**V-search/Usearch pipeline.**

**Reason for use**

Vsearch v2.16.0 and Usearch v11.0 on HPC were employed in this pipeline. Usearch and Vsearch were chosen for this project because they are quick and simple to set up and use in terms of searching. Usearch was created by Robert Edgar (2010), but it was not open-source, and the free version had significant constraints built in—so if you have a huge dataset, several operations will be unavailable until you pay for the higher-end version.( Torbjørn Rognes et al., 2016) built Vsearch as a labour of love and possibly other emotions in order to address the demands of the community that Usearch could not, while still remaining entirely open-source and freely available. The reason to adoption of the pipeline is Usearch for some commands can perform tasks that Vsearch cannot as well as Vsearch. This only applies to commands where the limits of Usearch or Vsearch, which is freely available, should not be a problem for any dataset, including the one in use for this case. Also, it's really usual to end up hybridizing many tools into a workflow in order to take use of what you enjoy about each, and this serves as a working example of a workflow to follow when working with Oxford Nanopore long reads for 16srRNA. Of course, while working with your own data, you should never mindlessly follow any pipeline.

**WORKFLOW.**

1. **Filtering and sorting rRNA Sequences with SortMeRNA**

SortMeRNA is a program that allows you to filter, map, and pick OTUs from NGS readings in Metatranscriptomics and metagenomic data. The basic technique uses estimated seeds to enable for quick and sensitive nucleotide sequence analysis. SortMeRNA's main function is to filter ribosomal RNA from Metatranscriptomics data.. (Martinez et al., 2016).This tool was used to filter the rRNA. The command below was used:

*for file in ./\*.fq;*

*do*

*name=$(basename ${file} .fq)*

*sortmerna --ref $silva-bac-16s-database-id85.fasta*

*--reads file prefix$file \*

*--aligned file prefix$name*

*--other file prefix$name \*

*--fastx $name --threads 9*

*rm -rfv ./sortmerna/run/kvdb*

*done*

1. **Renaming identifier lines from the files.**

Awk is a scripting language for data manipulation and report generation. Variables, numeric functions, string functions, and logical operators are all available in the awk command programming language, which does not require compilation. (Weinberger et al., 1987) .The command below was used to rename the lines;

*awk '{if( (NR-1)%4 ) print; else printf("@CF3-%d\n",cnt++)}' $file > $new.fq*

1. **Merging the files.**

The command below was used:

*cat \*.fq > $name.fq*

1. **Quality filtering.**

Vsearch command was used to perform this, as well as a quality filter based on the maximum predicted error. Filtering for quality is crucial in lowering the number of false sequences and their impact. All sequencing technologies (including polymerases) have an inherent error rate that will consistently produce some sequences that differ from their genuine biological origin, which can significantly inflate metrics like richness and diversity. One of the techniques taken to mitigate the problem is quality filtering.( Torbjrn Rognes et al., 2016)

The command below was used:

*vsearch -fastq\_filter $file.fq*

*--fastq\_stripleft 0 \*

*--fastq\_stripright 0 \*

*--fastq\_maxee 600 \*

*--fastaout QCd\_$file.fa*

1. **Dereplication.**

The dereplication stage reduces all identical sequences to one and just counts how many there were in the first place. This can save a significant amount of time in subsequent processing processes. The Vsearch -derep fulllength command was used to perform the dereplication step:

*vsearch --derep\_fulllength QCd\_$file.fa -sizeout -relabel Uniq -output unique\_seqs.fa*

1. **Generating ASVs**

ASVs stand for "amplicon sequence variants." These are created using single-nucleotide resolution methods, which, as the name implies, allows you to distinguish between sequences that differ by only one base pair. This usually entails similarity values of greater than 99 percent, resulting in separate units that would otherwise be destroyed if OTU clustering. True biological sequences are represented by ASVs.

The Usearch –unoise3 command was used:

*usearch -unoise3 unique\_seqs.fa -zotus ASVs.fa -minsize 2*

(Bokulich et al.,2013) recommend using a 0.005% minimum abundance level, which in this case the instance come out to ~2, therefore it was specified. This command also got rid of any chimeric sequences.

1. **Generating count table.**

When mapping the merged sequences to the ASVs to generate the count table Vsearch’s –usearch \_global command was used to accomplish this at a threshold of 85%:

First,the Zotu were renamed to ASVs using this command:

*sed -i.tmp 's/Zotu/ASV\_/' ASVs.fa*

*Then this command was used to generate the count table:*

*vsearch -usearch\_global QCd\_$file.fa --db ASVs.fa --id 0.85\*

*--otutabout ASV\_counts.txt*

1. **Assigning taxonomy.**

The final step in this sequence processing was to assign taxonomy to the ASV sequences. Here Usearch’s -Sintax program was used with the RDP training set reference fasta.

*usearch -sintax ASVs.fa -db $rdp\_16s\_v18.fa -tabbedout ASV\_tax\_raw.txt -strand both -sintax\_cutoff 0.1*

*#this script is used to convert the output by removing the digit values.*

*bash convert\_usearch\_tax.sh*

*#!/bin/bash*

*cut -f1 ASV\_tax\_raw.txt > ASV\_IDs*

*cut -f4 ASV\_tax\_raw.txt > raw\_tax*

*cut -d : -f2 raw\_tax | cut -d , -f1 | sed 's/\"//g' > domain*

*cut -d : -f3 raw\_tax | cut -d , -f1 | sed 's/\"//g' > phylum*

*cut -d : -f4 raw\_tax | cut -d , -f1 | sed 's/\"//g' > class*

*cut -d : -f5 raw\_tax | cut -d , -f1 | sed 's/\"//g' > order*

*cut -d : -f6 raw\_tax | cut -d , -f1 | sed 's/\"//g' > family*

*cut -d : -f7 raw\_tax | cut -d , -f1 | sed 's/\"//g' > genus*

*cut -d : -f8 raw\_tax | cut -d , -f1 | sed 's/\"//g' > species*

*paste ASV\_IDs domain phylum class order family genus species > tax\_temp*

*echo -e "\tDomain\tPhylum\tClass\tOrder\tFamily\tGenus\tSpecies" > header*

*cat header tax\_temp > ASV\_tax.txt*

*rm ASV\_IDs raw\_tax domain phylum class order family genus species tax\_temp header*