16S rRNA microbial biodiversity analysis using GVL-Galaxy

Part 1: workflow for detecting paired-end overlaps

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, 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:

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_nonoverlapPE	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 1 Overlap detection.

WARNING: Filename formats

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; *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 this tutorial, the **16S_overlap_detection** workflow is used to detect the paired-end overlapping status for each library.

Below is the metadata for this dataset (

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*". We have added extra dummy metadata that will come in use for the second workflow (**16S_biodiversity_for_overlapPE**).

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 (cheese)	Replicate_Group
F3D0	0	Early	None	Group1
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

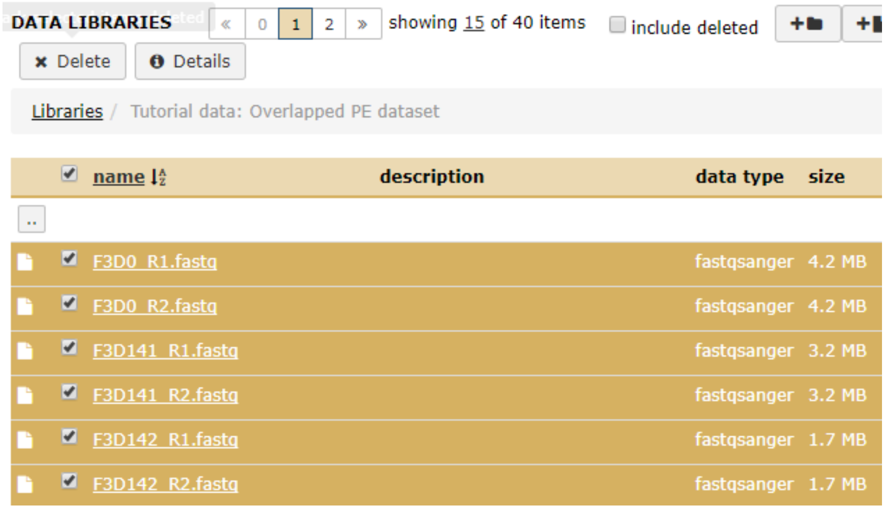
[1] The original dataset can be downloaded from https://zenodo.org/record/165147#.WYvbjfmqpBc

4 Data preparation

Before we start using the workflow we need to prepare the dataset in a format that is required by the workflow. This *16_overlap_detection* workflow in Galaxy is designed to take in a *List of Dataset Pairs* as input. The following steps show you how to create this list.

4.1 Import example dataset

1. Click on **Shared Data** from the top menu
2. Select **Data Libraries**
3. Click on **Tutorial data: Overlapped PE dataset** link
4. Check the box next to **name,** which will select *all* the FASTQ files. **Note:** make sure you select all the FASTQ files.



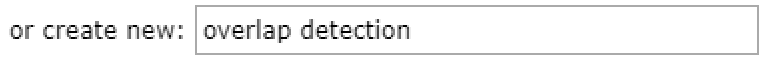
DATA LIBRARIES « 0 1 2 » showing 15 of 40 items □ include deleted

✖ Delete ❶ Details

Libraries / Tutorial data: Overlapped PE dataset

		name	description	data type	size
	☑	F3D0_R1.fastq		fastqsanger	4.2 MB
	☑	F3D0_R2.fastq		fastqsanger	4.2 MB
	☑	F3D141_R1.fastq		fastqsanger	3.2 MB
	☑	F3D141_R2.fastq		fastqsanger	3.2 MB
	☑	F3D142_R1.fastq		fastqsanger	1.7 MB
	☑	F3D142_R2.fastq		fastqsanger	1.7 MB

5. Click to History
6. Type a history name (e.g., Overlap detection) in the textbox



or create new: overlap detection

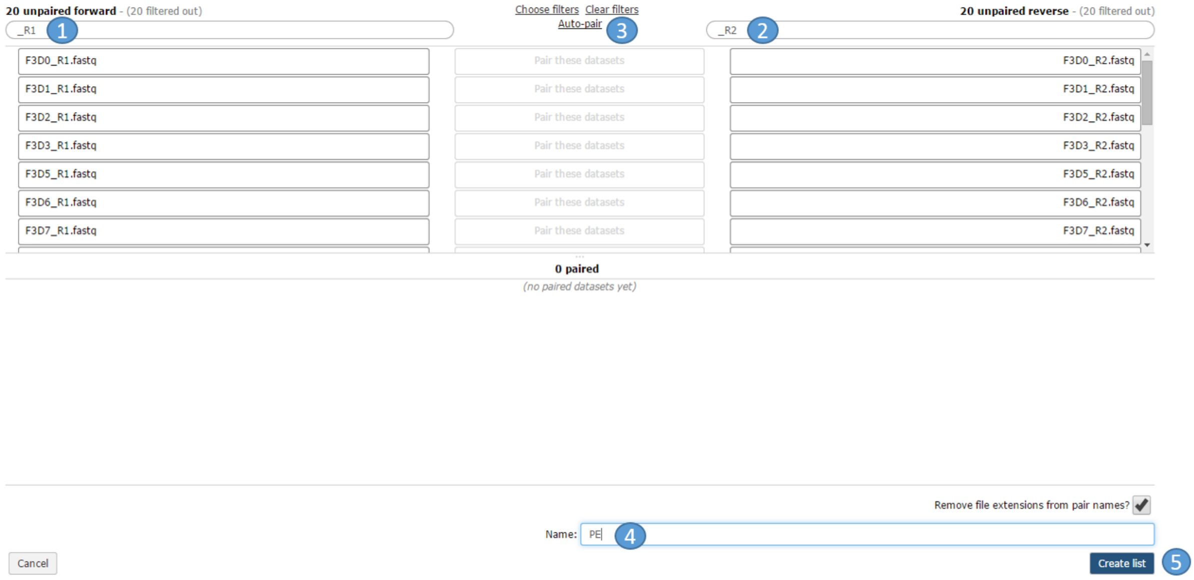
7. Click on **Import**
8. Click on **Analyze Data** from the top menu

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 ("overlap detection") 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 1
   - 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)

_R1

Choose filters Clear filters

Auto-pair

20 unpaired reverse - (20 filtered out)

_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? ☑

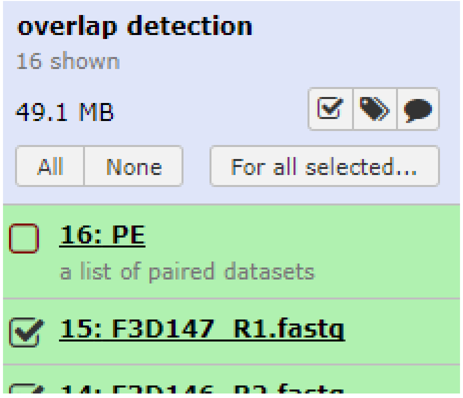
Name: PE

Cancel

Create list

Figure 1 Build a list of Dataset Pairs

You should now see a new dataset appear in the history panel.



overlap detection

16 shown

49.1 MB

All None For all selected...

□ 16: PE

a list of paired datasets

☑ 15: F3D147_R1.fastq

We are now ready for the workflow analysis.

5 Overlap detection workflow

This workflow randomly selects 1000 sequences from each paired-end fastq file and calculates how many reads overlap by at least 10bp. The average percentage is returned by the tool and a recommendation of which workflow is suitable. Table 4 describes the six steps used in the workflow, which is visually represented in Figure 2.

Table 4 Components in overlapping statistic paired-end workflow.

Step	Description
1. Input dataset collection	Dataset collection type
2. Seqtk subsample	Subset 1000 sequences from FASTQ files
3. FASTQC	Quality checking before adapter removal
4. Trimmomatic	Remove adapters
5. FASTQC	Quality checking after adapter removal
6. PEAR	Merging paired-end data
7. PEAR statistic	Generate a statistic log for all merged pairs

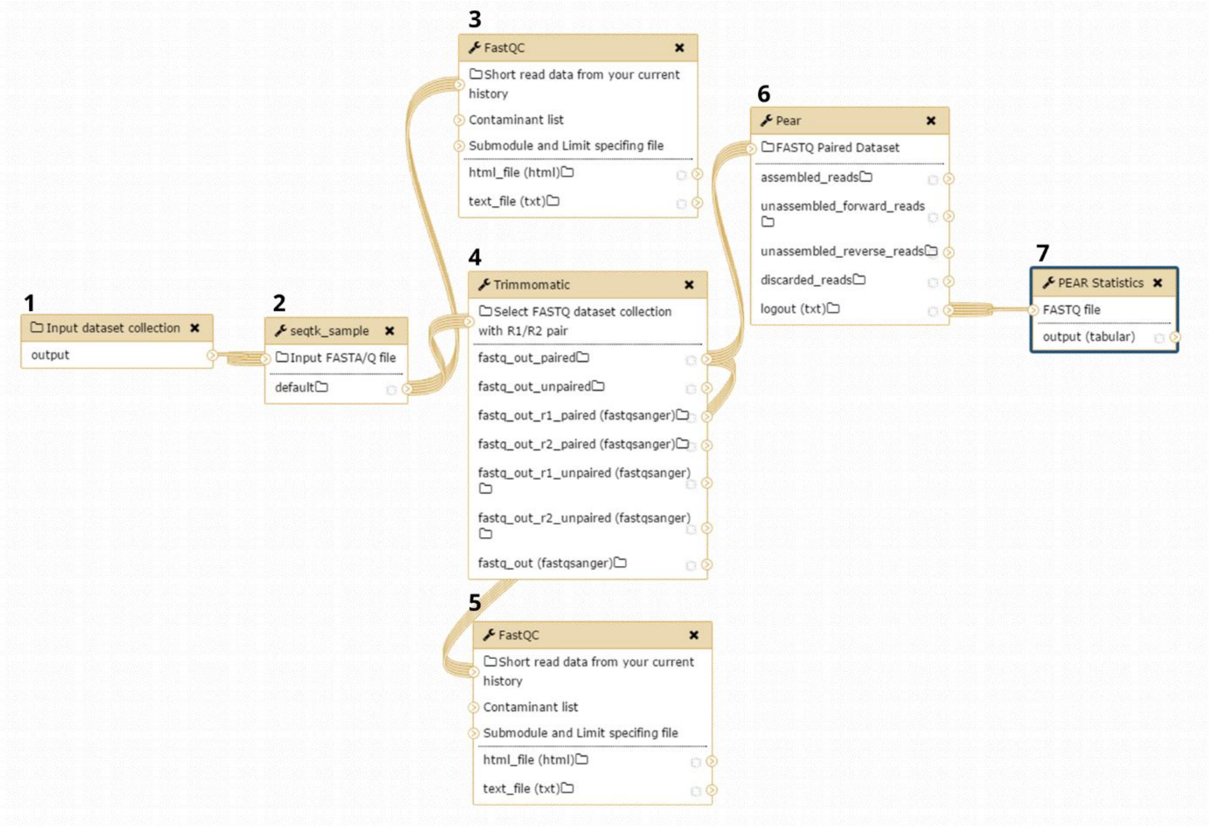
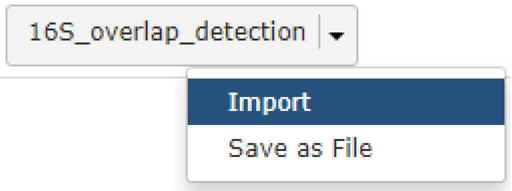

1
Input dataset collection
output
2
seqtk_sample
Input FASTA/Q file
default
3
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
html_file (html)
text_file (txt)
4
Trimmomatic
Select FASTQ dataset collection with R1/R2 pair
fastq_out_paired
fastq_out_unpaired
fastq_out_r1_paired (fastqsanger)
fastq_out_r2_paired (fastqsanger)
fastq_out_r1_unpaired (fastqsanger)
fastq_out_r2_unpaired (fastqsanger)
fastq_out (fastqsanger)
5
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
html_file (html)
text_file (txt)
6
Pear
FASTQ Paired Dataset
assembled_reads
unassembled_forward_reads
unassembled_reverse_reads
discarded_reads
logout (txt)
7
PEAR Statistics
FASTQ file
output (tabular)

Figure 2 The statistic workflow of overlapping paired-end reads.

5.1 Import workflow

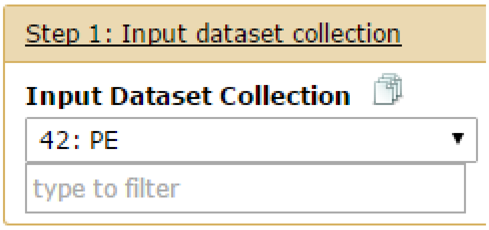
If you are a new user to Galaxy, follow this section to import the shared Galaxy workflow into your workspace. You can skip this section if you have already imported the workflow before.

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



5.2 Run workflow

1. Click on **Workflow** from the top menu
2. Click on **imported: 16S_overlap_detection**. If you cannot find this workflow, rerun Section 5.1 Import workflow.
3. Click on **Run** on the dropdown menu
4. Select the name of the **Build List of Dataset Pairs** in the previous section (e.g., PE)

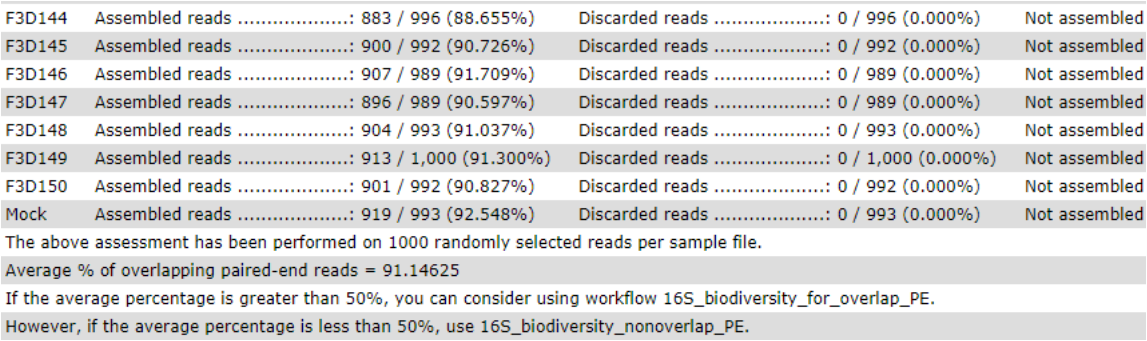

Step 1: Input dataset collection

Input Dataset Collection

42: PE

5. Click on **Run workflow** at the bottom of the page
6. The final output of this workflow is shown in Figure 3.

The last 4 lines provide a summary of the output:



```
F3D144   Assembled reads ..................: 883 / 996 (88.655%)     Discarded reads ..................: 0 / 996 (0.000%)     Not assembled
F3D145   Assembled reads ..................: 900 / 992 (90.726%)     Discarded reads ..................: 0 / 992 (0.000%)     Not assembled
F3D146   Assembled reads ..................: 907 / 989 (91.709%)     Discarded reads ..................: 0 / 989 (0.000%)     Not assembled
F3D147   Assembled reads ..................: 896 / 989 (90.597%)     Discarded reads ..................: 0 / 989 (0.000%)     Not assembled
F3D148   Assembled reads ..................: 904 / 993 (91.037%)     Discarded reads ..................: 0 / 993 (0.000%)     Not assembled
F3D149   Assembled reads ..................: 913 / 1,000 (91.300%)   Discarded reads ..................: 0 / 1,000 (0.000%)   Not assembled
F3D150   Assembled reads ..................: 901 / 992 (90.827%)     Discarded reads ..................: 0 / 992 (0.000%)     Not assembled
Mock     Assembled reads ..................: 919 / 993 (92.548%)     Discarded reads ..................: 0 / 993 (0.000%)     Not assembled
The above assessment has been performed on 1000 randomly selected reads per sample file.
Average % of overlapping paired-end reads = 91.14625
If the average percentage is greater than 50%, you can consider using workflow 16S_biodiversity_for_overlap_PE.
However, if the average percentage is less than 50%, use 16S_biodiversity_nonoverlap_PE.
```

Figure 3 The statistic of overlapping paired-end data

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