Demo workflow for 16S libraries prepared as described in Kozich et al, AEM 2013 using qiime1.8

12/19/14 - zongzhi

General note

 # Color code in the document
green/red: output blue:command you can copy and paste into the terminal black:notes

 # To make firefox work on omega
you need to close your local firefox and login with -Y like below
ssh -Y zl99@omega.hpc.yale.edu
As the last resort, you might need to copy the directory to your computer.

 # Usage of less to view txt files
q to quit, h for help, arrows/pgup/pgdn/home/end to navigate,
-S to nowrap, -x to set tab stops

Qiime1.8 setup:

 # ssh log in to omega like: ssh -Y zl99@omega.hpc.yale.edu

[zl99@login-0-0 ~]$

 # back up the .bash_profile and .qiime_config

```
cp .bash_profile bash_profile.0
cp .qiime_config qiime_config.0
```

 # seting up bashrc and qiime_config

```
shared=/home/mdi/goodman/shared
echo "source $shared/bashrc_180" >> .bash_profile
cp $shared/qiime_config_180 .qiime_config
mkdir -p ~/scratch/qiime_tmp
exit
```

 # test the settings after relogin to omega like: ssh -Y zl99@omega.hpc.yale.edu

```
print_qiime_config.py -t
```

…

```
   QIIME library version:  1.8.0
   QIIME script version:  1.8.0
```

…

```
----------------------------------------------------------------------
Ran 35 tests in 7.493s
OK
```

Demo setup:

 # work interactively on a compute node

```
qsub -q mdi -IX -d $PWD -l walltime=10:00:00
```

[zl99@compute-33-10 ~]$

 # set up workDir in scratch for demo data, and the output files

```
shared=/home/mdi/goodman/shared
cp -r $shared/qiime18_demo_Kozich ~/scratch
cd ~/scratch/qiime18_demo_Kozich
```

[zl99@compute-26-10 qiime18_demo_Kozich]$

```
# check out the sample mapping file
validate_mapping_file.py -m Demo_master_mapfile.txt
```

No errors or warnings were found in mapping file.

```
masterFile=$PWD/Demo_master_mapfile_corrected.txt
```

```
less -S $masterFile  #q to quit, h for help, arrows to navigate
```

```
#SampleID	BarcodeSequence	LinkerPrimerSequence	ReversePrimer	Plate
1	ATCGTACGAACTCTCG	TATGGTAATTGTGTGCCAGCMGCCGCGGTAA	AGTCA
2	ACTATCTGAACTCTCG	TATGGTAATTGTGTGCCAGCMGCCGCGGTAA	AGTCA
...
```

Preprocessing

 # cat the splited fastq files to four files (R1-4)

```
for ((i=1; i<=4; i++)); do
    zcat raw_data/*R${i}_00?*.fastq.gz > R${i}.fastq
done
```

```
ls *.fastq
```

```
R1.fastq  R2.fastq  R3.fastq  R4.fastq
```

 # join the paired barcodes; change ending 3 to 1 to satisfy the following step

```
join_paired_barcodes.py R3.fastq R2.fastq \
    | sed '1~4 s/ 3:N:0:$/ 1:N:0:/' \
    > paired_barcodes.fastq
```

```
less paired_barcodes.fastq
```

```
@MISEQ:113:000000000-A9GBL:1:1107:26519:22285 1:N:0:
TCTTTCCCGGAGACTA
+
BBBBBDFF1B3>111>
@...
```

 # assemble paired reads, with a min overlaping of 175nt

```
join_paired_ends.py -f R1.fastq -r R4.fastq -b paired_barcodes.fastq \
    -o joined175 --min_overlap 175
```

```
cd joined175
```

```
[zl99@compute-26-10 joined175]$
```

 # count number of reads in joined and not

```
wc -l *.fastq | awk '{print $1/4,$2}'
```

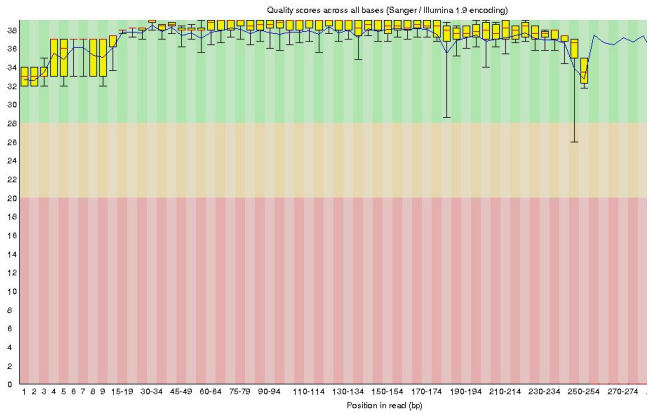
12742 fastqjoin.join_barcodes.fastq
12742 fastqjoin.join.fastq
3258 fastqjoin.un1.fastq
3258 fastqjoin.un2.fastq
32000 total

#QC plots of the joined fastq files

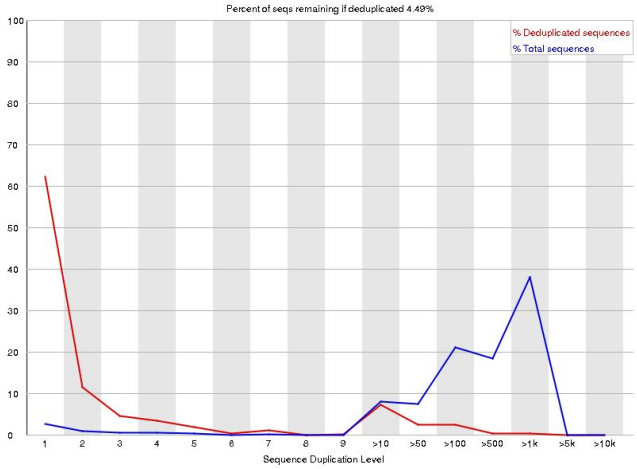
```
fastqc *.join.fastq
```

```
firefox *.join_fastqc.html  #you might have to close your local firefox
```

Per base sequence quality



Sequence Duplication Levels



##split_libraries and filtering by Q value>20

```
split_libraries_fastq.py -v -q 19 -m $masterFile --phred_offset 33 \
     --barcode_type 16 -b *.join_barcodes.fastq \
     -i *.join.fastq -o splitQ20
```

```
cd splitQ20
```

[zl99@compute-38-3 splitQ20]$

```
less histograms.txt
```

```
Length  Count
…
247.0   11338
257.0   0
--
```

```
less seqs.fna
```

```
>170_0 MISEQ:113:000000000-A9GBL:1:1107:15843:22285 1:N:0: orig_bc=GGATATCTATACTTCG
new_bc=GGATATCTATACTTCG bc_diffs=0
TACGTAGGGTGCAAGCGTTATCCGGAATTATTGGGCGTAAAGGGCTCGTAGGCGGTTCGTCGCGTCCGGT
…
```

OTU picking and taxonomy assignment

```
# set references files
```

```
shared=/home/mdi/goodman/shared
```

```
refFa=$shared/DB/gg_13_5_otus/rep_set/97_otus.fasta
```

```
refTaxonomy=$shared/DB/gg_13_5_otus/taxonomy/97_otu_taxonomy.txt
```

```
# pick OTUs with default method (uclust) and make phylogeny
```

```
pick_open_reference_otus.py --suppress_taxonomy_assignment -r $refFa \
    -i seqs.fna -o otuUclust
```

```
Unsupported or depreciated options passed to pynast: temp_dir
  blast_db, max_e_value, and addl_blast_params are depreciated and will be removed in PyNAST 1.2.
```

This step will take a while and the above warning can be ignored for now. For a big dataset, you might want to add -a -O for parallel computing like below:

```
echo "pick_open_reference_otus.py --suppress_taxonomy_assignment -r $refFa \
    -i seqs.fna -o otuUclust -a -O 8" > pickOtus.qsub

qsub -q mdi -d . pickOtus.qsub
```

cd otuUclust

[zl99@compute-38-3 otuUclust]$

less -S -x 11 final_otu_map_mc2.txt

851733 143_1652 143_1968 143_2048 143_2550 143_3214 143_4297 141_6365
4347520 16_309 105_386 105_688 98_1041 153_1256 148_1746 16_1760
…

less rep_set.fna

>1004910 59_2725
TACGGAAGGTCCAGGCGTTATCCGGATTTATTGGGTTTAAAGGGAGTGTAGGCGGTTTGTTAAGCGTGTTGTGAA
ATTTAGATGCTCAACATTTAACTTGCAGCGCGAACTGGCGAACTTGAGTGCACACAACGTATGCGGAATTCATGGT
GTAG…

Taxonomy assignment with default method (uclust)

assign_taxonomy.py -t $refTaxonomy -r $refFa -i rep_set.fna -o taxUclust

cd taxUclust

[zl99@compute-38-3 taxUclust]$

less -S *tax_assignments.txt

851733 k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Lactobacillaceae;
4347520 k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__; g__; s__
364538 k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Porphyromona
…

add taxonomy back to OTU table

```
biom add-metadata --sc-separated taxonomy \
    --observation-header OTUID,taxonomy \
    --observation-metadata-fp *tax_assignments.txt \
    -i ../otu_table_mc2.biom -o otu_table_mc2.tax.biom
```

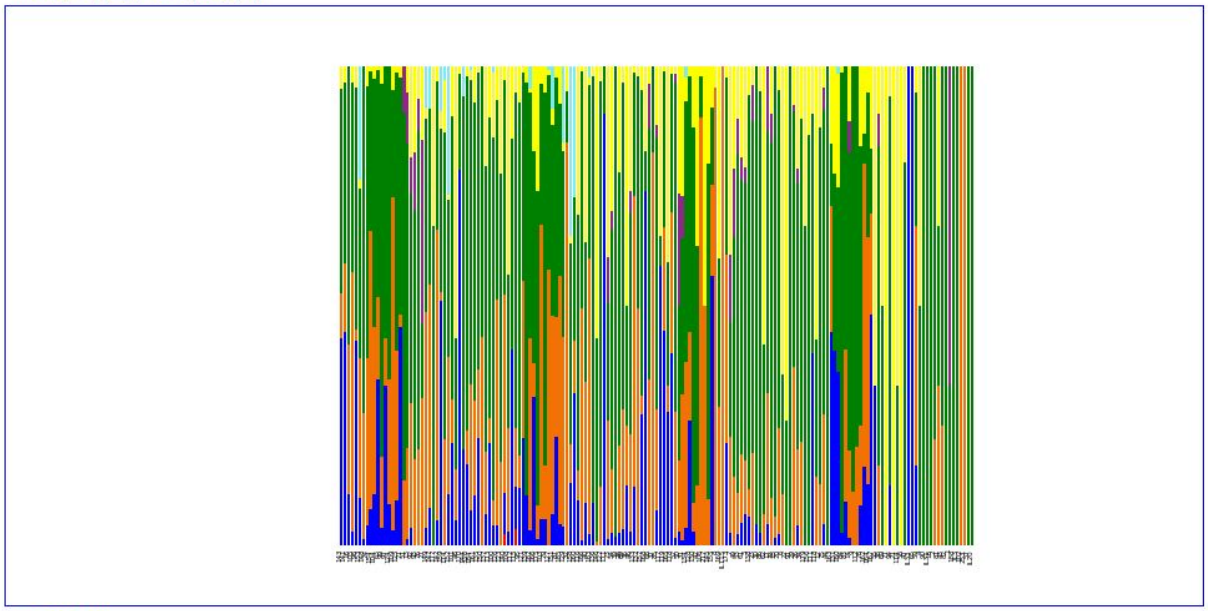
summarize taxanomy

#to plot by category include –m mapping file and –c <category>

```
summarize_taxa_through_plots.py \
    -i otu_table_mc2.tax.biom -o summarize_taxa_out
```

```
firefox summarize_taxa_out/taxa_summary_plots/bar_charts.html
```

View Figure (.pdf) View Legend (.pdf)



View Table (.txt)

		Total		143	141	16	105	98	153	148	24	154	104	63	99	97	124	109	125	71	51	13	92	93	76	27	14
Legend	Taxonomy	count	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
	k__Archaea;p__Euryarchaeota	0	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0
	k__Bacteria;p__Actinobacteria	20	11.6%	43.2%	44.4%	10.7%	2.8%	42.7%	10.0%	1.3%	4.1%	7.6%	10.7%	34.5%	3.7%	33.3%	8.7%	3.2%	9.5%	45.7%	0.0%	1.4%	3.6%	0.0%	0.0%	0.0%	3.7
	k__Bacteria;p__Bacteroidetes	36	21.2%	9.5%	14.4%	31.1%	54.2%	2.2%	29.1%	26.3%	35.1%	58.1%	34.8%	17.3%	14.8%	10.0%	26.0%	69.4%	31.1%	2.5%	13.5%	18.9%	26.2%	18.0%	20.0%	30.8%	45.1
	k__Bacteria;p__Firmicutes	86	50.3%	42.6%	37.8%	58.2%	39.6%	50.6%	35.5%	72.4%	56.7%	33.3%	51.8%	47.3%	77.8%	56.7%	65.4%	22.6%	58.1%	49.4%	76.9%	63.5%	44.0%	51.7%	66.7%	15.4%	40.2
	k__Bacteria;p__Fusobacteria	3	1.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	9.6%	10.8%	7.1%	12.4%	6.7%	38.5%	0.0
	k__Bacteria;p__Proteobacteria	24	13.9%	4.7%	3.3%	0.0%	3.5%	4.5%	1.8%	0.0%	4.1%	1.0%	2.7%	0.9%	3.7%	0.0%	0.0%	4.8%	1.4%	2.5%	0.0%	5.4%	19.0%	18.0%	6.7%	15.4%	2.4
	k__Bacteria;p__Verrucomicrobia	2	1.1%	0.0%	0.0%	0.0%	0.0%	0.0%	23.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	8.5

you might want to use compare_taxa_summaries.py for differential enrichment

```
# convert biom to matrix/spreadsheet

biom convert -b --header-key taxonomy \
    -i otu_table_mc2.tax.biom -o otu_table_mc2.tax.txt

less -S otu_table_mc2.tax.txt

# Constructed from biom file
#OTU ID 143    141    16     105    98     153    148    24     154    104
851733  15.0   4.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0
4347520 0.0    0.0    6.0    3.0    1.0    2.0    2.0    1.0    2.0    2.0

# biom filtering

# normally singletons (OTUs represented by a single sequence) would be removed

filter_otus_from_otu_table.py -n 2 \
    -i otu_table_mc2.tax.biom -o otu_table_mc2.tax.no_singles.biom
```

Alpha diversity

```
#[zl99@compute-38-3 taxUclust]$
```

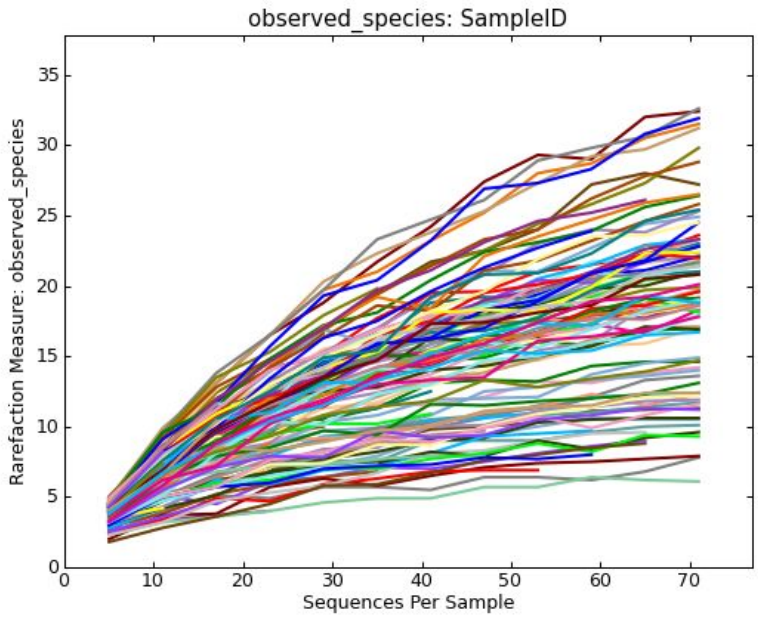
```
# Subsample data so each sample has equivalent sequencing depth
```

Note that real datasets should be rarified to 5000-10000 reads/sample, a depth (-d) of 5,000+ would be used

```
single_rarefaction.py -d 100 \
    -i otu_table_mc2.tax.biom -o otu_table_mc2.tax.100.biom

# Calculate alpha diversity of each sample at various sequencing depths
```

Note that many alpha diversity metrics can be specified by in a parameter file. An example is shown below.

```
echo 'alpha_diversity:metrics observed_species,PD_whole_tree' \
    > alpha_diversity.params

alpha_rarefaction.py -p alpha_diversity.params --min_rare_depth 5 -n 10 \
    -m $masterFile -t ../rep_set.tre \
    -i otu_table_mc2.tax.biom -o alpha_rarefaction_out

firefox alpha_rarefaction_out/alpha_rarefaction_plots/*.html
```



For larger datasets, the number of steps should be more to make a smooth curve. And -a -O can used for parallel computing like below. Compare_alpha_diversity.py can be used for downstream analysis.

```
echo "alpha_rarefaction.py -p alpha_diversity.params --min_rare_depth 5 -n 10 \
    -m $masterFile -t ../rep_set.tre \
    -i otu_table_mc2.tax.biom -o alpha_rarefaction_out \
    -a -O 8" > alphaRarefaction.qsub

qsub -q mdi -d . alphaRarefaction.qsub
```

Beta diversity

#[zl99@compute-38-3 taxUclust]$

Filter OTU table to just include samples of interest for beta-diversity analysis

Note that OTU table can also be subsampled based on multiple columns in the mapping file using multiple passes of filter_samples_from_otu_table.py

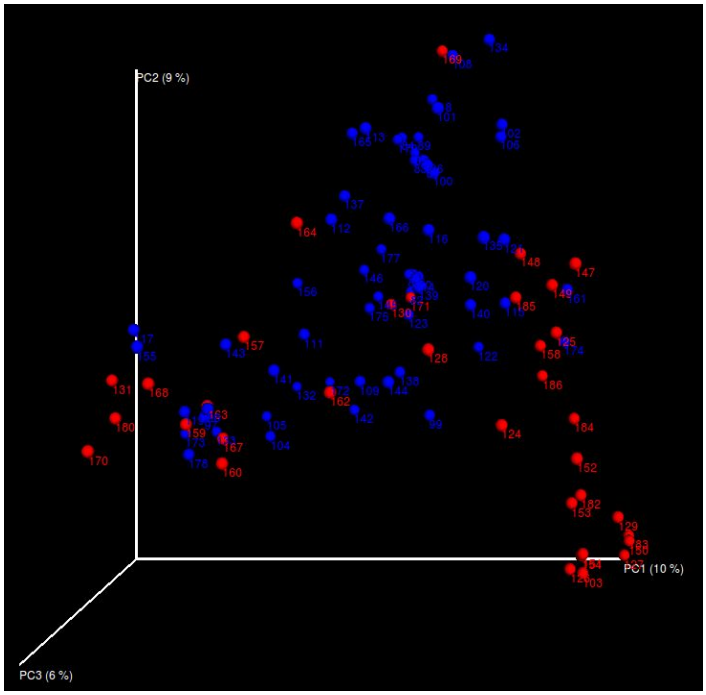
```
filter_samples_from_otu_table.py -m $masterFile -s 'Stool_Sputum:STOOL' \
    -i otu_table_mc2.tax.biom -o otu_table_mc2.tax.stool_samples.biom
```

Calculate beta diversity and make plots

Note that many beta diversity metrics can be specified in a param file.

```
echo 'beta_diversity:metrics hellinger' > beta_diversity.params

beta_diversity_through_plots.py -p beta_diversity.params \
    -t ../rep_set.tre  -m $masterFile \
    -i otu_table_mc2.tax.stool_samples.biom -o stool_samples_beta_div_out
```

you might want to use --seqs_per_sample=10000 for even sampling
To view the plots, you need to copy to your computer to make it works.
firefox stool_samples_beta_div_out/hellinger_emperor_pcoa_plot/*.html



For a larger dataset, -a -O can be used for parallel computing like below.

```
echo "beta_diversity_through_plots.py -p beta_diversity.params \
    -t ../rep_set.tre  -m $masterFile \
    -i otu_table_mc2.tax.stool_samples.biom -o stool_samples_beta_div_out \
    -a -O 8" > betaDiversity.qsub

qsub -q mdi -d . betaDiversity.qsub
```

Final note

this is not an exhaustive demo of what is available in QIIME. Resources for more info:

http://qiime.org/documentation/index.html

http://qiime.org/scripts/index.html

https://groups.google.com/forum/#!forum/qiime-forum