16S rRNA microbial diversity using GVL-Galaxy

Part 3: workflow for analysing non-overlapping paired-end reads

Version 1.1

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. 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_biodivesity_BIOM	Handle BIOM file and generate plots

This tutorial covers workflow 3 16S rRNA biodiversity analysis for nonoverlap PE.

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 data for this tutorial was generated using the Illumina MiSeq sequencer, using PE-sequencing with reads of 2 x 300bp, however, this is a non-overlapping dataset. There are 24 pairs of fastq files (Table 3) with the metadata provided below in Table 2. There are three replicates per *Tissue-Protein* group which we will be interested in looking at further.

This dataset is used to demonstrate using the third workflow **16S_biodiverisy_for_nonoverlap_PE** that does not require merging of paired end reads using the PEAR[1] tool. Instead the workflow will only perform the analysis using the R1 (or forward) read files only.

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 non-overlapping paired-end dataset

#SampleID	Water	Tissue	Protein	Replicate_Group	Expected_Effect	SeqID
Sample13	No	Lung	None	Group5	No	R13
Sample14	No	Lung	None	Group5	No	R14
Sample15	No	Lung	None	Group5	No	R15
Sample16	No	Lung	Protein3	Group6	Yes	R16
Sample17	No	Lung	Protein3	Group6	Yes	R17
Sample18	No	Lung	Protein3	Group6	Yes	R18
Sample19	No	Lung	Protein2	Group7	Yes	R19
Sample20	No	Lung	Protein2	Group7	Yes	R20
Sample21	No	Lung	Protein2	Group7	Yes	R21
Sample22	No	Lung	Protein1	Group8	Yes	R22
Sample23	No	Lung	Protein1	Group8	Yes	R23
Sample24	No	Lung	Protein1	Group8	Yes	R24
Sample37	No	Feces	None	Group13	No	R37
Sample38	No	Feces	None	Group13	No	R38
Sample39	No	Feces	None	Group13	No	R39
Sample40	No	Feces	Protein3	Group14	Yes	R40
Sample41	No	Feces	Protein3	Group14	Yes	R41
Sample42	No	Feces	Protein3	Group14	Yes	R42

[1] https://sco.h-its.org/exelixis/web/software/pear/

#SampleID	Water	Tissue	Protein	Replicate_Group	Expected_Effect	SeqID
Sample43	No	Feces	Protein2	Group15	Yes	R43
Sample44	No	Feces	Protein2	Group15	Yes	R44
Sample45	No	Feces	Protein2	Group15	Yes	R45
Sample46	No	Feces	Protein1	Group16	Yes	R46
Sample47	No	Feces	Protein1	Group16	Yes	R47
Sample48	No	Feces	Protein1	Group16	Yes	R48

Table 3 non-overlapping paired-end data filename

Library	Filename(Forward)	Filename(Reverse)
Sample13	Sample13_R1.fastq	Sample13_R2.fastq
Sample14	Sample14_R1.fastq	Sample14_R2.fastq
Sample15	Sample15_R1.fastq	Sample15_R2.fastq
Sample16	Sample16_R1.fastq	Sample16_R2.fastq
Sample17	Sample17_R1.fastq	Sample17_R2.fastq
Sample18	Sample18_R1.fastq	Sample18_R2.fastq
Sample19	Sample19_R1.fastq	Sample19_R2.fastq
Sample20	Sample20_R1.fastq	Sample20_R2.fastq
Sample21	Sample21_R1.fastq	Sample21_R2.fastq
Sample22	Sample22_R1.fastq	Sample22_R2.fastq
Sample23	Sample23_R1.fastq	Sample23_R2.fastq
Sample24	Sample24_R1.fastq	Sample24_R2.fastq
Sample37	Sample37_R1.fastq	Sample37_R2.fastq
Sample38	Sample38_R1.fastq	Sample38_R2.fastq
Sample39	Sample39_R1.fastq	Sample39_R2.fastq
Sample40	Sample40_R1.fastq	Sample40_R2.fastq
Sample41	Sample41_R1.fastq	Sample41_R2.fastq
Sample42	Sample42_R1.fastq	Sample42_R2.fastq
Sample43	Sample43_R1.fastq	Sample43_R2.fastq
Sample44	Sample44_R1.fastq	Sample44_R2.fastq
Sample45	Sample45_R1.fastq	Sample45_R2.fastq
Sample46	Sample46_R1.fastq	Sample46_R2.fastq
Sample47	Sample47_R1.fastq	Sample47_R2.fastq
Sample48	Sample48_R1.fastq	Sample48_R2.fastq

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

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 non-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. As a guideline, if less than 50% of your reads are overlapping, you can run this tutorial. If more than 50% of the reads are overlapping then we suggest you run workflow 2 (see the **02_16S_OverlapPE_tutorial** guide).

This workflow is nearly identical to the second workflow **16S_biodiversity_for_overlapPE**, where boxes 1 and 4 to 7, shown in the figure below are the same. Boxes 2 and 3 are slightly different in that this workflow *does not* perform a merging step of the reads. Instead, it is replaced by a concatenating step (blue box 3) where all forward reads from all samples are combined into one FASTQ file.

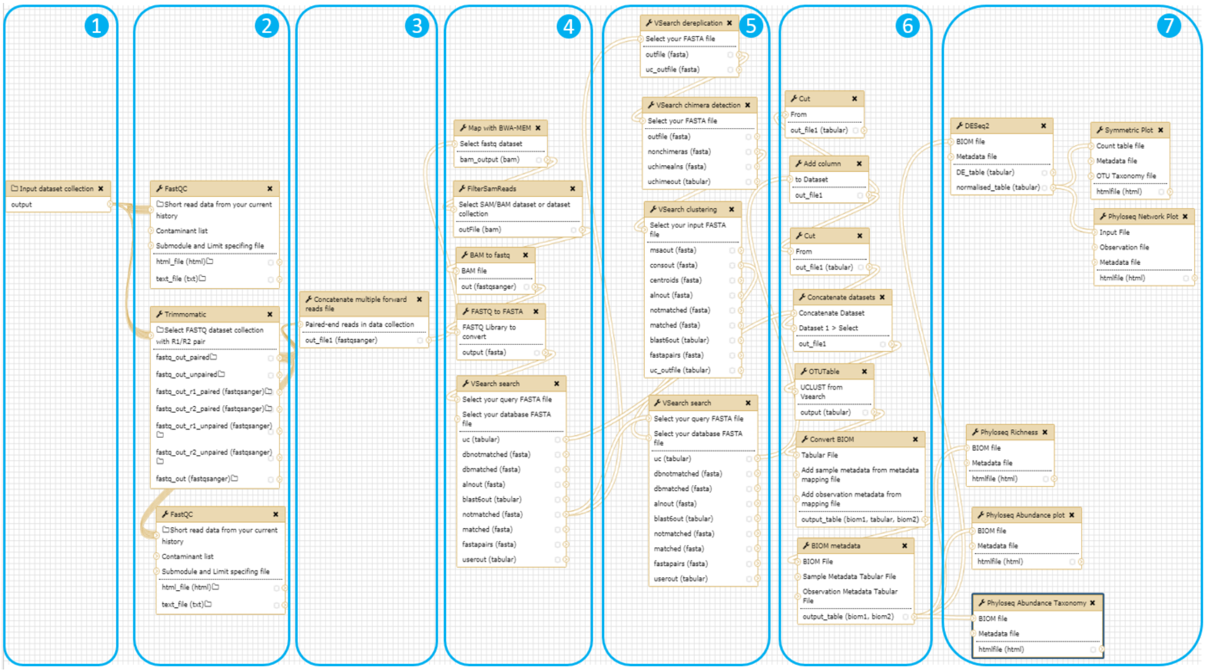


Figure 1 16S workflow without PEAR

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. Concatenate dataset	Combines forward reads of all samples into one FASTQ file.
4. Remove host genome and search against known data (eg., Greengenes)	This section of the workflow filter 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.
5. 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.
6. 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.
7. Run differential analysis testing and plots creation	This last section of the workflow generates the following output for data interpretation: • Normalised table • Diversity plot • Abundance plots • Symmetric plot • Network plot

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

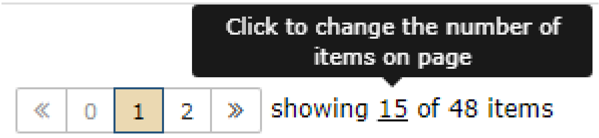
4.1 Workflow data preparation

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.

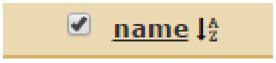
A reminder, if you are unfamiliar with the Galaxy interface, we recommend you have a look at this **Introduction to Galaxy** quick start guide.

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: Non-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 48 and press enter. The page should update and show all 48 items on one page.



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

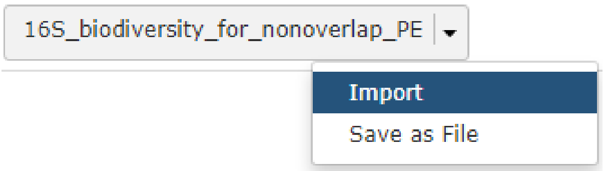


6. Click **to History** from the top menu
7. Type a history name e.g., **16S_biodiversity_for_nonoverlap_PE**
8. Repeat step 1 to step 5 to import Greengenes taxonomy file (**gg_13_5_taxonomy.txt** from the **GreenGenes** folder) into the same history that was created in step 6 (**16S_biodiversity_for_nonoverlap_PE**).
9. Repeat Steps 1 to 5 again, but this time click on the folder **metadata** to import **metadata_NonOverlappedPE.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_nonoverlap_PE"** into your workspace. You can skip this section if you have already imported the workflow before.

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



4.2 Creating a List 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 ("16S_biodiversity_for_nonoverlap_PE") 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**

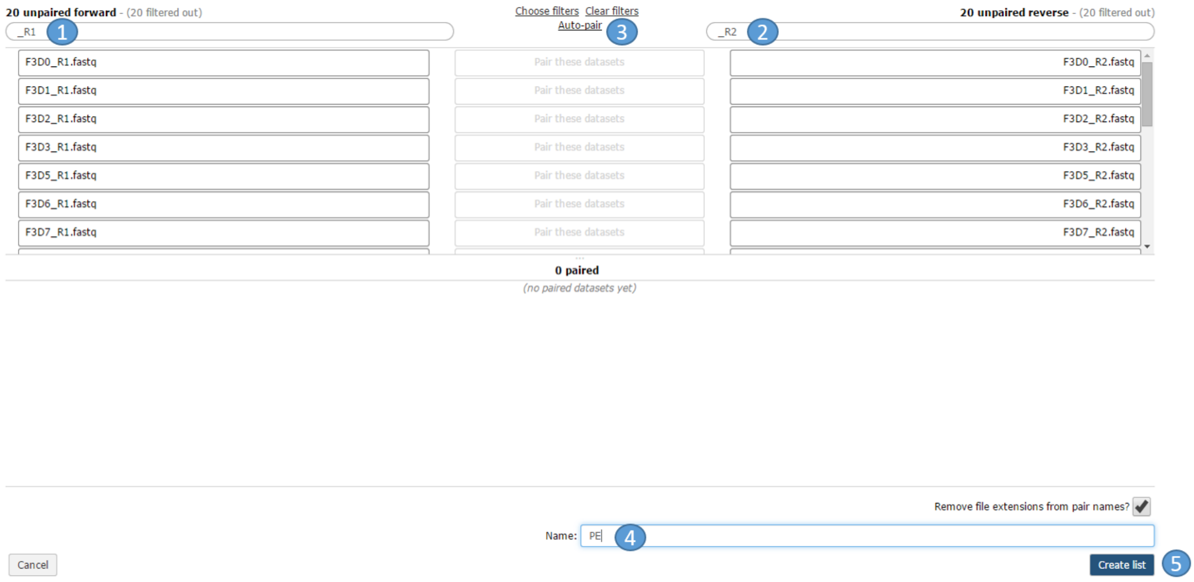


Figure 2 Build a list of dataset pairs

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 originates. 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 read belongs to which library and perform the quantification step.

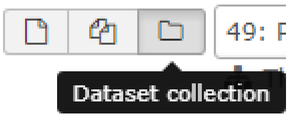
A fastq header before applying the **reheader** tool will look like this:

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

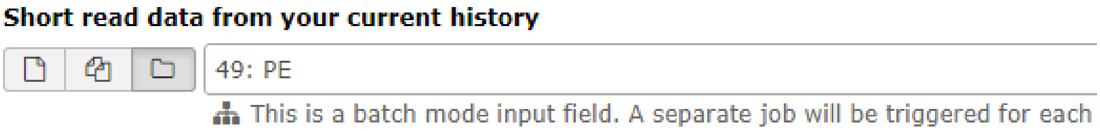
A fastq header after applying the **reheader** tool will look like this:

@M00967:43:000000000-A3JHG:1:1101:18327:1699**_F3D0/1**

1. Click on **reheader** in the *Tools: Data processing* at the left panel in galaxy.
2. Select **Data collection** icon



3. Select the newly created *Dataset Pairs* (e.g., **49:PE**) Yours may be named differently.



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

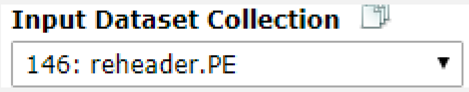
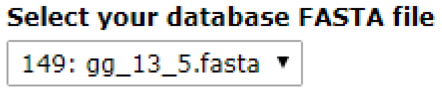
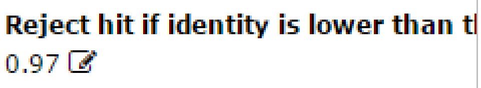
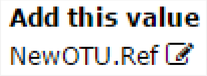
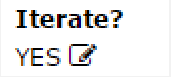
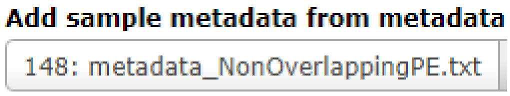
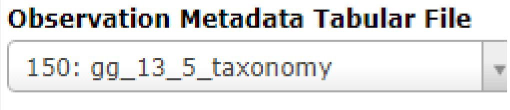
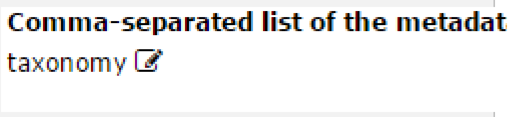
4. Click **Execute**

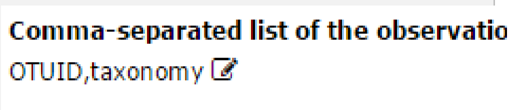
4.4 Run workflow

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

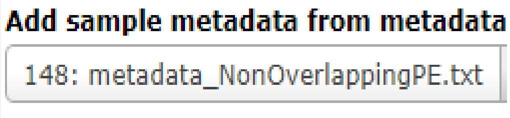
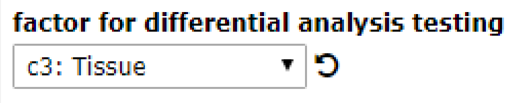
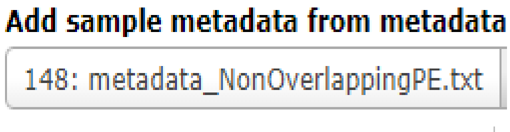
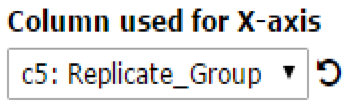
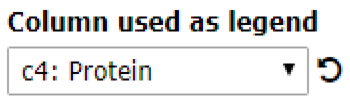
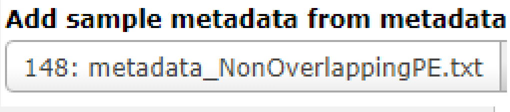
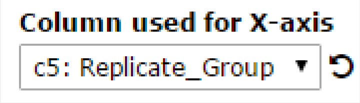
4.5 Specify parameters

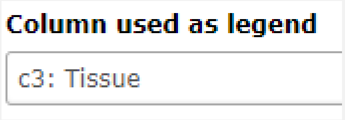
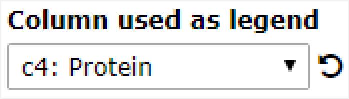
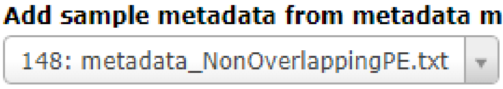
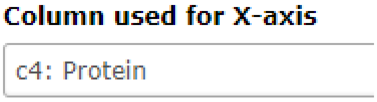
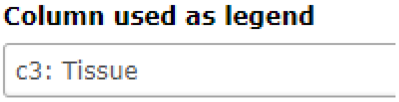
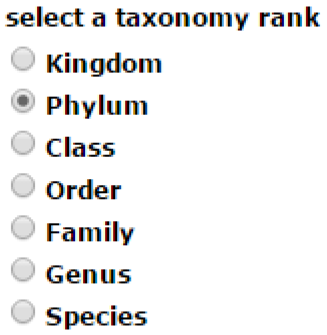
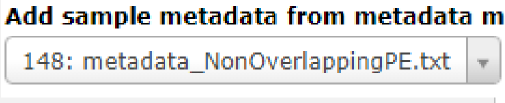
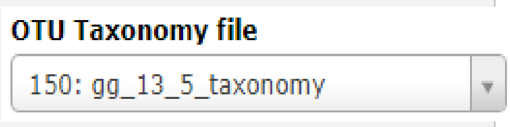
While the entire workflow consists of 27 total steps, you do not need to specify the parameters for all steps. Follow the following step-by-step guide which highlights the parameters that need modification.

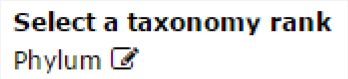
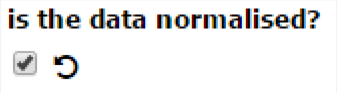
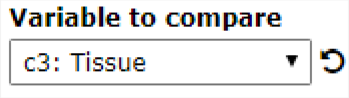
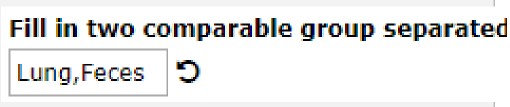
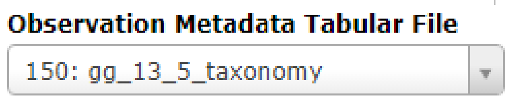
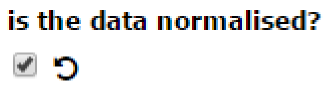
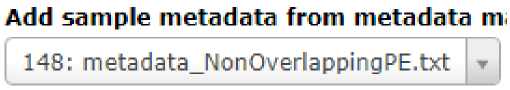
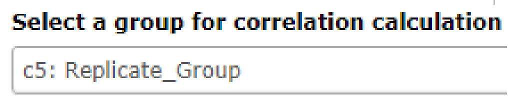
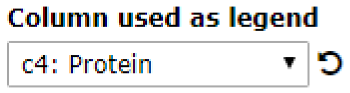
Step	Actions	Screenshot
Step 1: Input dataset collection	- Select **reheader.PE** from the *"Input Dataset Collection"*	 **Input Dataset Collection** 146: reheader.PE
Step 10: Vsearch search	- Select green genes reference sequences **gg_13_5.fasta** from the "*Select your datasbase FASTA file"* dropdown menu - 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** 149: gg_13_5.fasta  **Reject hit if identity is lower than t** 0.97
Step 15: 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 20: 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** 148: metadata_NonOverlappingPE.txt
Step 21: BIOM metadata	- 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*"	 **Observation Metadata Tabular File** 150: gg_13_5_taxonomy  **Comma-separated list of the metadat** taxonomy

Step	Actions	Screenshot
	• Type “**OTUID**” and “**taxonomy**” separated by comma in the textbox under “*Comma-separated list of the observation metadata field names*”	 Comma-separated list of the observatio OTUID,taxonomy

From Step 22 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 22: DESeq2	• Select metadata file (e.g., metadata_NonOverlappingPE.txt) from the dropdown menu • Select “**c3: Tissue**” from the “*factor for differential analysis testing*”	 Add sample metadata from metadata 148: metadata_NonOverlappingPE.txt  factor for differential analysis testing c3: Tissue
Step 23: Phyloseq Richness	• Select metadata file (e.g., metadata_NonOverlappingPE.txt) from “*Metadata file*” the dropdown menu • Select “**c5: Replicate_Group**” from the “*Column used for X-axis*” • Select “**c4: Protein**” from the “*Column used as legend*” dropdown menu	 Add sample metadata from metadata 148: metadata_NonOverlappingPE.txt  Column used for X-axis c5: Replicate_Group  Column used as legend c4: Protein
Step 24: Phyloseq Abundance plot	• Select metadata file (e.g., metadata_NonOverlappingPE.txt) from “*Metadata file*” the dropdown menu • Select “**c5: Replicate_Group**” from the “*Column used for X-axis*”	 Add sample metadata from metadata 148: metadata_NonOverlappingPE.txt  Column used for X-axis c5: Replicate_Group

Step	Actions	Screenshots
	• Select "**c3: Tissue**" from the "*Column used as legend*" dropdown menu • Select "**c4: Protein**" from the "*Column used as factor 1*" dropdown menu	 Column used as legend c3: Tissue  Column used as legend c4: Protein
Step 25: Phyloseq Abundance Kingdom	• Select metadata file (e.g., metadata_NonOverlappingPE.txt) from "Metadata file" the dropdown menu • Select "**c4: Protein**" from the "*Column used for X-axis*" dropdown menu • Select "**c3: Tissue**" from the "*Column used as legend*" dropdown menu • Click on "**Phylum**" in the "select a taxonomy rank"	 Add sample metadata from metadata m 148: metadata_NonOverlappingPE.txt  Column used for X-axis c4: Protein  Column used as legend c3: Tissue  select a taxonomy rank ○ Kingdom ● Phylum ○ Class ○ Order ○ Family ○ Genus ○ Species
Step 26: Symmetric Plot	• Select metadata file (e.g., metadata_NonOverlappingPE.txt) from "*Metadata file*" the dropdown menu • Select green genes annotation file "**gg_13_5_taxonomy**" from the "*OTU taxonomy file*" dropdown menu	 Add sample metadata from metadata m 148: metadata_NonOverlappingPE.txt  OTU Taxonomy file 150: gg_13_5_taxonomy

Step	Actions	Screenshots
	- Select “**Phylum**” under “*Select a taxonomy rank*” - Check the checkbox under “**is the data normalised?**” - Select “**c3: Tissue**” from the “*Variable to compare” dropdown menu*” - Type “**Lung,Feces**” in the textbox under “*Fill in two comparable group separated by comma*”	 **Select a taxonomy rank** Phylum  **is the data normalised?** ☑  **Variable to compare** c3: Tissue  **Fill in two comparable group separated** Lung,Feces
Step 27: 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_NonOverlappingPE.txt) from the “*Metadata file*” dropdown menu - Select “**c5: Replicate_Group**” from the “*Select a group for correlation calculation”* dropdown menu - Select “**c4:Protein**” from the “*Column used as legend*” dropdown menu	 **Observation Metadata Tabular File** 150: gg_13_5_taxonomy  **is the data normalised?** ☑  **Add sample metadata from metadata m** 148: metadata_NonOverlappingPE.txt  **Select a group for correlation calculation** c5: Replicate_Group  **Column used as legend** c4: Protein

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 eg. after quality checking, after filtering steps, after plotting.

Step	Brief Description	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 o Sample13 ▪ Forward ▪ Reverse o Sample14 ▪ Forward ▪ Reverse o Etc… • reheader.PE.log o Sample13 ▪ Forward ▪ Reverse o Sample14 ▪ Forward ▪ Reverse o Etc… In this tutorial there were 24 x 2 = 48 total input files, so there will be 48 x 2 = 96 output files.

Step	**Brief Description**	**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) o Sample13 ▪ Forward ▪ Reverse o Sample14 ▪ Forward ▪ Reverse o Etc… • FASTQC on collection RawData (text) o Sample13 ▪ Forward ▪ Reverse o Sample14 ▪ Forward ▪ Reverse Etc…
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*	The collections have the following hierarchical structure: • *Trimmomatic across collection XXXX* (paired) o Sample13 ▪ Forward ▪ Reverse o Sample14 ▪ Forward ▪ Reverse

Step	Brief Description	Output files
	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.	○ Etc… • *Trimmomatic across collection XXXX* (unpaired) ○ Sample13 ▪ Forward ▪ Reverse ○ Sample14 ▪ Forward ▪ Reverse Etc…
Step 4: Concatenate multiple datasets	The output is a FASTQ file after merging forward reads from all input samples.	"*Concatenate multiple forward reads file merged*"
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: 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 data XXXX (mapped reads in BAM format)*
Step 7: FilterSamReads	Reads that mapped to the host genome in Step 6 are removed, generated one BAM file with the naming structure shown on the right.	*"FilterSamReads on data XXX: filtered BAM"*

Step	Brief Description	Output files
Step 8: 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 9: FASTQ to FASTA	The FASTQ file (step 9) is converted to a FASTA file with the naming structure shown on the right.	*"FASTQ to FASTA on data XXX"*
Step 10: 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 11: Vsearch dereplication	This steps remove all duplicate sequences and generates one FASTA output listing the unique sequences.	*"Vsearch dereplication on data XXXX"*
Step 12: 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"*

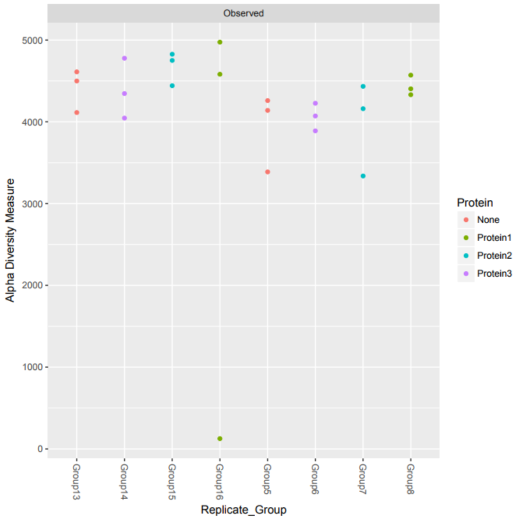
Step	Brief Description	Output files
Step 13: Vsearch clustering	This tool clusters the sequences into groups and generates a FASTA output that contains a list of consensus sequences with the naming structure shown on the right	"*Vsearch clustering on data XXXX: Consensus Sequences*"
Step 14: Vsearch search	Similar to step 10: Vsearch search This time we are searching the remaining reads from Step 10, output 3, against a new database created from Step 13. Only two outputs are generated in this step following the naming structure shown on the right.	1) *"VSearch search on data XXXX and XXXX: UCLUST like output"* 2) *"VSearch search on data XXXX and XXXX: Matching query sequences*"
Step 15: 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 16: 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 17: 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"*

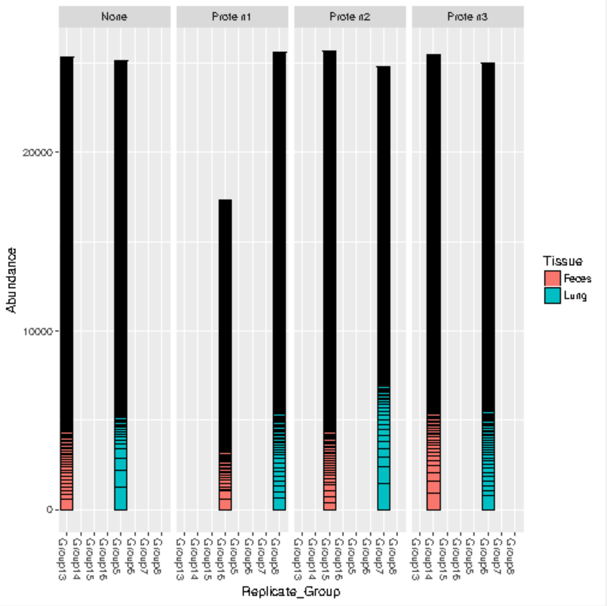
Step	**Brief Description**	**Output files**
Step 18: Concatenate datasets	This step merges the two outputs from step 12 with step 19 and generates a combined tabular file with the naming structure shown on the right. Reminder: Step 12 – is performing a search against Greengenes (known sequences) Step 19 – is the output after performing a search against novel sequences	*"Concatenated datasets on data XXXX and XXXX"*
Step 19: 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*"

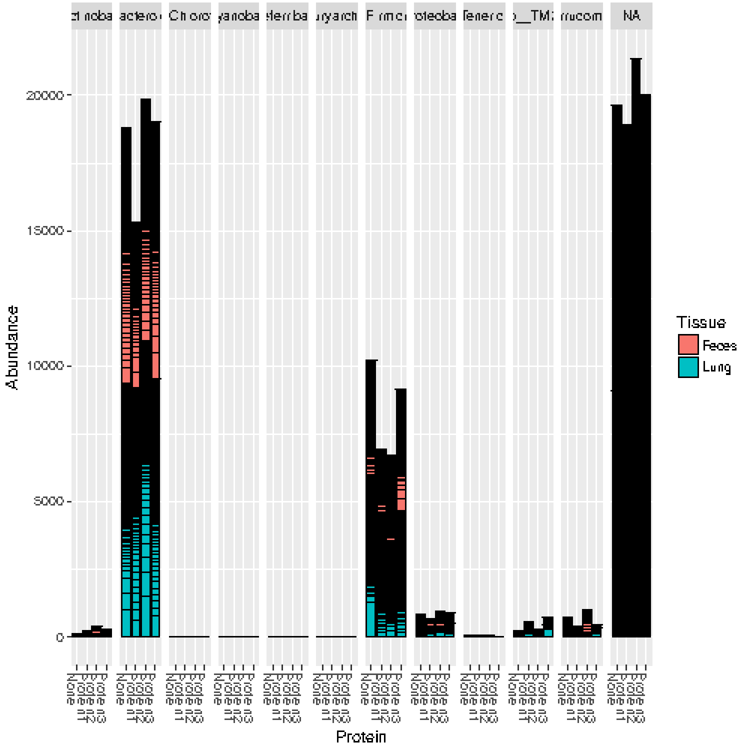
Note: Repeating differential analysis without post-processing dataset

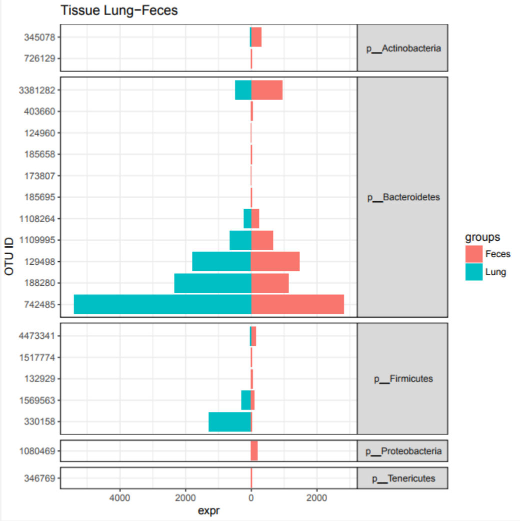
Now the dataset is transformed into an OTU Table in step 19, you can use this output in the fourth workflow (**04_16S_biodiversity_BIOM**) directly to perform other comparisons, without re-running all the data processing steps.

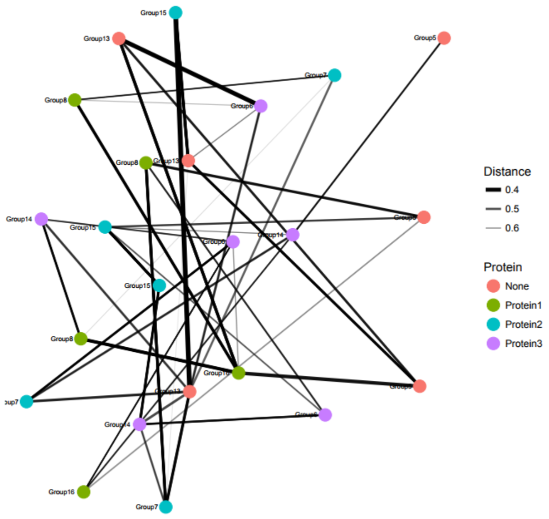
Step	**Brief Description**	**Output files**
Step 20: convert BIOM	This step converts the output from step 21 into a BIOM file following the naming structure shown on the right.	*"Convert BIOM on data XXXX and XXXX"*

Step	Brief Description	Output files
Step 21: 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 22: DESeq2	This step takes the annotated BIOM file from step 23 and generates two output files: 1) Normalised count table and A table listing the significant OTU results between the groups selected for comparison (in this tutorial "*Tissue"* was selected)	*1) "DESeq2 Normalised Table.txt"* 2) *"DESeq2 DE.txt"*
Step 23: 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	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 24: Phyloseq Abundance plot	This is an abundance plot of all samples grouped by the protein type. Within each group, the bars are coloured by the different tissue type (Feces vs Lung). The horizontal (x) axis are the groups 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 25: Phyloseq Abundance taxonomy plot	This is an abundance plot of all samples grouped by the taxonomy "phylum" as selected in Step 27 of the workflow. The bars are coloured by the tissue type (Feces vs Lung). The horizontal (x) axis are the protein type 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 Taxonomy.html"*

Step	Brief Description	Output files
		
Step 26: Symmetric plot	This symmetric plot shows the normalised counts abundance between the two tissue types (Feces vs Lung). The results shown is only for the dataset under taxonomy "phylum" as selected in step 28 of the workflow.	*"Symmetric Plot SymmetricPlot.html"*

Step	**Brief Description**	**Output files**
		
Step 27: Phyloseq Network plot	This plot shows a correlation network of samples based on the microbiome profiles using the normalised dataset.	*"Phyloseq Network Plot.html"*

Step	Brief Description	Output files
		

6 Version History

Version	Date	Modified by	Description
1.0	2017-09-20	QFAB (Mike, Xin-Yi)	• Initial version
1.1	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