Instructions for the processing of Grubisha Lab Illumina MiSeq sequences

Samantha Nellis[1], Josh Moyer[2], and Lisa C. Grubisha[3]

January 17, 2017

Updated February 24, 2017

Sequences are received from the University of Wisconsin Biotechnology Center as **demultiplexed, zipped fastq files** (.fastq.gz). These files need to be quality filtered, combined into one file and assigned taxonomy (Figure 1).

Before processing or analyzing the data, go through the tutorials (requires Python Notebook) on the QIIME website to familiarize yourself with writing QIIME codes. The tutorial from the Werner lab was also helpful (www.wernerlab.org).

The following analysis follows the pipeline by Nguyen et al (2015) and is intended for soil community analyses but could be adapted for other studies.

1.Download and install Oracle VM Virtual Box

2.Open Oracle VM Virtual Box , initiate QIIME program

3.Import the *.fastq.gz* files and save in a new directory (the other soil samples are saved on the VirtualBox desktop under Illumina Soil). Make sure to save the raw data files from the lab in two places (one folder for processing and one with the original files). The files can either be downloaded directly from the website using Chrome in the VirtualBox or can be put into the "Shared_Folder". The Shared_Folder is the only folder that can be accessed in the Virtual machine as well as the lab PC. It is labeled "Shared_Folder" in the VirtualBox and is under the C: drive -> "QIIME virtual box" on the PC. If necessary, look on the QIIME website for more information on creating or editing shared folders. The QIIME website contains detailed explanations of the codes used. Descriptions can also be brought up in the terminal by using the –h command (ex. pick_otus.py –h). Full file names should be used for all of inputs unless the user has a good understanding of how to move between different directories. (For example: /home/qiime/Desktop/Illumina_soil/ITS/processing/235_R1_trimmed.good.fasta)

[1]Environmental Science & Policy Graduate Program, Department of Natural & Applied Sciences, University of Wisconsin – Green Bay, Green Bay, WI 54311

[2]Biology, Cell & Molecular Biology. Department of Natural & Applied Sciences, University of Wisconsin – Green Bay, Green Bay, WI 54311

[3]Department of Natural & Applied Sciences, University of Wisconsin – Green Bay, Green Bay, WI 54311

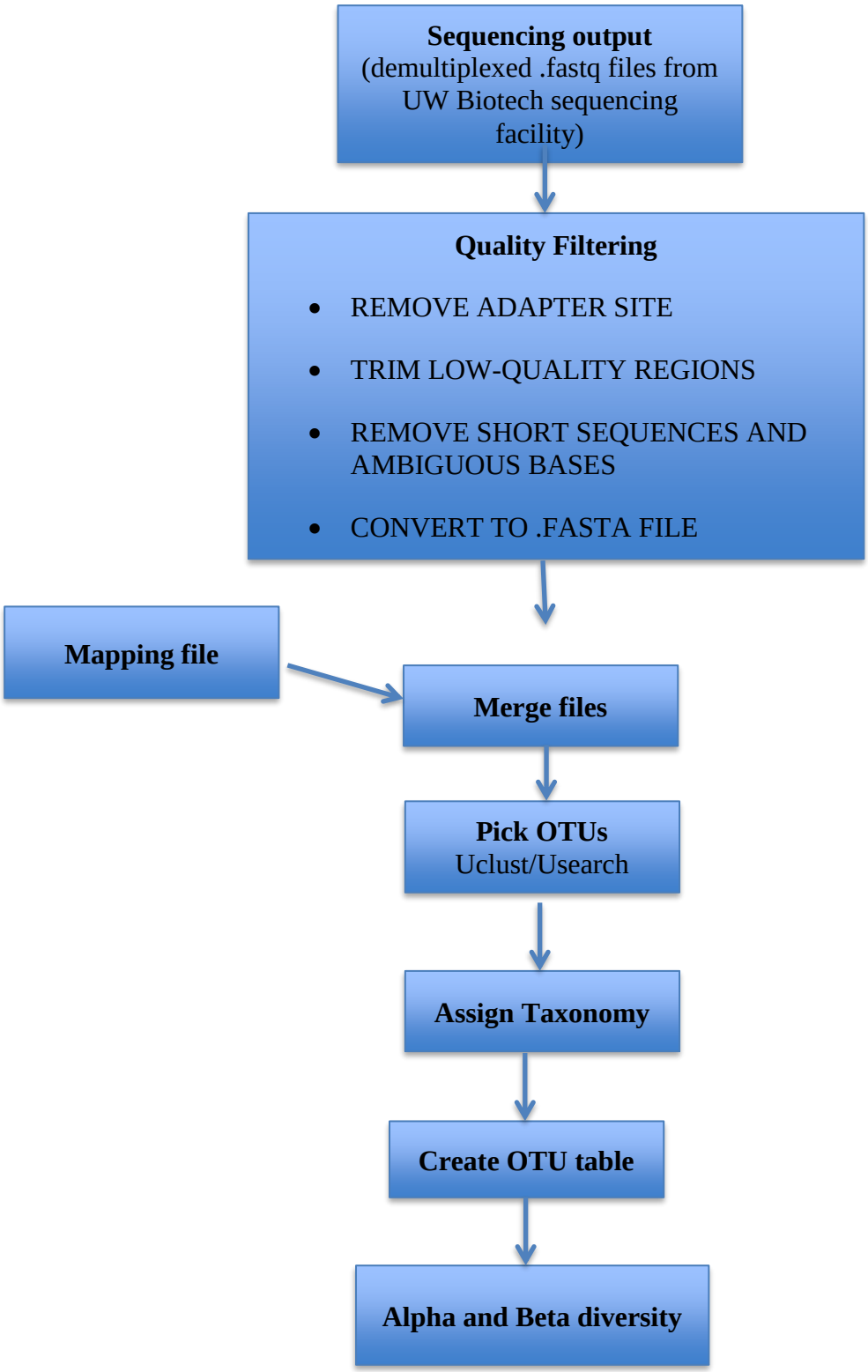

Figure 1. Flowchart showing the steps taken in processing Illumina sequence data from the University of Wisconsin Biotechnology Center (Madison, WI) using the Qiime pipeline. Adapted from Kuczynski et al. (2011).

Analyzing the fungal (ITS) files:

Note: Steps 1-4 should be performed on *each* .fastq.gz file from the sequencing lab. The following shows an example of one file. The same code can be used but the input and output filenames must be changed for each. Process all files with the code in Step 1 before moving onto Step 2. Only the codes and selected outputs are shown.

#1. Remove adapter site.
Only use forward read files (files that end in _R1_001). Per Nguyen 2015 the forward reads are more accurate but you may want to try pairwise at a later date. **Use cutadapt**. Example:

qiime@qiime-190-virtual-box:~$ cutadapt -a GATCTCTTGGNTCTNGCATCGATGAAGAACG -q 20 -e 0.2 /home/qiime/Desktop/Illumina_soil/ITS/UWGB_ITS_235_GTCATTCACGAG_L001_R1_001.fastq.gz -o /home/qiime/Desktop/Illumina_soil/ITS/processing/235_R1.fastq

#2. Trim low-quality regions.
Example:

qiime@qiime-190-virtual-box:~$ java -jar /usr/local/bin/Trimmomatic/trimmomatic-0.36.jar SE -phred33 /home/qiime/Desktop/Illumina_soil/ITS/processing/235_R1.fastq /home/qiime/Desktop/Illumina_soil/ITS/processing/235_R1_trimmed.fastq LEADING:20 TRAILING:20 MINLEN:125

#3. Convert fastq to fasta files.
Example:

qiime@qiime-190-virtual-box:~$ mothur

mothur > fastq.info(fastq=/home/qiime/Desktop/Illumina_soil/ITS/processing/235_R1_trimmed.fastq)

Output File Names:
/home/qiime/Desktop/Illumina_soil/ITS/processing/235_R1_trimmed.fasta
/home/qiime/Desktop/Illumina_soil/ITS/processing/235_R1_trimmed.qual

#4. Filter out short seqs and ones with ambiguous bases (forward and/or reverse reads).
Example:

mothur >
screen.seqs(fasta=/home/qiime/Desktop/Illumina_soil/ITS/processing/235_R1_trimmed.fasta, maxambig=0, minlength=125, maxhomop=9, processors=2)

Output File Names:
/home/qiime/Desktop/Illumina_soil/ITS/processing/235_R1_trimmed.good.fasta
/home/qiime/Desktop/Illumina_soil/ITS/processing/235_R1_trimmed.bad.accnos

mothur > quit

#5. Move *all* trimmed.good.fasta files to one directory.
Example:

/home/qiime/Desktop/Illumina_soil/ITS/processing/Processed/235_R1_trimmed.good.fasta

#6. Create a mapping file in Excel.
Example:

#SampleID	BarcodeSequence	LinkerPrimerSequence	InputFileName	Description
Soil05	ATTCTCTCACGT	GGCTTGGTCATTTAGAGGAAGTAA	201_R1_trimmed.good.fasta	W1-8/03/15
Soil06	CGACTCTAAACG	GGCTTGGTCATTTAGAGGAAGTAA	202_R1_trimmed.good.fasta	W2-8/03/15
Soil07	GTCTTCAGCAAG	GGCTTGGTCATTTAGAGGAAGTAA	203_R1_trimmed.good.fasta	W3-8/03/15
Soil08	CGGATAACCTCC	GGCTTGGTCATTTAGAGGAAGTAA	204_R1_trimmed.good.fasta	W4-8/11/15

Soil09	AGGGTGACTTTA	GGCTTGGTCATTTAGAGGAAGTAA	205_R1_trimmed.good.fasta	W5-8/11/15
Soil11	GGATAGCCAAGG	GGCTTGGTCATTTAGAGGAAGTAA	219_R1_trimmed.good.fasta	W7-8/13/15
Soil12	TGGTTGGTTACG	GGCTTGGTCATTTAGAGGAAGTAA	220_R1_trimmed.good.fasta	W8-8/13/15

First 4 headers need to be in the following order: "#SampleID | BarcodeSequence | LinkerPrimerSequence | InputFileName" then you can add additional headers to help with identifying samples later on (e.g., "Description" header creates a column where we can put sample IDs, metadata etc.). The "Description" header must be the last header included. Once the spread sheet is made, save it as a ".txt" file to the virtual box share drive (e.g. mapping file from above is "Soil_Map.txt") and then move the map file from the share drive to your working directory (ex. "/home/qiime/Desktop/Illumina_soil/ITS/processing/Processed/"). Check the mapping file using the code "validate_mapping_file.py". This code will make sure that the file is in the correct format and that no information is missing (it will not tell you, for example, if you put in the wrong linker primer sequence). See the QIIME website for additional tips on creating your mapping file.

#7. Add qiime labels and combine all sequences into 1 file.
Example:
qiime@qiime-190-virtual-box: $ add_qiime_labels.py -i ~/Desktop/Illumina_soil/ITS/processing/Processed/ -m Soil_Map.txt -c InputFileName -o ~/Desktop/Illumina_soil/ITS/processing/Processed/

Output File Names:

~/Desktop/Illumina_soil/ITS/processing/Processed/combined_seqs.fna ("Each new file gets the same name every time you run the command, think about renaming file (e.g. "combinedSoil_seqs.fna")."

#8a. OTU picking. The first step uses the UCLUST OTU picking method and the latest UNITE database. It should be noted that there are many OTU picking methods that will produce different results. The following method is known as

multi-step OTU picking and uses two OTU pickers chained together in an attempt to cluster the sequences more tightly.

qiime@qiime-190-virtual-box: $ parallel_pick_otus_uclust_ref.py -i ~/Desktop/Illumina_soil/ITS/processing/Processed/combinedSoil_seqs.fna -o ~/Desktop/Illumina_soil/ITS/processing/Processed/parallel_pick/ -r /usr/local/bin/developer/sh_refs_qiime_ver7_97_31.01.2016_dev.fasta

#8b. Pick rep set. A representative set of sequences (a file containing one sequence from each OTU, a rep set) is selected from the output of the first OTU picker. This rep set can be passed through another OTU picker.

qiime@qiime-190-virtual-box:~$ pick_rep_set.py -i /home/qiime/Desktop/Illumina_soil/ITS/processing/Processed/parrallel_pick/combinedSoil_seqs_otus.txt -f /home/qiime/Desktop/Illumina_soil/ITS/processing/Processed/combinedSoil_seqs.fna -o /home/qiime/Desktop/Illumina_soil/ITS/processing/Processed/parrallel_pick/ITS_rep_set.fna

#8c. Second OTU picking method using USEARCH and the latest UNITE database:

qiime@qiime-190-virtual-box:~$ pick_otus.py -i /home/qiime/Desktop/Illumina_soil/ITS/processing/Processed/parrallel_pick/ITS_rep_set.fna -m usearch --word_length 64 --db_filepath /usr/local/bin/developer/sh_refs_qiime_ver7_97_31.01.2016_dev.fasta --minsize 2 -o /home/qiime/Desktop/Illumina_soil/ITS/processing/Processed/parrallel_pick/pick_otus/

#8d. Merge OTU maps. The following command combines the outputs from both OTU pickers into one file.

qiime@qiime-190-virtual-box:~$ merge_otu_maps.py -i /home/qiime/Desktop/Illumina_soil/ITS/processing/Processed/parrallel_pick/combinedSoil_seqs_otus.txt,/home/qiime/Desktop/Illumina_soil/ITS/processing/Processed/parrallel_pick/pick_otus/ITS_rep_set_otus.txt -o /home/qiime/Desktop/Illumina_soil/ITS/processing/Processed/parrallel_pick/pick_otus/merged_otus.txt

#8e. Pick a final rep set. This rep set can be used for further analysis and processing.

qiime@qiime-190-virtual-box:~$ pick_rep_set.py –I /home/qiime/Desktop/Illumina_soil/ITS/processing/Processed/parrallel_pick/pick_otus/merged_otus.txt -f /home/qiime/Desktop/Illumina_soil/ITS/processing/Processed/combinedSoil_seqs.fna -o /home/qiime/Desktop/Illumina_soil/ITS/processing/Processed/parrallel_pick/pick_otus/ITSrep_set.fasta

#9a. Assign taxonomy:

qiime@qiime-190-virtual-box:~$ assign_taxonomy.py -i /home/qiime/Desktop/Illumina_soil/ITS/processing/Processed/parrallel_pick/pick_otus/ITSrep_set.fasta -r /usr/local/bin/developer/sh_refs_qiime_ver7_97_31.01.2016_dev.fasta -t /usr/local/bin/developer/sh_taxonomy_qiime_ver7_97_31.01.2016_dev.txt -o /home/qiime/Desktop/Illumina_soil/ITS/processing/Processed/parrallel_pick/pick_otus/assign_tax/

#9b. Make OTU table.

qiime@qiime-190-virtual-box:~$ make_otu_table.py -i /home/qiime/Desktop/Illumina_soil/ITS/processing/Processed/parrallel_pick/pick_otus/merged_otus.txt -t /home/qiime/Desktop/Illumina_soil/ITS/processing/Processed/parrallel_pick/pick_otus/assign_tax/ITSrep_set_tax_assignments.txt -o /home/qiime/Desktop/Illumina_soil/ITS/processing/Processed/parrallel_pick/pick_otus/assign_tax/ITS_otu_table.biom

#9c. Summarize biom table

qiime@qiime-190-virtual-box:~$ biom summarize-table -i /home/qiime/Desktop/Illumina_soil/ITS/processing/Processed/parrallel_pick/pick_otus/assign_tax/ITS_otu_table.biom

Num samples: 20
Num observations: 41
Total count: 56960
Table density (fraction of non-zero values): 0.466
Counts/sample summary:
 Min: 33.0
 Max: 7306.0
 Median: 2748.000
 Mean: 2848.000
 Std. dev.: 1899.233
 Sample Metadata Categories: None provided
 Observation Metadata Categories: taxonomy
Counts/sample detail:
 Soil20: 33.0
 Soil05: 397.0
 Soil06: 563.0
 Soil16: 967.0
 Soil15: 1316.0
 Soil10: 1555.0
 Soil13: 1687.0
 Soil08: 1871.0
 Soil11: 2041.0
 Soil22: 2690.0
 Soil17: 2806.0
 Soil21: 2830.0
 Soil18: 3517.0
 Soil19: 3598.0
 Soil12: 3962.0
 Soil04: 4151.0
 Soil09: 4182.0
 Soil07: 5550.0
 Soil14: 5938.0
 Soil03: 7306.0

#10. Alpha and Beta diversity analyses. There are several ways to do this and more codes may be needed depending on the research question. The command below, core_diversity_analyses.py, will perform a host of analyses on the data provided, including alpha and beta diversity analyses. Note that the "-e 33" comes from the "min count" in the previous output. This value will vary with each new

biom table provided. Also, be sure to use the mapping file associated with the current biom table/workflow.

qiime@qiime-190-virtual-box:~$ core_diversity_analyses.py -i /home/qiime/Desktop/Illumina_soil/ITS/processing/Processed/parrallel_pick/pick_otus/assign_tax/ITS_otu_table.biom -m /home/qiime/Desktop/Illumina_soil/ITS/processing/Processed/Soil_Map.txt -o /home/qiime/Desktop/Illumina_soil/ITS/processing/Processed/parrallel_pick/pick_otus/assign_tax/core_diversity/ -e 33 --nonphylogenetic_diversity

Analyzing the bacterial (16S) files:

Note: Steps 1-4 should be performed on *each* .fastq.gz file from the sequencing lab. The following shows an example from one file. The same code can be used but the input and output filenames must be changed for each. Process all files with the code in Step 1 before moving onto Step 2. Only the codes and selected outputs are shown.

#1. Remove adapter site. Only use forward read files (files that end in _R1_001). Per Nguyen 2015 the forward reads are more more accurate but you may want to try pairwise at a later date. Used cutadapt. Example:

qiime@qiime-190-virtual-box:~$ cutadapt -a ATTAGAWACCCBDGTAGTCCGGCTGACTGACT -q 20 -e 0.2 /home/qiime/Desktop/Illumina_soil/16s/UWGB_16S_146_AACGTATCGCCA_L001_R1_001.fastq.gz -o /home/qiime/Desktop/Illumina_soil/16s/processing/146_R1.fastq

#2. Trim low-quality regions.
Example:
qiime@qiime-190-virtual-box:~$ java -jar /usr/local/bin/Trimmomatic/trimmomatic-0.36.jar SE -phred33 /home/qiime/Desktop/Illumina_soil/16s/processing/146_R1.fastq /home/qiime/Desktop/Illumina_soil/16s/processing/146_R1_trimmed.fastq LEADING:20 TRAILING:20 MINLEN:100

#3. Convert fastq to fasta files.
Example:

qiime@qiime-190-virtual-box:~$ mothur

mothur >
fastq.info(fastq=/home/qiime/Desktop/Illumina_soil/16s/processing/146_R1_trimmed.fastq)

Output File Names:
/home/qiime/Desktop/Illumina_soil/16s/processing/146_R1_trimmed.fasta
/home/qiime/Desktop/Illumina_soil/16s/processing/146_R1_trimmed.qual

#4. Filter out short seqs and ones with ambiguous bases (forward and/or reverse reads).
Example:
mothur>
screen.seqs(fasta=/home/qiime/Desktop/Illumina_soil/16s/processing/146_R1_trimmed.fasta, maxambig=0, minlength=100, maxhomop=7, processors=2)

Output File Names:
/home/qiime/Desktop/Illumina_soil/16s/processing/146_R1_trimmed.good.fasta
/home/qiime/Desktop/Illumina_soil/16s/processing/146_R1_trimmed.bad.accnos

mothur > quit

#5. Put all trimmed.good.fasta into new folder (e.g. "Processed").

#6. Create mapping file. See ITS example and explanation. Map files can be viewed in the text editor in VirtualBox.

#7. Add QIIME labels and combine sequences into one file:

qiime@qiime-190-virtual-box:~$ add_qiime_labels.py -i /home/qiime/Desktop/Illumina_soil/16s/processing/processed -m /home/qiime/Desktop/Illumina_soil/16s/Soil_16s_map_corrected.txt -c InputFileName -o /home/qiime/Desktop/Illumina_soil/16s/processing/processed/

#8a. Pick OTUs: This step uses an open reference OTU picking method where sequences are matched to a database (GreenGenes) and any sequences that are not matched to the database are assigned using de novo OTU picking method. Alternatively, the steps from the ITS OTU picking can be used, just be sure to use either the Greengenes or RDP databases as the reference database (-r).

qiime@qiime-190-virtual-box: $ pick_open_reference_otu.py –i /home/qiime/Desktop/Illumina_soil/16s/processing/processed/combined_seqs.fna –o /home/qiime/Desktop/Illumina_soil/16s/processing/processed/open_ref_otu_jrm/ -r /usr/local/bin/97_otus.fasta –m uclust

#8b. Summarize biom table.

qiime@qiime-190-virtual-box: $ biom summarize-table –i /home/qiime/Desktop/Illumina_soil/16s/processing/processed/open_ref_otu_jrm/otu_table_mc2.biom

#9. Alpha and Beta diversity analyses.

qiime@qiime-190-virtual-box: $ core_diversity_analyses.py –i /home/qiime/Desktop/Illumina_soil/16s/processing/processed/open_ref_otu_jrm/otu_table_mc2_w_tax.biom –o /home/qiime/Desktop/Illumina_soil/16s/processing/processed/open_ref_otu_jrm/core_diversity_output –m /home/qiime/Desktop/Illumina_soil/16s/Soil_16S_map_corrected.txt –e 121741 –nonphylogenetic_diversity

References

Caporaso, J.G., et al. 2010. QIIME allows analysis of high throughput community sequencing data. *Nature Methods*, 7 (5): 335-336. DOI:10.1038/NMETH.F.303

Kuczynski, J., et al. 2011. Using QIIME to analyze 16S rRNA gene sequences from Microbial Communities. *Curr Protoc Bioinformatics*. CHAPTER: Unit10.7. doi:10.1002/0471250953.bi1007s36

Nguyen, N. H., Smith, D. P., Peay, K. G., & Kennedy, P. G. 2015. Parsing ecological signal from noise in next generation amplicon sequencing. *New Phytologist* 205:1389-1393. DOI: 10.1111/nph.12923