Illumina (post-MR DNA) processing pipeline : QIIME

Kylie Langlois[1]

[1]State University of New York at Stony Brook

dx.doi.org/10.17504/protocols.io.hk2b4ye

Collier Lab

Kylie Langlois

ABSTRACT

This pipeline is for datasets generated from Illumina (2x300) sequencing by MR DNA. MR DNA runs sequences through a proprietary pipeline for quality control of sequences and OTU clustering. This pipeline begins with the OTU abundnace table and utilizes the representative sequence set generated by the MR DNA pipeline. **Through trial and error, I have discovered the easiest and best practice for my lab is to use a combination of QIIME (full installation) and mothur. This protocol is the QIIME portion.** This protocol does not include data visualization.

As of 4/5/17, the following versions were used: QIIME v.1.9.1-amd64.vdi and mothur v.1.39.5

PROTOCOL STATUS

Working

Working for QIIME1. Our lab currently uses another protocol for QIIME2.

Create a mapping file

1 In QIIME, use the alpha rarefaction script to rarefy the OTU abundance table according to specific alpha diversity metrics. Before that, you need to create a mapping file, a parameters file, and convert the OTU table into the biom format.

Create a mapping file and validate it.

- Create a mapping file following these requirements http://qiime.org/documentation/file_formats.html
- Validate the mapping file with this script http://qiime.org/scripts/validate_mapping_file.htm

validate_mapping_file.py -m CCWT_16S_Jan_map.txt -o validate_mapping_file_output -p -b

COMMAND

validate_mapping_file.py -m CCWT_16S_Apr_map.txt -o CCWT_16S_Apr_validate_mp -p -b

This error will appear:

Convert the OTU abundance table to biom format

2 Convert the OTU table to biom format

- QIIME uses biom files, more info can be found http://biom-format.org/documentation/
- Excel wookbooks (.xl or .xls), tab-delimited (.txt or .tsv), comma-delimited (.csv), and biom (.bioim) files are not interchangeable. To work in QIIME, you need to convert your file to the biom format and you'll need to convert it out of the bioma format after your analyses are concluded. Those scripts can be found here http://biom-format.org/documentation/biom_conversion.html

COMMAND

biom convert -i CCWT_16S_Apr_OTU.txt -o CCWT_16S_Apr_OTU.biom --to-hdf5 --table-type="OTU table"

The biom website list the above arguments in different orders, but the order above is the only way that works

Get initial sequencing effort

3 Extract information from the biom table to see how many sequences and OTUs were in your original dataset and how many sequences were initially in each sample.

COMMAND

```
biom summarize-table -i OTU_table.biom -o OTU_table_summary.txt
```

Create a parameter file

4 In order to use the alpha_rarefaction.py script, you need a parameters file to specify which alpha diversity metric you wish to you. Once you create a parameters file, you can reuse it in various scripts by changing the parameters line.
http://qiime.org/documentation/qiime_parameters_files.html

- For alpha diversity metrics, the first line might look like the table below. In the first line without hashtags (#), the script that is running, a colon, the script that will be called, a space, and the metric you wish to pass. I've found that trying to run multiple metrics at once simply returns an error message from QIIME.

alpha_rarefaction:alpha_diveresity shannon	

- For beta diversity metrics, the first line might look like the table below.

beta_diversity_through_plots:beta_diversity unifrac	

Remove singletons

5 Remove OTUs that are only 1 sequence

COMMAND

```
filter_otus_from_otu_table.py -i otu_table.biom -o otu_table_no1.biom -n 2
```

Determination of the rarefied sequencing depth and alpha diversity

6 The alpha rarefaction script rarefies the OTU abundance table to multiple sequencing depths and plots that against the alpha diversity at each sequencing depth for each sample. The result directory contains interactive and static line graphs (and corresponding data tables) showing the progression of alpha diversity at different sequencing depths. From here, I typically choose the highest sequencing abundance that still retains all samples and use that in my downstream analysis.

I typically use "observed OTUs" as my alpha diversity metric because the calculated metrics (such as Shannon and Simpson) have gone to saturation very quickly in every dataset I've had.

COMMAND

```
alpha_rarefaction.py -i CCWT_16S_Apr_OTU.biom -p CCWT_16S_Apr_parameter.txt -o CCWT_16S_Apr_arare_osd/ -m CCWT_16S_Apr_map_corrected.txt
```

All other defaults were used

Rarefaction of the OTU abundance table

7 Take the rarefied sequencing depth from Step 4 and use that for the single rarefaction script. Step 4 will only give you graphs, not the OTU abundance table at different depths.

COMMAND

```
single_rarefaction.py -i CCWT_16S_Apr_OTU.biom -o CCWT_16S_Apr_even.biom -d 46644
```

All other defaults were used

Align representative set of sequences

8 In order to run the beta diversity script that creates graphs as the final output, you need to use the representative set of sequences for your entire dataset.

After you locate this file, the QIIME SOP is to align the sequences to a reference alignment, filter them, make a phylogeny, then calculate the beta diversity. The default reference alignment is the Greengenes reference alignment, but you can use other alignments such as RDP and SILVA (but SILVA requires additional curating). You can download the newest Greengenes references alignment (13_8) from here http://qiime.org/home_static/dataFiles.html or here https://www.mothur.org/wiki/Greengenes-formatted_databases. The Greengenes

website only has version 13_5 available for download, but it can be found here http://greengenes.secondgenome.com/downloads/database/13_5.

For 18S sequences, you need to use a SILVA alignment, which can be found here https://www.arb-silva.de/download/archive/qiime/.

PyNast is the default aligner and is a python implementation of the NAST alignment algorithm. Based on phylogenetic analysis, PyNast may not be the best tool to use for Illumina sequencing, which can have sequences in the reverse order or very 'weird' sequences (most likely environmental DNA fragments) that PyNast and QIIME do not detect and filter out.

COMMAND

align_seqs.py -i CCWT_16S_rep_set.fa -o CCWT_16S_pynast_aligned/

All other defaults were used. The default aligner is PyNast.

COMMAND

align_seqs.py -i CCWT_18S_rep_set.fa -t Silva_119_rep_set97_aligned_18S_only.fa -o CCWT_18S_rep_set_aligned/

FOR 18S SEQUENCES

Filter alignment from Step 6

9 A required step if you use PyNast to align the sequences.

For 18S sequences, you need to suppress the lane mask or the alignment will not be filtered correctly.

COMMAND

filter_alignment.py -i CCWT_16S_rep_set_aligned.fasta -o CCWT_16S_filtered_alignment/

All other defaults were used

COMMAND

filter_alignment.py -i CCWT_18S_rep_set_aligned.fasta -s -o CCWT_18S_rep_set_aligned_filtered/

FOR 18S SEQUENCES: use parameter

Create phylogenetic tree from alignment made in Step 8

10 The default tree building tool is FastTree, but clearcut can be implemented by use the command "-t clearcut" at the terminal end of the command below. **If sequences are flipped or "weird" (see step 7), the resulting phylogenetic tree may be incorrect.**

Note: the default mothur method to building phylogenetic trees is clearcut.

COMMAND

make_phylogeny.py -i CCWT_16S_rep_set_aligned_pfiltered.fasta -o CCWT_16S_phylo.tre

All other defaults were used

Determine beta diversity

11 Using the files generated in Steps 8, 9, 1, and the mapping file from Step 1, you can calculate beta diversity. Even with a parameter file, both weighted and unweighted UniFrac will be used to calculate beta diversity. Weighted UniFrac is a quantitative measurement while unweighted UniFrac is a qualitative measurement. Both metrics calculate dissimiliarity between samples based on phylogenetic information.

This script results in a distance matrix and a coordinates matrix (for PCoA) of your sample points, as well as premade PCoA plots. From here, you can take the coordinates matrix and plot any PCs against each other in R.

COMMAND

beta_diversity_through_plots.py -i CCWT_16S_Apr_OTU_even.biom -o CCWT_16S_Apr_beta_div_plots/ -t CCWT_16S_phylo.tre -m CCWT_16S_Apr_map_corrected.txt

All other defaults were used

Convert rarefied OTU abundance table back to text file

12 To view and use the rarefied OTU abundance table, you need to convert it out of the biom format

COMMAND

```
biom convert -i CCWT_16S_Apr_OTU_even.biom -o CCWT_16S_Apr_OTU_even.txt --to-tsv
```

The biom website list the above arguments in different orders, but the order above is the only way that works