
<table>
<tr><th>Step</th><th>Brief description</th><th>Example output files</th></tr>
<tr><td>Step 28:
Symmetric plot</td><td>This symmetric plot shows the normalised counts abundance between the two time points (early vs late). The results shown is only for the dataset under taxonomy “phylum” as selected in step 28 of the workflow.</td><td>

“Symmetric Plot SymmetricPlot.html”



</td></tr>
<tr><td>Step 29:
Phyloseq Network plot</td><td>This plot shows a correlation network of samples based on the microbiome profiles using the normalised dataset.</td><td>

“Phyloseq Network Plot.html”



</td></tr>
</table>

6 Version History

Version	Date	Modified by	Description
1.0	2017-09-12	QFAB (Mike, Xin-Yi)	• Initial version
1.1	2017-09-20	QFAB (Xin-Yi)	• Add in version history table • Formatting and minor edits
1.2	2017-10-10	QFAB (Mike, Xin-Yi)	• Extra step (4) in Section 4.1.1 to change the number of items to show on one page • Add in Section 4.2 Creating a List of Dataset Pairs