**METAGENOMICS WORKFLOW**

1. **Read-based analysis**
   1. **Microbial Taxa (Kraken2)**

module load kraken/2.0.8-beta

kraken2 --db <path\_to\_database> --confidence 0.1 --paired --threads $threads --output $output.txt --report $report.txt --gzip-compressed $read\_1\_file.fastq.gz $read\_2\_file.fastq.gz

* 1. **Papillomavirus (HPViewer)**

module load bowtie/2-2.4.5

module load samtools/1.16.1

module load bedtools/2.30.0

python HPViewer.py -p $threads -1 $read\_1\_file.fastq.gz -2 $read\_2\_file.fastq.gz -o $output

* 1. **Antimicrobial Resistance Genes (eautils/shortbred)**

# join read files

module load ea-utils/bd148d4

fastq-join $read\_1\_file.fq.gz $read\_2\_file.fq.gz -o $joined\_reads.fq.gz

# shortbred quantify (using markers prepared with ‘shortbred identify’ for CARD database with Uniprot90 as reference)

# ran with singularity container (shortbred\_quantify.py v0.9.5; ‘SB\_CARD\_markers\_u90.faa’ is located in ‘/supporting\_files’)

shortbred\_quantify.py --threads $threads --markers SB\_CARD\_markers\_u90.faa –wgs $joined\_reads.fastq.gz --results $output.txt --tmp $tmpdir

1. **Metagenomic assembly**
   1. **Assemble reads into contigs (Spades)**

module load spades/3.15.5

spades.py -1 $read\_1\_file.fq.gz -2 $read\_1\_file.fq.gz -o $output\_dir --meta -t $threads -m 1000 #(-m is optional; larger memory needed for some co-assemblies)

* 1. **Binning (Metawrap)**

# ran with singularity container (MetaWRAP v=1.2.2)

metawrap binning --metabat2 --maxbin2 --concoct -m 180 -l 2500 -t $threads -a $scaffolds.fasta -o $bins\_dir $read\_1\_file.fastq $read\_2\_file.fastq

1. **Prokaryote MAG recovery**
   1. **Bin refinement (MetaWRAP)**

# ran with singularity container (MetaWRAP v=1.2.2)

metawrap bin\_refinement -t $threads -o $bins\_dir/bins\_refined\_dir -A $bins\_dir/metabat2\_bins/ -B $bins\_dir/maxbin2\_bins/ -C $bins\_dir/concoct\_bins/ -c 50 -x 10

* 1. **Dereplication (dRep)**

# ran with singularity container (dRep v2.6.2)

dRep dereplicate -p $threads $bins\_refined\_dir/drep\_dir -g $bins\_refined\_dir/\*fa -pa 0.9 -sa 0.95 -nc 0.30 -cm larger -comp 50 -con 5

* 1. **Taxonomic assignment (GTDB-Tk)**

# ran with conda environment (GTDB-Tk v1.3.0)

source $USER\_path/conda/etc/profile.d/conda.sh

conda activate gtdbtk

gtdbtk classify\_wf --cpus $threads --genome\_dir $drep\_dir/dereplicated\_genomes --out\_dir $gtdbtk\_dir -x fa

* 1. **Phylogenetic tree (FastTree2)**

module load FastTree/2.1.10

FastTree < $gtdbtk\_dir/align/gtdbtk.bac120.user\_msa.fasta > $output\_tree.newick

1. **Eukaryote MAG recovery**
   1. **Quality assessment (EukCC)**

# ran with singularity container (EukCC version 2.1.1)

eukcc folder --out $euk\_output\_dir --db <path\_to\_downloaded\_database>/eukccdb/eukcc2\_db\_ver\_1.1 --threads $threads $concoct\_bins\_dir # (CONCOCT bins from MetaWRAP are the input)

* 1. **Dereplication (dRep)**

# ran with singularity container (dRep v2.6.2)

dRep dereplicate -p $threads $euk\_output\_dir/euk\_drep\_dir -g $euk\_output\_dir/\*fa -pa 0.95 -ms 10000 --genomeInfo $quality\_info.csv -comp 50 -con 5 # (quality info manually created from the eukcc outputs, sub-setting eukaryote genomes based on completeness/contamination)

* 1. **Taxonomic assignment (mash/mummer)**

module load mash/2.3

mash dist -p $threads $mash\_sketch\_genbank\_fungi.msh $bin > $mash\_out

sort -gk3 $mash\_out | sed -n 1p > $mash\_sort

module load mummer/4.0.0beta2

dnadiff $reference\_genome.fa $euk\_bin.fa -p $euk\_bin\_ANI

1. **Viral genome recovery**
   1. **Identify sequences from assembled contigs (seqkit, VirSorter2, VirFinder)**

# filter assembled contigs to minimum 500bp

module load seqkit/2.1.0

seqkit seq -m 5000 $scaffolds.fasta -o $scaffolds.filt.fasta

# ran virsorter with conda environment (Virsorter2 v2.1)

source $USER\_path/conda/etc/profile.d/conda.sh

conda activate vs2

virsorter run -w $vs2\_out\_dir -i $scaffolds.filt.fasta --hallmark-required --include-groups dsDNAphage,ssDNA,lavidaviridae -j $threads

# ran virfinder with conda environment and R-script (‘/virfinder\_R\_steps’ is located in ‘/supporting\_files’)

source $USER\_path/conda/etc/profile.d/conda.sh

conda activate base

cd <path\_to\_virfinder>/vf\_out

mkdir $scaffolds.filt

cd $scaffolds.filt

cp <path\_to\_virfinder\_R\_steps>/vf\_script\_model.R .

cp <path\_to\_virfinder\_R\_steps>/VF.modEPV\_k8.rda .

cp <path\_to\_assemblies>/$scaffolds.filt.fasta .

mv $scaffolds.filt.fasta filt\_scaffolds.fasta

Rscript vf\_script\_model.R

mv table.txt $scaffolds.filt.txt

mv table\_qval.txt $scaffolds.filt\_qval.txt

# subset contigs with q<0.05

# merge viral contigs identified with virsorter and virfinder using cat

* 1. **Quality assessment (CheckV)**

# ran with conda environment (CheckV v0.7.0)

export CHECKVDB=,path\_to\_database>/checkv-db-v0.6

checkv end\_to\_end $virus\_contigs.fasta $checkv\_out\_dir -t $threads

* 1. **Cluster viral sequences (CD-HIT)**

module load cd-hit/4.6.8

cd-hit-est -i $virus\_contigs.fasta -o $virus\_contigs.clust.fasta -c 0.90 -aS 0.75 -G 0 -T 0 -M 50000

* 1. **Taxonomic assignment (Demovir)**

# ran with conda environment (CheckV v0.7.0)

source /data/$USER/conda/etc/profile.d/conda.sh

conda activate base

export PATH="$HOME/<path\_to\_demovir\_dir>/Demovir/usearch11.0.667\_i86linux32:$PATH"

module load prodigal/2.6.3

sh demovir.sh $virus\_contigs.clust.fasta $threads

* 1. **Papillomavirus analysis**
     1. **Translate sequences (Prodigal)**

module load prodigal/2.6.3

prodigal -i $hpv\_virus\_contigs.fa -o $hpv\_genes.fa -a $hpv\_proteins.faa -p meta

* + 1. **Extract L1 sequence (muscle/hmmer)**

# build hmm for L1 sequences from the PAVE database

module load muscle/5.0.1428

muscle -align $PAVE\_L1.faa -output $PAVE\_L1.aln

module load hmmer/3.3.2

hmmbuild $PAVE\_L1.hmm $PAVE\_L1.aln

# scan all contigs

hmmsearch --cpu $threads --tblout $L1\_proteins\_out --noali $PAVE\_L1.hmm $hpv\_proteins.faa

* + 1. **Align L1 of identified HPVs with those from PAVE (muscle)**

# subset L1 protein(s) from each hpv contig based on table output

# merge L1 proteins with those from PAVE

module load muscle/5.0.1428

muscle -align $hpv\_L1\_with\_PAVE\_L1.faa -output $L1\_aln.afa

* + 1. **Phylogenetic tree (FastTree2)**

module load FastTree/2.1.10

FastTree < $l1\_aln.afa > $output\_tree.newick

1. **Read mapping**
   1. **Genome collection**

# create directory with prokaryote MAGs, eukaryote MAGs, and viral contigs (note: final collection of MAGs is dereplicated from SMGC, UHGG, and ELGG, and viral contigs are # clustered from SMGC)

# rename headers (python script from <https://github.com/Finn-Lab/MGS-gut/tree/master/scripts>)

python rename\_multifasta\_prefix.py -f $genome.fa -p $genome.fa > header/$genome.fa

# merge all ‘genomes’ with adjusted headers using cat to generate final $genome\_collection.fasta

# build index for collection

module load bwa/0.7.17

bwa index $genome\_collection.fasta

* 1. **Map reads (*SEE SEPARATE SCRIPT ‘map\_metagenome\_reads.sh’ located in ‘/supporting \_files’*)**
  2. **Stats (samtools)**

module load samtools/1.16.1

samtools flagstat $mapped\_file.bam -@ $threads > $output.txt