From Fadrosh et al., 2014
https://github.com/igsbma/MiSeq16S/blob/master/README

```
fastx_trimmer -h
     usage: fastx_trimmer [-h] [-f N] [-l N] [-z] [-v] [-i INFILE] [-o
OUTFILE]

     version 0.0.6
        [-h]         = This helpful help screen.
        [-f N]       = First base to keep. Default is 1 (=first base).
        [-l N]       = Last base to keep. Default is entire read.
        [-z]         = Compress output with GZIP.
        [-i INFILE]  = FASTA/Q input file. default is STDIN.
        [-o OUTFILE] = FASTA/Q output file. default is STDOUT.

#1. Trim first 6 nt as barcode, form a new barcode file

fastx_trimmer -i Undetermined_S0_L001_R1_001.fastq -f 1 -l 6 -Q 33 -o
R1_barcode.fastq
fastx_trimmer -i Undetermined_S0_L001_R2_001.fastq -f 1 -l 6 -Q 33 -o
R2_barcode.fastq

#read barcode file data, pipe output to fq_mergelines.pl, create temp
barcode file

cat R1_barcode.fastq | perl fq_mergelines.pl > R1_barcode_temp
cat R2_barcode.fastq | perl fq_mergelines.pl > R2_barcode_temp

#paste contents of temp barcode files side-by-side, tab separated. Pipe
to awk

paste R1_barcode_temp R2_barcode_temp | awk -F"\t" '{print
$5"\t"$2$6"\t"$3"\t"$4$8}' | perl fq_splitlines.pl > R1R2_barcode.fastq

10255350  : R1R2_barcode.fastq (Sequence lengths (mean +/- std): 12.0000
+/- 0.0000)
10255350  : Total

#trim 6nt barcodes off the left end of each raw read sequence
./seqtk trimfq -b 6 Undetermined_S0_L001_R1_001.fastq >
R1_trimmed_seq.fastq
./seqtk trimfq -b 6 Undetermined_S0_L001_R2_001.fastq >
R2_trimmed_seq.fastq

#2. assemble paired end read using either pandaseq, FLASH, or SeqPrep.
QIIME 1.8 version also implement the function of merging reads ends.
# note that different merging algorithm has its own issues and
disadvantages, choose carefully based on your needs (efficiency,
overlapping length, quality, etc.)
# await evaluating ea-utils

#a)using pandaseq, note that the overlapping length needed to be
determined based on read length and amplicon size
```

```
#Run Pandaseq to assemble MiSeq1 paired-end reads (output as fastq, min
overlap of 140, max overlap of 300, allow sequences to lack barcodes)

pandaseq -F -o 140 -O 300 -B -f R1_trimmed_seq.fastq -r
R2_trimmed_seq.fastq > PandaAssembly.fq

9461515  : PandaAssembly.fq (Sequence lengths (mean +/- std): 341.0907
+/- 59.2750)
9461515  : Total

#4) fix the pandaseq headers to make them compatible with QIIME
#first fix the extra text in the headers.
sed 's/:0$/ 2:N:0:0/g' PandaAssembly.fq > pandAssFixed.fastq

sed -n '1~4'p pandAssFixed.fastq | sed -e 's/^@//g' >
pandAssFixed.keep.header

#Filter the barcodes file, so that only barcodes present in your
sequences are included.

filter_fasta.py -f R1R2_barcode.fastq -o filtered_Barcodes.fastq -s
pandAssFixed.keep.header

#Pull out your header files from this new filtered barcodes fastq

sed -n '1~4'p filtered_Barcodes.fastq | sed 's/^@//g' >
headerLines_filtered_Barcodes.txt

#run diff -s to make sure that all the headers are going to be the same.
It will tell you if it's identical, or show you the offending sequence.

diff -s headerLines_filtered_Barcodes.txt pandAssFixed.keep.header

#demultiplexing (quality filtering turned off)

split_libraries_fastq.py -i pandAssFixed.fastq -b filtered_Barcodes.fastq
--barcode_type 12 -o slout_q20 -m mappingfileBermudaAE15wspacers.txt -q 0
-n 999 -r 999 -p .0001

8133355  : slout_q20/seqs.fna (Sequence lengths (mean +/- std): 336.1627
+/- 59.1550)
8133355  : Total

5. extract 16S and 18S reads into separate files
less slout_q20/seqs.fna |pcregrep -M "^>[A-Z][1-9]?[1-
9].18S[^>]+$">Berm15_18s.fasta
less slout_q20/seqs.fna |pcregrep -M "^>[A-Z][1-9]?[1-
9].16S[^>]+$">Berm15_16s.fasta

6. remove chimeras with usearch 6.1

identify_chimeric_seqs.py -i Berm15_16s.fasta -m usearch61 -o
usearch_checked_chimeras16s/ -r
```

```
~/Documents/Silva123.1/Silva_123_provisional_release/SILVA123_QIIME_relea
se/rep_set/rep_set_16S_only/97/97_otus_16S.fasta

filter_fasta.py -f Berm15_16s.fasta -o 16sseqs_chimeras_filtered.fna -s
usearch_checked_chimeras16s/chimeras.txt -n

5361308  : 16sseqs_chimeras_filtered.fna (Sequence lengths (mean +/-
std): 296.7290 +/- 2.9214)
5361308  : Total

identify_chimeric_seqs.py -i Berm15_18s.fasta -m usearch61 -o
usearch_checked_chimeras18s/ -r
~/Documents/Silva123.1/Silva_123_provisional_release/SILVA123_QIIME_relea
se/rep_set/rep_set_18S_only/97/97_otus_18S.fasta

filter_fasta.py -f Berm15_18s.fasta -o 18sseqs_chimeras_filtered.fna -s
usearch_checked_chimeras18s/chimeras.txt -n

2521797  : 18sseqs_chimeras_filtered.fna (Sequence lengths (mean +/-
std): 424.1804 +/- 4.2959)
2521797  : Total

7. Tag cleaning using tagcleaner

perl tagcleaner.pl -fasta
~/Documents/Berm15taxf/16sseqs_chimeras_filtered.fna -out 16Sclean -
line_width 0 -verbose -log tagclean16S -nomatch 3 -tag5
GTGYCAGCMGCCGCGGTAA -mm5 3 -tag3 ATTAGATACCCVNGTAGTC -mm3 3 -trim_within
24

perl tagcleaner.pl -fasta
~/Documents/Berm15taxf/18sseqs_chimeras_filtered.fna -out 18Sclean -
line_width 0 -verbose -log tagclean18S -nomatch 3 -tag5
CCAGCASCYGCGGTAATTCC -mm5 3 -tag3 ATCAAGAACGAAAGT -mm3 3 -trim_within 25

8. pick otus and assign taxonomy
pick_open_reference_otus.py -i ~/Documents/Berm15taxf/16Sclean.fasta -m
usearch61 -r
~/Documents/Silva123.1/Silva_123_provisional_release/SILVA123_QIIME_relea
se/rep_set/rep_set_16S_only/97/97_otus_16S.fasta -p
~/Documents/Berm15taxa/silva-16s_paramsBerm15.txt -o
~/Documents/Berm15taxf/Berm15_16Sotus/ -f

pick_open_reference_otus.py -i ~/Documents/Berm15taxf/18Sclean.fasta -m
usearch61 -r
~/Documents/Silva123.1/Silva_123_provisional_release/SILVA123_QIIME_relea
se/rep_set/rep_set_18S_only/97/97_otus_18S.fasta -p
~/Documents/Berm15taxa/silva-18s_paramsBerm15.txt -o
~/Documents/Berm15taxf/Berm15_18Sotus/ -f

9. biom summarize-table -i
Berm15_16Sotus/otu_table_mc2_w_tax_no_pynast_failures.biom
Num samples: 24
Num observations: 6127
```

```
Total count: 5334461
Table density (fraction of non-zero values): 0.285

Counts/sample summary:
 Min: 109410.0
 Max: 302092.0
 Median: 232346.500
 Mean: 222269.208
 Std. dev.: 54914.100
 Sample Metadata Categories: None provided
 Observation Metadata Categories: taxonomy

Counts/sample detail:
 Q44.16S: 109410.0
 Q6.16S: 140825.0
 S1.16S: 145646.0
 Q1.16S: 147525.0
 Q39.16S: 170236.0
 Q17.16S: 173061.0
 Q48.16S: 173574.0
 S31.16S: 188647.0
 S34.16S: 195110.0
 Q42.16S: 209702.0
 S26.16S: 221644.0
 S25.16S: 227103.0
 Q11.16S: 237590.0
 Q56.16S: 240440.0
 S29.16S: 240980.0
 Q54.16S: 255015.0
 S4.16S: 263929.0
 Q37.16S: 269705.0
 Q64.16S: 277221.0
 S21.16S: 278961.0
 Q58.16S: 282406.0
 S7.16S: 283879.0
 Q35.16S: 299760.0
 D2.16S: 302092.0

   biom summarize-table -i
Berm15_18Sotus/otu_table_mc2_w_tax_no_pynast_failures.biom
Num samples: 20
Num observations: 8429
Total count: 2439804
Table density (fraction of non-zero values): 0.233

Counts/sample summary:
 Min: 40659.0
 Max: 258132.0
 Median: 105990.000
 Mean: 121990.200
 Std. dev.: 57871.339
 Sample Metadata Categories: None provided
 Observation Metadata Categories: taxonomy
```

```
Counts/sample detail:
 Q48.18S: 40659.0
 S34.18S: 50322.0
 Q37.18S: 60844.0
 S26.18S: 64348.0
 S31.18S: 70693.0
 Q11.18S: 86027.0
 Q6.18S: 88773.0
 S4.18S: 94774.0
 Q56.18S: 98673.0
 Q39.18S: 99349.0
 Q54.18S: 112631.0
 Q58.18S: 116469.0
 Q17.18S: 127798.0
 S21.18S: 160734.0
 Q35.18S: 167505.0
 Q44.18S: 168014.0
 S7.18S: 168019.0
 Q64.18S: 173724.0
 D2.18S: 232316.0
 S25.18S: 258132.0
```

10. #discard metazoan otus, discard control otus

```
filter_taxa_from_otu_table.py -i
Berm15_16Sotus/otu_table_mc2_w_tax_no_pynast_failures.biom -o
Berm15_16Sotus/otu_table_controls_filtered.biom -n "D_1__Deinococcus-
Thermus"

filter_taxa_from_otu_table.py -i
Berm15_18Sotus/otu_table_mc2_w_tax_no_pynast_failures.biom -o
Berm15_18Sotus/18Sotu_table_Metazoa_controls_filtered_Syndinialeskeep.bio
m -n "D_3__Metazoa
(Animalia)",D_2__Chloroplastida,D_9__Schizosaccharomycetaceae
```

11. results

```
biom summarize-table -i
Berm15_18Sotus/18Sotu_table_Metazoa_controls_filtered_Syndinialeskeep.bio
m

Num samples: 20
Num observations: 7843
Total count: 2094122
Table density (fraction of non-zero values): 0.238

Counts/sample summary:
 Min: 33245.0
 Max: 249403.0
 Median: 89849.500
 Mean: 104706.100
 Std. dev.: 54089.729
 Sample Metadata Categories: None provided
```

```
 Observation Metadata Categories: taxonomy

Counts/sample detail:
 Q48.18S: 33245.0
 S34.18S: 36162.0
 Q39.18S: 51194.0
 S26.18S: 53217.0
 Q37.18S: 56141.0
 S31.18S: 62054.0
 Q56.18S: 76951.0
 Q11.18S: 82769.0
 S4.18S: 84725.0
 Q6.18S: 86861.0
 Q54.18S: 92838.0
 Q58.18S: 102331.0
 Q17.18S: 114628.0
 S21.18S: 126363.0
 S7.18S: 135245.0
 Q35.18S: 139766.0
 Q64.18S: 160190.0
 Q44.18S: 163446.0
 D2.18S: 186593.0
 S25.18S: 249403.0

core_diversity_analyses.py -i
Berm15_18Sotus/18Sotu_table_Metazoa_controls_filtered_Syndinialeskeep.bio
m -o 18scdoutkeepSyndiniales/ -m mappingfileBermudaAE15wspacers.txt -t
Berm15_18Sotus/Berm15_18S_rep_set.tre -e 33245

summarize_taxa_through_plots.py -i
Berm15_18Sotus/18Sotu_table_Metazoa_controls_filtered_Syndinialeskeep.bio
m -o taxsumkeepSynd/

summarize_taxa.py -i
Berm15_18Sotus/18Sotu_table_Metazoa_controls_filtered_Syndinialeskeep.bio
m -o 18S_filtered_keepSynd_summ_abs/ -a
```