Pipeline for 454 Cattle with QIIME

1. Validate Mapping file

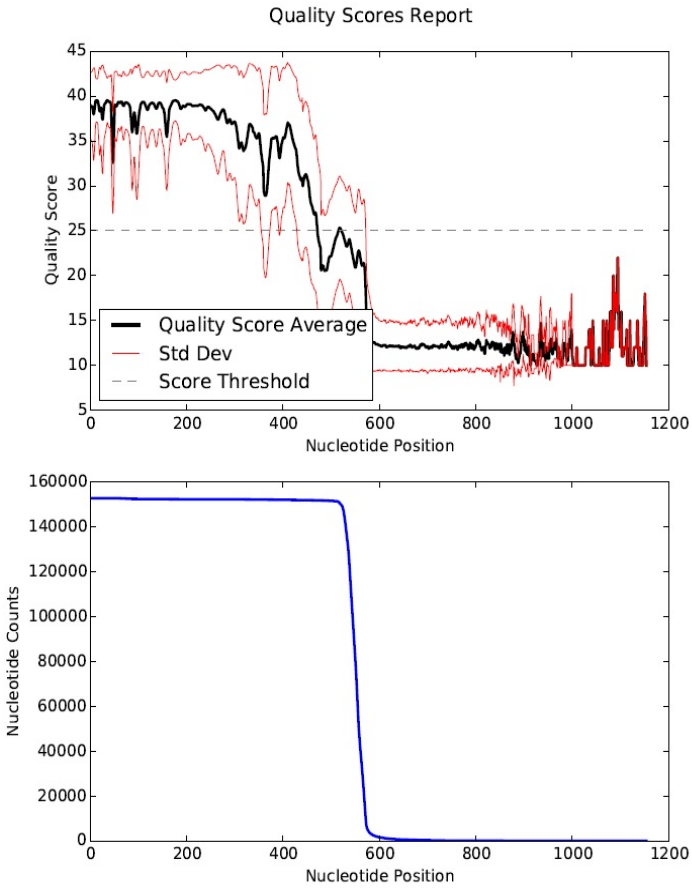
```
validate_mapping_file.py -m 'map'$i'.txt' -o map_corrected;
```

2. Process Sff file

```
process_sff.py -i 'Run'$i'.sff' -o sff_processed --make_flowgram;
```

3. Quality score plots

```
quality_scores_plot.py -q 'sff_processed/Run'$i'.qual' -o quality_histograms/;
```



In this plot we can see that read quality drops off after the 600 bp nucleotide position. Most of our read lengths will fall below this threshold with subsequent processing in the next step.

---

4. Filter sequences

```
split_libraries.py -m 'map_corrected/map'$i'_corrected.txt' -f 'sff_processed/Run'$i'.fna' -q 'sff_pro
cessed/Run'$i'.qual' -o split_default -b variable_length -z truncate_only -n $((i * 1000000));
```

5. Loop it for all plates/runs

```
for i in {1..28};
do
 ...
done
```

Run this in linux command prompt

```
for i in {1..28};
do
echo $i &>> '/media/sf_S/Brian Nohomovich/Cattle/Necessary_Resources/log.txt';
cd '/media/sf_S/Brian Nohomovich/Cattle/'$i;
validate_mapping_file.py -m 'map'$i'.txt' -o map_corrected;
process_sff.py -i 'Run'$i'.sff' -o sff_processed --make_flowgram;
quality_scores_plot.py -q 'sff_processed/Run'$i'.qual' -o quality_histograms/;
split_libraries.py -m 'map_corrected/map'$i'_corrected.txt' -f 'sff_processed/Run'$i'.fna' -q 'sff_pro
cessed/Run'$i'.qual' -o split_default -b variable_length -z truncate_only -n $((i * 1000000));
done
```

Sequence Quality Assessment

Goal: Provide depth of sample after quality assessment.

1. Take last lines of log-file; skip firest 29

```
tail -n +29 '/media/sf_S/Brian Nohomovich/Cattle/'$i'/split_default/split_library_log.txt'
```

2. Starting at top of remaining lines; ignore first line

```
head -n -1
```

3. Print the sample ID and read count for sample

```
awk '{ print $1"\t" $2 }'
```

4. Sort by sequence ID and output to counts for sample

```
sort -k 1 > '/media/sf_S/Brian Nohomovich/Cattle/'$i'/split_default/counts.txt'
```

5. Loop it for all plates/runs

```
for i in {1..28};
do
 ...
done
```

Run this in linux command prompt

```
for i in {1..28};
do tail -n +29 '/media/sf_S/Brian Nohomovich/Cattle/'$i'/split_default/split_library_log.txt' | head -
n -1 | awk '{ print $1"\t" $2 }' | sort -k 1 > '/media/sf_S/Brian Nohomovich/Cattle/'$i'/split_default
/counts.txt';
done
```

Denoise

```
denoise_wrapper.py -i sff_processed/Run$i.txt -f split_default/seqs.fna -m 'map_corrected/map'$i'_corr
ected.txt' --titanium --force_overwrite -n 200;

inflate_denoiser_output.py -c centroids1.fna,centroids2.fna -s singletons1.fna,singletons2.fna -f seqs
1.fna,seqs2.fna -d denoiser_mapping1.txt,denoiser_mapping2.txt -o denoised_seqs.fna
```

Run the denoise_wrapper.py for each sequencing run then inflate them all with inflate_denoiser.py. This will create a combined denoised dataset.

Vsearch (instead of Usearch)

Usearch is only usable on small datasets unless the licensed version is purchased. Vsearch and other softwares are also permissable but require their own commands to get files that can be then loaded back into QIIME. Vsearch performs better in benchmarking.

The following steps are important to format read names correctly for use in Vsearch

```
sed 's/\_[0-9]*//g' '/media/sf_S/Brian_Nohomovich/Cattle/merged/denoised_seqs_all.fna' > removeID.fna
```

Remove trailing _numerical read identifintication one each read. This is added by qiime

```
sed 's/\./\_/g' '/media/sf_S/Brian_Nohomovich/Cattle/merged/removeID.fna' > finalReads.fna
```

Converts all . to _ in sampleID.

```
sed -e 's/>/>sample=/' /media/sf_S/Brian_Nohomovich/Cattle/merged/finalreads.fna
```

Identifies sampleID in read name.

Further QC with Vsearch

1. Dereplication

```
vsearch -derep_fulllength ../removeID.fna --sizeout --relabel derep_ --minuniquesize 2 --output unique
_seqs.fna --log derep.log.txt
```

2. Chimera Detection and removal (denovo and reference)

```
vsearch --uchime3_denovo unique_seqs.fna --sizein --sizeout --nonchimeras denovo.fna --log denovo_chim
era_log.txt;
```

Denovo chimera removal using uchime3 algorithm.

```
vsearch --uchime_ref denovo.fna --db /media/sf_S/Brian_Nohomovich/Cattle/Necessary_Resources/gold.fa -
-nonchimeras refchi.fna -log ref_chimera_log.txt;
```

Reference based chimeric read removal

3. OTUs and rep_set

Cluster for OTUs at given id %; id=0.X parameter

97 (Genus)

```
vsearch --cluster_size refchi.fna --id 0.97 --centroids 97/rep_set.fna --log 97/rep_set.log --sizein -
-xsize --relabel OTU_  --biomout 97/otu_table.biom --uc 97/all.clustered.uc -log 97/otu_log.txt;
```

99 (Species)

```
vsearch --cluster_size refchi.fna --id 0.99 --centroids 99/rep_set.fna --log 99/rep_set.log --sizein -
```

```
-xsize --relabel OTU_  --biomout 99/otu_table.biom --uc 99/all.clustered.uc -log 99/otu_log.txt;
```

4. Mapping

Maps reads to OTUs to generate table.

97

```
vsearch -usearch_global ../../finalreads.fna --sizein --sizeout --db rep_set.fna --id 0.97 -log otu_ma
pping_log.txt -uc otu_table_mapping.uc --biomout otu_table_mapped.biom;
```

99

```
vsearch -usearch_global ../finalreads.fna --sizein --sizeout --db rep_set.fna --id 0.99 -log otu_mappi
ng_log.txt -uc  otu_table_mapping.uc --biomout otu_table_mapped.biom;
```

5. Assign taxonomy to OTUs

Assigns taxonomy (will most likely need HPCC for this) using the Silva database (v132).

```
parallel_assign_taxonomy_uclust.py -i rep_set.fna -o uclust_assign_tax300 --jobs_to_start 20 --referen
ce_seqs_fp "/mnt/research/manninglab/Brian/Cattle/Necessary_Resources/Silva_132_release/SILVA_132_QIIM
E_release/rep_set/rep_set_16S_only/99/silva_132_99_16S.fna" --id_to_taxonomy_fp "/mnt/research/manning
lab/Brian/Cattle/Necessary_Resources/Silva_132_release/SILVA_132_QIIME_release/taxonomy/16S_only/99/ta
xonomy_7_levels.txt" --similarity 0.8
```

6. Add metadata

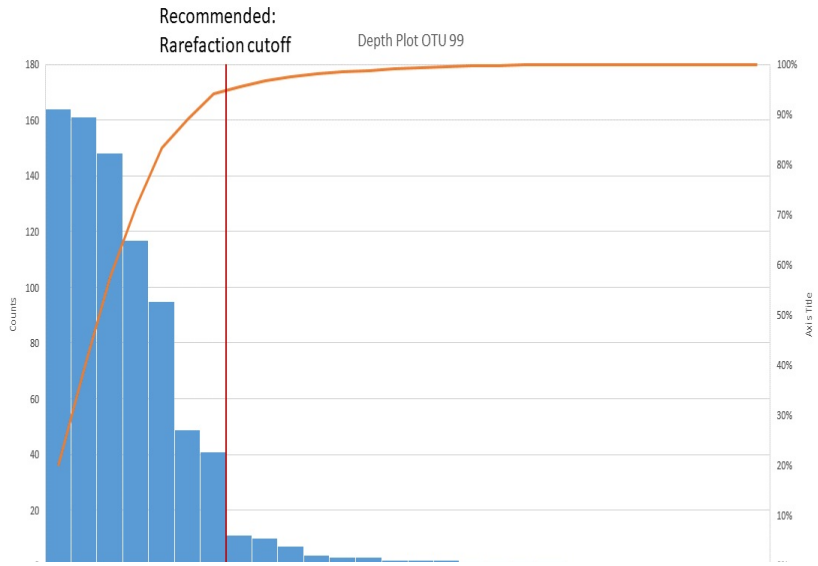
Adds metadata (otu and ID) to biom table.

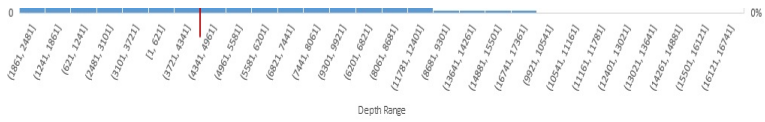
```
biom add-metadata -i otu_table_mapped.biom -o otu_table_w_tax.biom --observation-metadata-fp uclust_as
sign_tax/rep_set_tax_assignments.txt --sc-separated taxonomy --observation-header OTUID,taxonomy
```

7. Summarize and convert to text

Summarizes depth of each sample ID.

```
biom summarize-table -i otu_table_w_tax.biom -o otu_table_w_tax.summary.txt
```





Greater than 90% of the data is found between samples with a depth of 1-4341. ~25% of the samples have a depth less than 700. It is recommended that rarefaction be considered to optimize the diversity and sample size of the dataset. Analysis might need to be ran several times to find the optimal tradeoff. See otu_table_w_tax.summary for the raw data.

---

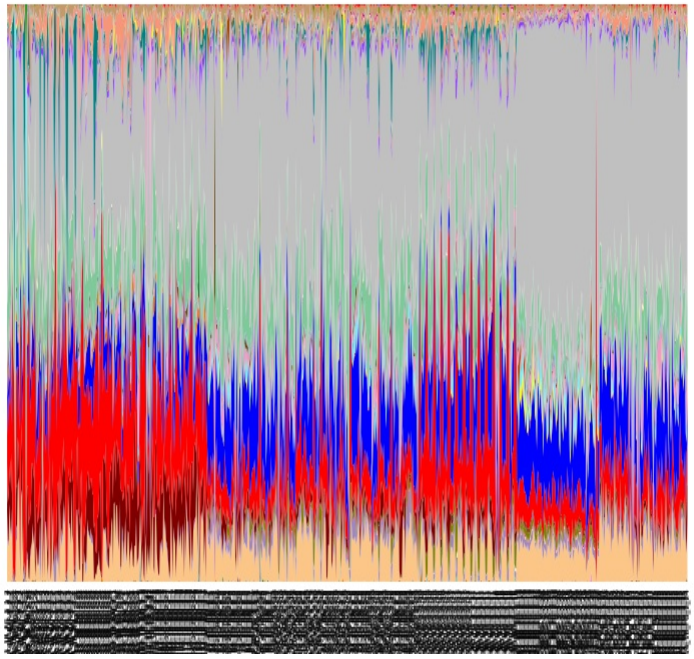
8. Convert to text

Converts biom table to text file.

```
biom convert -i otu_table_w_tax.biom -o otu_table_w_tax.txt --to-tsv --header-key taxonomy
```

9. Preliminary Analysis

```
summarize_taxa_through_plots.py -i '/media/sf_S/Brian_Nohomovich/Cattle/merged/otus_v/99/otu_table_w_tax.biom' -o taxa_summary -m '/media/sf_S/Brian_Nohomovich/Cattle/Combined/merged_mapping.txt'
```



Further investigation is needed but preliminary analysis appears to highlight some trends.

==========

Misc.

Merge sequences, quality, and map files

Map files

1. Find all map files and create list

```
find '/media/sf_S/Brian Nohomovich/Cattle' -name map*_corrected.txt maps.csv
```

2. Merge map list using qiimes built in function

2. Merge map file using qiime built in function

```
merge_mapping_files.py -m '/media/sf_S/Brian Nohomovich/Cattle/1/map_corrected/map1_corrected.txt','/media/sf_S/Brian Nohomovich/Cattle/10/map_corrected/map10_corrected.txt','/media/sf_S/Brian Nohomovich/Cattle/11/map_corrected/map11_corrected.txt','/media/sf_S/Brian Nohomovich/Cattle/16/map_corrected/map16_corrected.txt','/media/sf_S/Brian Nohomovich/Cattle/19/map_corrected/map19_corrected.txt','/media/sf_S/Brian Nohomovich/Cattle/2/map_corrected/map2_corrected.txt','/media/sf_S/Brian Nohomovich/Cattle/22/map_corrected/map22_corrected.txt','/media/sf_S/Brian Nohomovich/Cattle/23/map_corrected/map23_corrected.txt','/media/sf_S/Brian Nohomovich/Cattle/24/map_corrected/map24_corrected.txt','/media/sf_S/Brian Nohomovich/Cattle/25/map_corrected/map25_corrected.txt','/media/sf_S/Brian Nohomovich/Cattle/26/map_corrected/map26_corrected.txt','/media/sf_S/Brian Nohomovich/Cattle/27/map_corrected/map27_corrected.txt','/media/sf_S/Brian Nohomovich/Cattle/28/map_corrected/map28_corrected.txt','/media/sf_S/Brian Nohomovich/Cattle/3/map_corrected/map3_corrected.txt','/media/sf_S/Brian Nohomovich/Cattle/4/map_corrected/map4_corrected.txt','/media/sf_S/Brian Nohomovich/Cattle/5/map_corrected/map5_corrected.txt','/media/sf_S/Brian Nohomovich/Cattle/6/map_corrected/map6_corrected.txt','/media/sf_S/Brian Nohomovich/Cattle/7/map_corrected/map7_corrected.txt','/media/sf_S/Brian Nohomovich/Cattle/8/map_corrected/map8_corrected.txt','/media/sf_S/Brian Nohomovich/Cattle/9/map_corrected/map9_corrected.txt' -o merged_mapping.txt -n 'NA'
```

Quality files

1. Find all qual files and combine

```
find '/media/sf_S/Brian Nohomovich/Cattle' -name Run*.qual -exec cat {} + > Combined/allqual.qual
```

2. Check that size matches

```
find '/media/sf_S/Brian Nohomovich/Cattle' -name Run*.qual -exec du -ch {} + |tail -1| cut -f 1
```

Seqs files

1. Find all seqs files and combine

```
find '/media/sf_S/Brian Nohomovich/Cattle' -name seqs.fna -exec cat {} + > Combined/allseqs.fna
```

2. Check that size matches

```
find '/media/sf_S/Brian Nohomovich/Cattle' -name seqs.fna -exec du -ch {} + |tail -1| cut -f 1
```

Now that we have merged sequences from each plate that has been quality controlled we can now find OTUs within these sequences. We will use a parameter file that has been setup to use *pick_de_novo_otus* in qiime.

More QIIME workflow (can be skipped)

The biom file generated from vsearch can be loaded into QIIME2. These steps are not

---

pick_de_novo_otus.py

```
pick_de_novo_otus.py -i '/media/sf_S/Brian Nohomovich/Cattle/Combined/allseqs.fna' -o '/media/sf_S/Brian Nohomovich/Cattle/Combined/otus_Silva' -p '/media/sf_S/Brian Nohomovich/Cattle/Necessary_Resources/parameter.txt'
```

Parameter File (the -p)

This text file informs qiime to utilize a different database, chimeric correction, and otu picking. It is also defines the alpha and beta diversity metrics to use.

```
beta_diversity:metrics  bray_curtis,euclidean,unweighted_unifrac,weighted_unifrac,binary_jaccard,chord
alpha_diversity:metrics shannon,PD_whole_tree,chao1,observed_otus,simpson
assign_taxonomy:reference_seqs_fp   /media/sf_Y_DRIVE/Brian/Cattle/Necessary_Resources/Silva_132_release/SILVA_132_QIIME_release/rep_set/rep_set_16S_only/97/silva_132_97_16S.fna
assign_taxonomy:id_to_taxonomy_fp   /media/sf_Y_DRIVE/Brian/Cattle/Necessary_Resources/Silva_132_release/SILVA_132_QIIME_release/taxonomy/16S_only/97/taxonomy_7_levels.txt
```

```
assign_taxonomy:assignment_method   rdp
assign_taxonomy:confidence  0.80
assign_taxonomy:rdp_max_memory  20000
parallel:jobs_to_start  6
pick_otus:otu_picking_method    usearch
pick_otus:similarity    0.97
pick_otus:percent_id_err    0.97
pick_otus:word_length   64
pick_otus:db_filepath   /media/sf_Y_DRIVE/Brian/Cattle/Necessary_Resources/gold.fa
pick_otus:perc_id_blast 0.97
```

assign taxonomy to OTUs

```
assign_taxonomy.py -i refchi.fna -o assign_taxa --reference_seqs_fp '/media/sf_S/Brian Nohomovich/Cattle/Necessary_Resources/Silva_132_release/SILVA_132_QIIME_release/rep_set/rep_set_16S_only/97/silva_132_97_16S.fna' --id_to_taxonomy_fp '/media/sf_S/Brian Nohomovich/Cattle/Necessary_Resources/Silva_132_release/SILVA_132_QIIME_release/taxonomy/16S_only/97/taxonomy_7_levels.txt' --assignment_method rdp --confidence 0.80 --rdp_max_memory 20000
```

align seqs with PYNAST

```
align_seqs.py -i '/media/sf_S/Brian Nohomovich/Cattle/1/split_default_remove/test/refchi.fna'
```

Filter alignment

```
filter_alignment.py -i seqs_rep_set_aligned.fasta -o filtered_alignment/
```

Create tree

```
make_phylogeny.py -i refchi.fna -o rep_phylo.tre

alpha_rarefaction.py -i otus/otu_table.biom -m Fasting_Map.txt -o arare -p alpha_params.txt -t otus/rep_set.tre
```

beta_diversity_through_plots.py -i otus/otu_table.biom -m Fasting_Map.txt -o bdiv_even146 -t otus/rep_set.tre -e 146

Unoise algorithm performs poorly on 454, pacbio data

```
vsearch --cluster_unoise unique_seqs.fna --centroids denoised_seqs.fna -log denoise_log.txt;
```