Bioinformatics Pipeline

Yisha Tan[*1,2,3]

[1]**Archaeal Biology Centre, Synthetic Biology Research Center, Shenzhen Key Laboratory of Marine Microbiome Engineering, Key Laboratory of Marine Microbiome Engineering of Guangdong Higher Education Institutes, Institute for Advanced Study, Shenzhen University, Shenzhen 518060, China**
[2]**HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen, China**
[3]**Department of Ocean Science, The Hong Kong University of Science and Technology, Hong Kong, SAR China**

Introduction

This is the bioinformatics pipeline used for the manuscript titled: *"Genetic diversity of cyanobacteria and their viruses reveals features of coevolution in cultured and natural environments."*

We focus on analyzing metagenomic data to detect and characterize cyanobacteria and associated cyanophages, especially in estuarine environments. The major steps are:

1. Metagenomic assembly, gene prediction, binning, and taxonomy.
2. Viral prediction, clustering, taxonomic/functional annotation.
3. Whole-genome analyses such as mapping, variant calling, and phylogenetics.

Below is a detailed explanation of each step, sample commands, and usage notes.

1 Metagenomic Assembly and Genome Binning

1.1 Trimming Sequences

```
sickle pe \
    -f raw_data1.fastq \
    -r raw_data2.fastq \
    -t sanger \
    -q 25 \
    -o raw_data1.trim.fastq \
    -p raw_data2.trim.fastq
```

*ytanap@connect.ust.hk

1.2 Assembly

```
~/SeqTools/interleave.pl \
    -fwd raw_data1.trim.fastq \
    -rev raw_data2.trim.fastq \
    -o total.trim.fa

idba_ud \
    --mink 65 \
    --maxk 145 \
    --step 10 \
    -r total.trim.fa \
    -o k65-145
```

1.3 Gene Prediction

```
prodigal \
    -f gff \
    -i sample.fa \
    -o sample.gff \
    -a prot.faa \
    -p meta

python2 emapper.py \
    -i prot.faa \
    -o prot.txt \
    -m diamond \
    --cpu 20
```

1.4 Mapping and Binning

```
# Build index
~/bowtie2-2.4.2/bowtie2-build sample.fa sample.fa

# Map reads
~/bowtie2-2.4.2/bowtie2 \
    -x sample.fa \
    -1 raw_data1.trim.fastq \
    -2 raw_data2.trim.fastq \
    -S sample.sam

# Convert & sort
samtools view -bS sample.sam > sample.bam
samtools sort sample.bam -o sample_sort.bam
samtools index sample_sort.bam
```

```
samtools flagstat sample_sort.bam > sample_flag.txt

# Coverage
~/jgi_summarize_bam_contig_depths --outputDepth sample_depth.txt
   sample_sort.bam

# Binning
metabat \
    -i sample.fa \
    -a sample_depth.txt \
    -o Binning
```

1.5 Quality Assessment

```
checkm lineage_wf \
    -x fa \
    -t 32 \
    checkm_result
```

1.6 Taxonomy

```
gtdbtk classify_wf \
    --genome_dir genomes \
    --out_dir output \
    --cpus 10

blastn \
    -db ~/SILVA/SILVA_132_SSURef_Nr99_tax_silva.fasta \
    -query bin.fasta \
    -out blastn_silva_result.txt \
    -evalue 1e-5 \
    -outfmt 6 \
    -max_target_seqs 1 \
    -num_threads 32
```

1.7 Abundance

```
coverm contig \
    -c raw_data1.trim.fastq raw_data2.trim.fastq \
    --reference cyano_silva138.fasta \
    -m rpkm \
    --output-file genes.rpkm.txt \
    -t 60
```

2 Viral Prediction and Analysis

2.1 Recovery of Viral Contigs

```
virsorter run \
    -w vs2 \
    -j 60 \
    --min-length 1500 \
    -i sample.fa \
    -d vs2_databse all

python dvf.py \
    -i sample.vir.fa \
    -l 1000 \
    -c 2

checkv end_to_end \
    sample.vir.fna \
    output_checkv \
    -t 16
```

2.2 Cyanophage Genomes and Clustering

```
cdhit \
    -i sample.vir.faa \
    -o output_cdhit \
    -c 0.95 \
    -aS 0.85 \
    -n 9 \
    -d 0 \
    -T 8
```

2.3 Viral Taxonomy and Genome Clustering

```
python ~/vRhyme/aux/coverage_table_convert.py \
    -i sample.vir.rpkm.txt \
    -o vir.coverage.tsv

vRhyme -i sample.vir.fasta -c vir.coverage.tsv

vcontact2_gene2genome \
    --proteins all_cluster.vir.faa \
    --output all_cluster.vir.g2g.csv \
    --source-type Prodigal-FAA

```

```
vcontact \
    --rel-mode Diamond \
    --pcs-mode MCL \
    --vcs-mode ClusterONE \
    --c1-bin ~/Tools/ \
    --db cyanophage_db \
    --verbose \
    --threads 30 \
    --raw-proteins all_cluster.vir.faa \
    --proteins-fp all_cluster.vir.g2g.csv \
    --output-dir result
```

2.4 Gene Prediction and Function

```
prodigal \
    -p meta \
    -a protein_seq.fasta \
    -m \
    -d nucleotide_seq.fasta \
    -o genome.gff \
    -f gff \
    -s stat \
    -i genome.fasta

python emapper.py \
    -i genome.faa \
    -o genome.txt \
    -m diamond \
    --cpu 20
```

2.5 Viral Abundance Mapping

```
makeblastdb \
    -in genome.faa \
    -dbtype prot

blastp \
    -query bin.faa \
    -db genome.faa \
    -out genome.txt \
    -outfmt 6 \
    -evalue 1e-3 \
    -max_target_seqs 1 \
    -num_threads 30

```

```
coverm contig \
    -c raw_data1.trim.fastq -2 raw_data2.trim.fastq \
    --reference vir.genome.fa \
    -m rpkm \
    --output-file genome.rpkm.txt \
    -t 60
```

2.6 ANI/AAI

```
comparem aai_wf \
    --cpus 40 \
    vir.fasta \
    ANI
```

2.7 Phylogenetic Trees

```
trimal \
    -in combined_genome.fasta \
    -out combined_genome.trim.fasta \
    -gt 0.95

muscle \
    -in combined_genome.trim.fasta \
    -out combined_genome.trim.aln.fasta \
    -maxiters 1 \
    -diags -sv \
    -distance1 kbit20_3

iqtree \
    -s combined_genome.trim.aln.fasta \
    -B 1000 \
    -nt AUTO \
    -m MFP \
    -pre trim
```

2.8 BLAST Searches

```
makeblastdb \
    -in cyanophage_ref.fa \
    -dbtype nucl \
    -out cyanophage_ref

nohup blastn \
    -max_target_seqs 1 \
```

```
    -db cyanophage_ref \
    -out ref.cyanophage.blast \
    -query lab.cyanophage.fa \
    -num_threads 10 \
    -outfmt 6 qseqid qlen qstart qend salltitles sseqid slen
       sstart send qcovs bitscore evalue pident
```

3 Whole-Genome Analyses

3.1 Alignment

```
~/bowtie2-2.4.2/bowtie2 \
    -x sample.fa \
    -1 raw_data1.trim.fastq \
    -2 raw_data2.trim.fastq \
    -S sample.sam \
    -p 10
```

3.2 Deduplicate Sequences

```
java -jar ~/Picard/picard.jar SortSam \
    -I sample.sam \
    -O sample_sort.sam \
    -SO coordinate

java -jar ~/Picard/picard.jar MarkDuplicates \
    -I sample_sort.sam \
    -O sample_mdup.bam \
    -M sample_mdup.metrics

java -jar ~/Picard/picard.jar AddOrReplaceReadGroups \
    -I sample_mdup.bam \
    -O sample_mdup_rg.bam \
    -SO coordinate \
    -LB sample \
    -PL illumina \
    -PU sample \
    -SM sample

~/samtools-1.12/samtools index sample_mdup_rg.bam
```

3.3 Mutation and Functional Analyses

```
~/gatk-4.2.0.0/gatk HaplotypeCaller \
    -I sample_mdup_rg.bam \
    -O sample.g.vcf \
    -R ref.fa \
    --sample-ploidy 1 \
    -ERC GVCF
```

Summary

This pipeline provides a comprehensive approach for:

- **Preprocessing & Assembly:** Trimming (Sickle), assembly (IDBA-UD), gene prediction (Prodigal).
- **Binning & Taxonomy:** Coverage mapping (Bowtie2, Samtools), binning (MetaBAT), and taxonomy (CheckM, GTDB-Tk).
- **Viral Discovery:** Identifying viral contigs (VirSorter, DeepVirFinder), evaluating completeness (CheckV), clustering and binning (CD-HIT, vRhyme, vContact2).
- **Abundance & Annotation:** RPKM calculations (CoverM), functional annotation (EggNOG-mapper).
- **Phylogenetic & Variant Analyses:** Tree building (trimal, MUSCLE, IQ-TREE), variant calling (GATK).

For questions or troubleshooting, please contact **Yisha Tan** at ytanap@connect.ust.hk.