#running on group 7 server: 10.32.204.57 Biome1874

#upload UniProt files to server

scp uniprot\_NorB.fasta.gz 425:/data/tn/NorB

scp uniprot\_NorB.tab.gz 425:/data/tn/NorB

#upload NorB refpkg to server

scp -r final\_outputs/ 425:/data/tn/NorB

#unzip files in working directory (/data/tn/NorB)

gunzip uniprot\_NorB.fasta.gz

gunzip uniprot\_NorB.tab.gz

#Classifying amino acid sequences

treesapp assign \

-n 4 \

--trim\_align \

--refpkg\_dir final\_outputs/ \

--fastx\_input uniprot\_NorB.fasta \

--output UniProt\_NorB\_assign/

#Create seqs\_to\_lineage files for update

echo -e "SeqID\tOrganism\tDomain\tPhylum\tClass\tOrder\tFamily\tGenus\tSpecies" >UniProt\_NorB.tsv

tail -q -n +2 uniprot\_NorB.tab >>UniProt\_NorB.tsv

#Update the refpkg with UniProt NorB sequences

treesapp update \

--fast \

--headless \

--overwrite \

--delete \

--cluster \

--trim\_align \

--min\_taxonomic\_rank o \

-n 4 \

--output NorB\_UniProt\_update/ \

--skip\_assign \

--treesapp\_output UniProt\_NorB\_assign/ \

--refpkg\_path final\_outputs/NorB\_build.pkl \

--seqs2lineage UniProt\_NorB.tsv

#Classifying ORFs predicted from genomes (SAGs) without --trim\_align

treesapp assign \

-n 4 \

--refpkg\_dir NorB\_UniProt\_update/final\_outputs/ \

--fastx\_input /mnt/datasets/2021w/saanich/SAGs/Concatenated\_Med\_Plus/Saanich\_Med\_Plus\_SAGs.fasta \

--output SI072\_SAGs\_assign/

#Update refpkg with SAG sequences

treesapp update \

--fast \

--headless \

--overwrite \

--delete \

--cluster \

-n 4 \

--output NorB\_SAG\_update/ \

--skip\_assign \

--seqs2lineage /mnt/datasets/2021w/saanich/seq2lineage\_Tables/Med\_Plus\_SAGs\_GTDB\_Taxonomies.tsv \

--treesapp\_output SI072\_SAGs\_assign/ \

--refpkg\_path NorB\_UniProt\_update/final\_outputs/NorB\_build.pkl

#Output after running treesapp update for SAGs with --trim\_align:

# Number of sequences: 88

# Longest sequence length: 480

# Shortest sequence length: 44

# Mean sequence length: 285.8

# Median sequence length: 275.5

#ERROR - commands, line 770:

**#No classified sequences exceed minimum length threshold of 486.**

#None of the SAG sequences pass the minimum quality control threshold sequence length of 486. Therefore the reference package cannot be updated with SAGs from saanich inlet (for NorB)

#Tried again removing --trim\_align and same result:

# Number of sequences: 88

# Longest sequence length: 478

# Shortest sequence length: 45

# Mean sequence length: 286.7

# Median sequence length: 276.0

#ERROR - commands, line 770:

#No classified sequences exceed minimum length threshold of 489.

#Classifying ORFs predicted from genomes (MAGs) without --trim\_align

treesapp assign \

-n 4 \

--refpkg\_dir NorB\_UniProt\_update/final\_outputs/ \

--fastx\_input /mnt/datasets/2021w/saanich/ MAGs/Concatenated/All\_SI072\_Metawrap\_MAGs.fa \

--output SI072\_MAGs\_assign/

#Update refpkg with MAG sequences without --trim\_align

treesapp update \

--fast \

--headless \

--overwrite \

--delete \

--cluster \

-n 4 \

--output NorB\_MAG\_update/ \

--skip\_assign \

--seqs2lineage /mnt/datasets/2021w/saanich/seq2lineage\_Tables/SI072\_MAGs\_All\_GTDB\_taxonomies.tsv \

--treesapp\_output SI072\_MAGs\_assign/ \

--refpkg\_path NorB\_UniProt\_update/final\_outputs/NorB\_build.pkl

#Same error output after running treesapp update for MAGs with and without --trim\_align.

Number of sequences: 23

Longest sequence length: 471

Shortest sequence length: 45

Mean sequence length: 312.0

Median sequence length: 456

ERROR - commands, line 770:

No classified sequences exceed minimum length threshold of 489.

#Run command to look at MAGs classification.tsv

less SI072\_MAGs\_update\_assign/final\_outputs/classifications.tsv

#Shows that there are 11 sequences with length over 450 (455-472). These sequences are likely NorB and show up in c\_\_Epsilonproteobacteria; o\_\_Campylobacterales and c\_\_Gammaproteobacteria

#With Connor, tried to change the treesapp command to lower the threshold to 450 length, however, there is a technical problem and treesapp was unable to process the command. The programmer of treesapp (Connor) was notified so that the issue can be resolved from the developer standpoint.

#Assign taxonomic labels to Saanich Inlet metagenomic contigs using the NorB reference #package

for f in /mnt/datasets/2021w/saanich/MetaG\_Assemblies/SI072\_\*m\_contig.fa; do sample=$( basename $f | sed 's/.fa//g'); treesapp assign -i $f --refpkg\_dir NorB\_UniProt\_update/final\_outputs/ --output SI072\_MetaG\_contigs\_NorB\_assign/${sample}\_assign --trim\_align -n 8; done

# overwrite the previous abundance values with the TPM values calculated from the seven #SI072 datasets

for f in SI072\_MetaG\_contigs\_NorB\_assign/SI072\_\*assign; do sample=$( basename $f | sed 's/\_contig\_assign//g'); treesapp abundance --treesapp\_output $f --reads /mnt/datasets/2021w/saanich/MetaG\_Trim\_QC\_Reads/${sample}\_pe.1.fq.gz --reverse /mnt/datasets/2021w/saanich/MetaG\_Trim\_QC\_Reads/${sample}\_pe.2.fq.gz -n 8 --report update; done

#calculate TPM values from the seven metatranscriptomes

for f in SI072\_MetaG\_contigs\_NorB\_assign/SI072\_\*assign; do sample=$( basename $f | sed 's/\_contig\_assign//g'); treesapp abundance --treesapp\_output $f --reads /mnt/datasets/2021w/saanich/MetaT\_Raw\_Reads/${sample}\_MetaT\_QC\_Filtered.fastq.gz -n 8 --pairing pe --metric tpm --report append; done

#annotate the NorB query sequences with their respective paralog at each depth

treesapp layer \

-o SI072\_MetaG\_contigs\_NorB\_assign/SI072\_100m\_contig\_assign \

--refpkg\_dir NorB\_UniProt\_update/final\_outputs/

treesapp layer \

-o SI072\_MetaG\_contigs\_NorB\_assign/SI072\_10m\_contig\_assign \

--refpkg\_dir NorB\_UniProt\_update/final\_outputs/

treesapp layer \

-o SI072\_MetaG\_contigs\_NorB\_assign/SI072\_120m\_contig\_assign \

--refpkg\_dir NorB\_UniProt\_update/final\_outputs/

treesapp layer \

-o SI072\_MetaG\_contigs\_NorB\_assign/SI072\_135m\_contig\_assign \

--refpkg\_dir NorB\_UniProt\_update/final\_outputs/

treesapp layer \

-o SI072\_MetaG\_contigs\_NorB\_assign/SI072\_150m\_contig\_assign \

--refpkg\_dir NorB\_UniProt\_update/final\_outputs/

treesapp layer \

-o SI072\_MetaG\_contigs\_NorB\_assign/SI072\_165m\_contig\_assign \

--refpkg\_dir NorB\_UniProt\_update/final\_outputs/

treesapp layer \

-o SI072\_MetaG\_contigs\_NorB\_assign/SI072\_200m\_contig\_assign \

--refpkg\_dir NorB\_UniProt\_update/final\_outputs/

#combine layered\_classification tables into one file

cat /data/tn/NorB/SI072\_MetaG\_contigs\_NorB\_assign/SI072\_\*m\_contig\_assign/final\_outputs/layered\_classifications.tsv | head -n 1 >SI072\_NorB\_layered\_classifications.tsv

tail -q -n +2 /data/tn/NorB/SI072\_MetaG\_contigs\_NorB\_assign/SI072\_\*m\_contig\_assign/final\_outputs/layered\_classifications.tsv >>SI072\_NorB\_layered\_classifications.tsv