**Homework 3.**

Kai Blumberg

ABE 587

In class, we went through analyzing 16S mock community data using usearch.

Based on what you learned, analyze your 16S data for your project and answer the following questions:

**a. What was the overall number reads in the file? (FYI these are 454 data so there only Forward reads)**

Running the following command:

## Get stats on reads merged using different parameters

./usearch -fastx\_info project1.fastq

I get that there are 72.7k seqs, which are the number of merged reads.

**b. What is the average EE value for the reads?**

The output from the previous command gives me an EE mean of 3.9.

**c. What is the average read length?**

Having run the command:

./usearch -fastq\_eestats2 project1.fastq -output project1.fastq.stats2 -length\_cutoffs 150,250,50 -ee\_cutoffs 0.5,1.0,2.0

I get 72728 reads with an average length of 554.3 bp.

**d. Using the following values for -fastq\_filter how many reads are retained after QC?**

* -fastq\_trunclen 400 -fastq\_maxee 1.0
* -fastq\_trunclen 300 -fastq\_maxee 1.0

Using a length of 400 and a max ee of 1, the command is:

./usearch -fastq\_filter project1.fastq -fastq\_trunclen 400 -fastq\_maxee 1 -fastqout project1-ee.1truc400.fq

I retain 54.8k reads.

Using a length of 300 and a max EE of 1.

I retain 65.7k reads.

**e. Using those same values how many OTU are recovered (after dereplicating, removing unique with <3 representatives)?**

* -fastq\_trunclen 400 -fastq\_maxee 1.0
* -fastq\_trunclen 300 -fastq\_maxee 1.0

Running the lines:

##Dereplicate sequences to find unique reads (removes unique clusters with <3 reads)

./usearch -fastx\_uniques project1-ee.1truc400.fq -fastaout project1-ee.1truc400.uniques.fa -minuniquesize 2 -sizeout -relabel Uniq

./usearch -fastx\_uniques project1-ee.1truc300.fq -fastaout project1-ee.1truc300.uniques.fa -minuniquesize 2 -sizeout -relabel Uniq

## cluster de-replicated sequences at 97% sequence identity

./usearch -cluster\_otus project1-ee.1truc400.uniques.fa -otus project1-ee.1truc400.uniques-rad3.fa -relabel OTU\_ -uparseout project1-ee.1truc400.uniques-rad3.txt

./usearch -cluster\_otus project1-ee.1truc300.uniques.fa -otus project1-ee.1truc300.uniques-rad3.fa -relabel OTU\_ -uparseout project1-ee.1truc300.uniques-rad3.txt

Using a length of 400 and a max EE of 1, I recover 164 OTUs.

Using a length of 300 and a max EE of 1, I recover 179 OTUs.

**f. Showing the results from the parameters you tried (at least one must differ from above), describe what are the best length truncation and EE cutoff values to be used for your data? Why?**

Running the quality control, removing unique clustering and finding unique OTU steps again with length 400 and max EE of 0.5:

Commands:

## Perform Quality control

##COMMAND QC merged reads using parameters chosen (400 or 300 bp length and max\_ee 1)

./usearch -fastq\_filter project1.fastq -fastq\_trunclen 300 -fastq\_maxee 0.5 -fastqout project1-ee.0.5truc300.fq

## Remove unique clusters

##Dereplicate sequences to find unique reads (removes unique clusters with <3 reads)

./usearch -fastx\_uniques project1-ee.0.5truc300.fq -fastaout project1-ee.0.5truc300.uniques.fa -minuniquesize 2 -sizeout -relabel Uniq

## Find unique OTUs.

## cluster de-replicated sequences at 97% sequence identity

./usearch -cluster\_otus project1-ee.0.5truc300.uniques.fa -otus project1-ee.0.5truc300.uniques-rad3.fa -relabel OTU\_ -uparseout project1-ee.0.5truc300.uniques-rad3.txt

I get 173 OTUs. Comparing this against our previous results of using length of 400 and a max EE of 1 and recovering 164 OTUs and using length of 300 and a max EE of 1 and recovering 179 OTUs.

Hence, I would say for this data that using a length cutoff of 300 was better than 400 in terms of recovering more OTUs. Additionally, using the EE of 0.5 which is of greater stringency than using 1, gives us 173 vs 179 OTU's recovered. This is only a reduction of 6 potential OTUs but using a much more conservative measure (EE value). Thus, I believe that for our data this would be the best way to proceed. I would rather be conservative and miss a couple real species in order to make sure I’m not including chimeras or other non-real artifacts.