DengueSeq Analysis Pipeline – pt2

—

https://github.com/grubaughlab/DENV_pipeline

Ammar Aziz – VIDRL

Doherty Institute

THE UNIVERSITY OF MELBOURNE

The Royal Melbourne Hospital

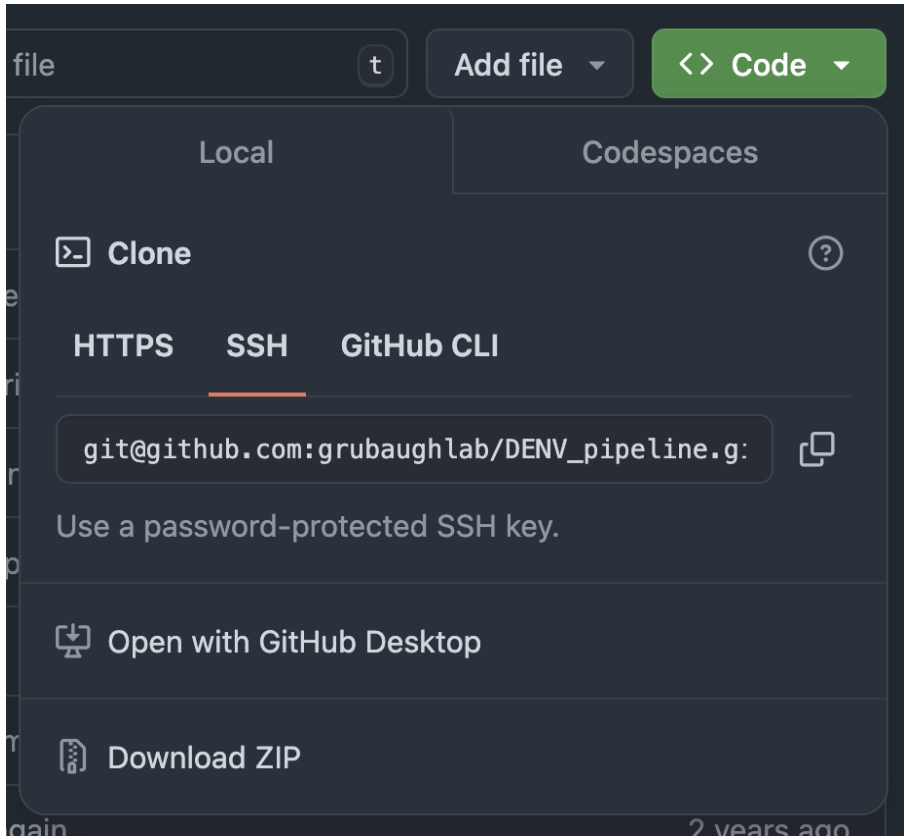
A joint venture between The University of Melbourne and The Royal Melbourne Hospital

Installing DengueSeq analysis pipeline

- Go to the code repository

https://github.com/grubaughlab/DENV_pipeline

- Click "Code" button
- Click "Download ZIP"
- Uncompress



Installing DENV_Analysis code

- Navigate to the location where the pipeline was downloaded and uncompressed
- Install the dependencies:

```
mamba env create -f environment.yml
```

- Activate the created environment

```
mamba activate analysis_env
```

- Install the toolkit

```
pip install .
```

Demo

Preparing inputs

- Your fastq files must be in the following structure:

```
.
├── sample245
│   ├── sample245.R1.fastq.gz
│   └── sample245.R2.fastq.gz
└── sample555
    ├── sample555.R1.fastq.gz
    └── sample555.R2.fastq.gz
```

- Note: R1 and R2 must be contain capital R

Running the pipeline - Options

```
usage: denv_pipeline [--config CONFIG] [--dry-run] [--symlink SYMLINK] [--indir INDIR] [--outdir OUTDIR]
                     [--reference-directory REFERENCE_DIRECTORY] [--depth DEPTH] [--threshold THRESHOLD]
                     [--temp] [--tempdir TEMPDIR] [--download] [--slurm] [--slurm-cores SLURM_CORES]
                     [--cores CORES] [--verbose] [--help] [--overwrite] [--ct-file CT_FILE]
                     [--ct-column CT_COLUMN] [--id-column ID_COLUMN]

optional arguments:
  --config CONFIG       config file containing all relevant arguments
  --dry-run             do all error checks and make files but don't run the pipeline.
  --symlink SYMLINK     argument for generating symlinks
  --indir INDIR         directory containing samples. Each sample must be a folder with the forward and reverse
                        runs in. Default is same as output directory
  --outdir OUTDIR       location where files will be stored.
  --reference-directory REFERENCE_DIRECTORY, -rd REFERENCE_DIRECTORY
                        location where bed files and reference genomes are
  --depth DEPTH         depth to map sequences to. Default=10
  --threshold THRESHOLD
                        threshold to call consensus positions at, default=0.75
  --temp                keep intermediate files
  --tempdir TEMPDIR     where the temporary files go
  --download            make a folder without bam files for download
  --slurm               flag for if running on HPC with slurm
  --slurm-cores SLURM_CORES
                        number of slurm cores to assign. Default is 10
  --cores CORES         number of non-slurm cores to assign. Default is 1
  --verbose, -v
  --help, -h
  --overwrite           overwrite current results
  --ct-file CT_FILE     to produce a plot of Ct against coverage, provide a csv file containing Ct information
                        by sample
  --ct-column CT_COLUMN
                        Name of Ct column in Ct file for plot
  --id-column ID_COLUMN
                        Name of ID column in Ct file to make Ct plot
```

- Lots of options! Most are optional.

Mandatory:

- --indir
- --outdir
- --cores

Running the pipeline

- Parameters we must set for the pipeline:

  --indir - the structured input directory with your fastq files

  --outdir - the name of the output directory

  --cores - cpus to use

Final command:

```
denv_pipeline --indir input --outdir results --cores 10
```

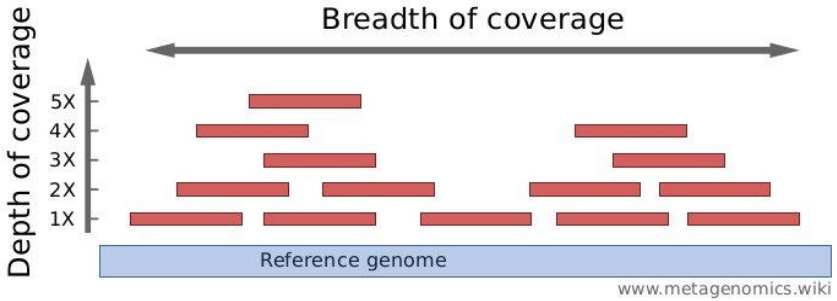
Pipeline Output

- Lots of outputs. Most important:
  - **virus_calls.tsv:** Contains virus calls per sample. I.e. those viruses with which the sample has more than 50% coverage
  - **top_virus_all_samples.tsv:** Contains the highest coverage virus per sample, regardless of coverage
  - **consensus folder** - consensus sequences of the called virus for each sample

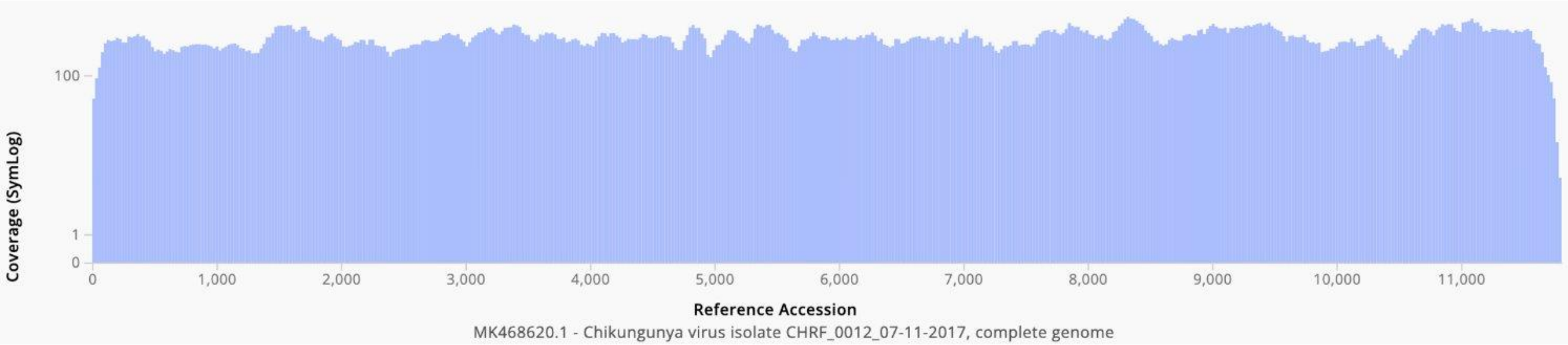
Pipeline Output - understanding

1. Check "virus_calls.tsv" file – open this in excel
   - Do the virus calls for each of your samples align with what you were expecting? For example, were you expecting mostly DENV2?
   - Any oddities, such as sylvatic (rare) detected?
2. Check "top_virus_calls.tsv" - open in excel
   - The virus chosen by the pipeline should be clear – that is no other dengue serotype is above the 50% apart from what was shown above
   - Example:
3. Check the how much of the genome is covered per sample
   - Does it match the Ct values?
   - Good rule: samples with Ct below 29 should provide good (>80%) genome coverage

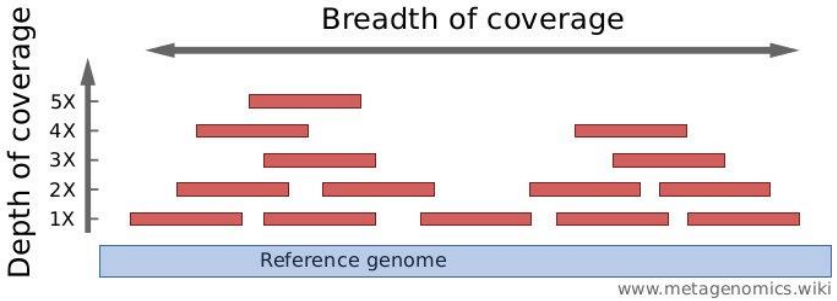
QC Metrics – Coverage plot



- **Coverage plot –** Coverage plots are good for looking at the breadth and depth of the assembled genome



QC Metrics – Coverage plot



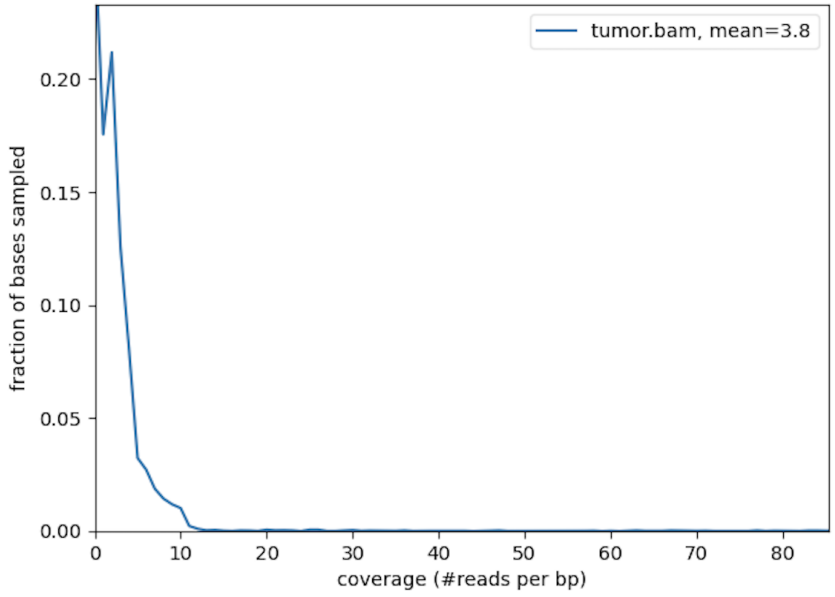
- **Coverage plot –** Coverage plots are good for looking at the breadth and depth of the assembled genome
- Depth:
  - Per-base coverage is the average number of times a base of a genome is sequenced. The coverage depth of a genome is calculated as the number of bases of all short reads that match a genome divided by the length of this genome. It is often expressed as 1X, 2X, 3X,… (1, 2, or, 3 times coverage).
- Breadth
  - Breadth of coverage is the percentage of bases of a reference genome that are covered with a certain depth. For example: "90% of a genome is covered at 1X depth; and still 70% is covered at 5X depth."

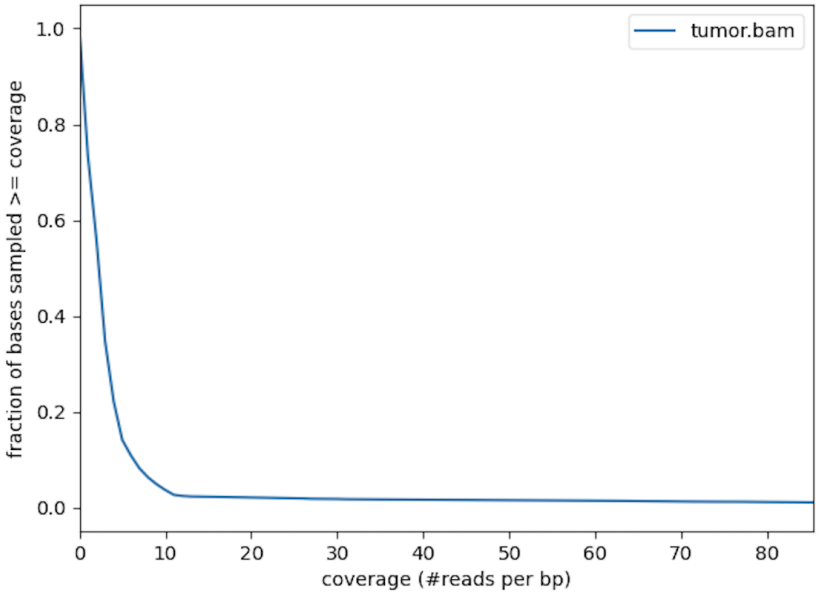
QC Metrics – Coverage plot

- Galaxy tool plotCoverage

plotCoverage assesses the sequencing depth of BAM/CRAM files
(Galaxy Version 3.5.4+galaxy0)

Run Tool





QC Metrics – Coverage plot

- Use different thresholds for different objectives
- Speciation – Which dengue serotype?
  - Genome coverage of >50% + E gene
- Phylogenetics – Whole genome/E gene
  - >80-90% genome coverage
- Clade/Lineage analysis
  - Use Nextclade internal metrics - Next lecture!

OtherQC Metrics

- **% Genome Called** - Refers to the percentage of the genome meeting thresholds for calling consensus bases. The closer this number is to 100%, the better.
- **SNPs** - Indicates the number of single nucleotide polymorphisms. SNPs represent single nucleotide variations between the reference accession and consensus genome.
- **Ambiguous bases** - If multiple sequencing reads support *more* than one nucleotide at a given site, those sites will be designated with an IUPAC ambiguity code. This metric specifies the number of non-C, T, G, A nucleotides in the consensus genome. The consensus genome pipeline only calls nucleotides that are detected at least at 75% frequency.
- **Mapped reads** - Refers to the total number of reads that mapped to the reference genome.

Questions? + Resources

- All the tools in DengueSeq are available in Galaxy, however there isn't a workflow that simplifies the proccess.
- If you use Galaxy, run the process for each reference (x6) and select the genome with the height coverage and depth.