**# Absolute path of NirK reference package (Immanuel updated from UniProt):**

/root/data/imman/uniprot\_data/NirK\_UniProt\_update/final\_outputs/NirK\_build.pkl

**# Create a new directory for UniProt-updated NirK processing ✅**

cd /root/data/kristi/

mkdir NirK\_processed\_revised

cd NirK\_processed\_revised

**# Copy UniProt\_NirK.fasta file to the new directory ✅**

cp /root/data/kristi/NirK\_processed/UniProt\_NirK.fasta /root/data/kristi/NirK\_processed\_revised/

**# Classifying amino acid sequences ✅**

cd /root/data/kristi/NirK\_processed\_revised/

treesapp assign \

  -n 4 \

  --trim\_align \

  --refpkg\_dir /root/data/imman/uniprot\_data/NirK\_UniProt\_update/final\_outputs/ \

  --fastx\_input UniProt\_NirK.fasta \

  --output /root/data/kristi/NirK\_processed\_revised/UniProt\_NirK\_assign/

**# Download the data package SI072\_sequence\_data.tar.gz from Zenodo and decompress it with tar ✅**

wget https://zenodo.org/record/6323402/files/SI072\_sequence\_data.tar.gz && \

tar -xzvf SI072\_sequence\_data.tar.gz

**# Classify ORFs predicted from genomes (SAGs). ✅**

treesapp assign \

  -n 4 \

  --trim\_align \

  --refpkg\_dir /root/data/imman/uniprot\_data/NirK\_UniProt\_update/final\_outputs/ \

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

  --output SI072\_SAGs\_assign/

# OUTPUT:

#[2022-04-14 21:53:08] DEBUG: Initial alignments: 162

Alignments discarded: 61

Fragmented alignments: 70

Inversions detected: 25

Alignments scaffolded: 10

Multi-alignments: 0

Sequences identified: 66

Number of markers identified:

NirK 66

# [2022-04-14 21:53:09] DEBUG: Number of query sequences in each marker's group:

NirK 0 58

NirK 1 4

NirK 2 3

NirK 3 1

**# Skip updating publicly available sequences**

**# Update NirK refpkg with SAG sequences ✅**

treesapp update \

  --fast \

  --headless \

  --overwrite \

  --delete \

  --cluster \

  --trim\_align \

  -n 4 \

  --output NirK\_SAG\_update/ \

  --skip\_assign \

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

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

  --refpkg\_path /root/data/imman/uniprot\_data/NirK\_UniProt\_update/final\_outputs/NirK\_build.pkl

# Output:

Sequence summary:

Number of sequences: 942

Longest sequence length: 568

Shortest sequence length: 264

Mean sequence length: 410.3

Median sequence length: 383.0

Extracting information from headers... done.

Reading cached lineages in '/root/data/kristi/NirK\_processed\_revised/NirK\_SAG\_update/intermediates/accession\_id\_lineage\_map.tsv'... done.

Clustering sequences with MMSeqs' Linclust... done.

Number of unique lineages:

root 1

domain 3

phylum 22

class 38

order 80

family 142

genus 330

species 634

Unclassified and incomplete lineages account for 157/905 (17.3%) references.

**# Classify ORFs predicted from genomes (MAGs) ✅**

treesapp assign \

-n 4 \

--trim\_align \

--refpkg\_dir /root/data/kristi/NirK\_processed\_revised/NirK\_SAG\_update/final\_outputs/ \

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

--output SI072\_MAGs\_assign/

**# Update the NirK reference package with MAG sequences ✅**

screen -S NirK\_refpkg\_MAG\_update

treesapp update \

--fast \

--headless \

--overwrite \

--delete \

--cluster \

--trim\_align \

-n 4 \

--output NirK\_MAG\_update/ \

--skip\_assign \

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

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

--refpkg\_path /root/data/kristi/NirK\_processed\_revised/NirK\_SAG\_update/final\_outputs/NirK\_build.pkl

# Output:

[2022-04-16 22:40:21] DEBUG: 0 classified sequences did not meet minimum LWR of

0.1 for updating

[2022-04-16 22:40:21] INFO: Number of sequences: 75

Longest sequence length: 508

Shortest sequence length: 78

Mean sequence length: 343.1

Median sequence length: 366

[2022-04-16 22:40:22] DEBUG: Identified and replaced invalid ambiguity character

s in 0 sequences.

[2022-04-16 22:40:22] INFO: Sequence summary:

Number of sequences: 965

Longest sequence length: 568

Shortest sequence length: 264

Mean sequence length: 410.6

Median sequence length: 383

[2022-04-16 22:40:29] INFO: Number of unique lineages:

root 1

domain 3

phylum 24

class 38

order 84

family 152

genus 342

species 636

**# Assign taxonomic labels to Saanich Inlet metagenomic contigs using the NirK 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 \

-n 8 \

--trim\_align \

--refpkg\_dir /root/data/kristi/NirK\_processed\_revised/NirK\_MAG\_update/final\_outputs/ \

-i $f \

--output SI072\_MetaG\_contigs\_NirK\_assign/${sample}\_assign; done

**# Overwrite previous abundance values with transcripts per million (TPM) values calculated from SI072 datasets ✅**

screen -S NirK\_abundance

for f in SI072\_MetaG\_contigs\_NirK\_assign/SI072\_\*assign; \

do sample=$( basename $f | sed 's/\_contig\_assign//g')

treesapp abundance \

-n 8 \

--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 \

--report update \

--metric tpm; done

# Output files in SI072\_MetaG\_contigs\_NirK\_assign:

**SI072\_100m\_contig\_assign** **SI072\_135m\_contig\_assign** **SI072\_200m\_contig\_assign**

**SI072\_10m\_contig\_assign** **SI072\_150m\_contig\_assign**

**SI072\_120m\_contig\_assign** **SI072\_165m\_contig\_assign**

**# Calculate TPM values for the 7 metatranscriptomes ✅**

screen -S NirK\_TPM\_7

for f in SI072\_MetaG\_contigs\_NirK\_assign/SI072\_\*assign; \

do sample=$( basename $f | sed 's/\_contig\_assign//g')

treesapp abundance \

-n 8 \

--treesapp\_output $f \

--reads /mnt/datasets/2021w/saanich/MetaT\_Raw\_Reads/${sample}\_MetaT\_QC\_Filtered.fastq.gz \

--pairing pe \

--metric tpm \

--report append; done

**# Annotate NirK query sequences with their respective paralog at all 7 depths ✅**

treesapp layer \

-o /root/data/kristi/NirK\_processed\_revised/SI072\_MetaG\_contigs\_NirK\_assign/SI072\_10m\_contig\_assign \

--refpkg\_dir /root/data/kristi/NirK\_processed\_revised/NirK\_MAG\_update/final\_outputs/

treesapp layer \

-o /root/data/kristi/NirK\_processed\_revised/SI072\_MetaG\_contigs\_NirK\_assign/SI072\_100m\_contig\_assign \

--refpkg\_dir /root/data/kristi/NirK\_processed\_revised/NirK\_MAG\_update/final\_outputs/

treesapp layer \

-o /root/data/kristi/NirK\_processed\_revised/SI072\_MetaG\_contigs\_NirK\_assign/SI072\_120m\_contig\_assign \

--refpkg\_dir /root/data/kristi/NirK\_processed\_revised/NirK\_MAG\_update/final\_outputs/

treesapp layer \

-o /root/data/kristi/NirK\_processed\_revised/SI072\_MetaG\_contigs\_NirK\_assign/SI072\_135m\_contig\_assign \

--refpkg\_dir /root/data/kristi/NirK\_processed\_revised/NirK\_MAG\_update/final\_outputs/

treesapp layer \

-o /root/data/kristi/NirK\_processed\_revised/SI072\_MetaG\_contigs\_NirK\_assign/SI072\_150m\_contig\_assign \

--refpkg\_dir /root/data/kristi/NirK\_processed\_revised/NirK\_MAG\_update/final\_outputs/

treesapp layer \

-o /root/data/kristi/NirK\_processed\_revised/SI072\_MetaG\_contigs\_NirK\_assign/SI072\_165m\_contig\_assign \

--refpkg\_dir /root/data/kristi/NirK\_processed\_revised/NirK\_MAG\_update/final\_outputs/

treesapp layer \

-o /root/data/kristi/NirK\_processed\_revised/SI072\_MetaG\_contigs\_NirK\_assign/SI072\_200m\_contig\_assign \

--refpkg\_dir /root/data/kristi/NirK\_processed\_revised/NirK\_MAG\_update/final\_outputs/

**# Concatenate the 7 classification files ✅**

cat /data/kristi/NirK\_processed\_revised/SI072\_MetaG\_contigs\_NirK\_assign/SI072\_\*m\_contig\_assign/final\_outputs/layered\_classifications.tsv | head -n 1 >SI072\_NirK\_layered\_classifications.tsv

tail -q -n +2 /data/kristi/NirK\_processed\_revised/SI072\_MetaG\_contigs\_NirK\_assign/SI072\_\*m\_contig\_assign/final\_outputs/layered\_classifications.tsv >>SI072\_NirK\_layered\_classifications.tsv

**# Copy SI072\_NirK\_layered\_classifications.tsv to local computer ✅**

scp root@10.32.204.58:/root/data/kristi/NirK\_processed\_revised/SI072\_NirK\_layered\_classifications.tsv <local file path>