Running NextFlow: ONT-Seq-Analysis & Mpox\_AmpSeq

For information about the pipelines please see here:

[CDCgov/ONT-Seq-analysis: ONT-Seq-analysis is a specialized pipeline designed for analyzing Oxford Nanopore Technologies (ONT) sequencing data from Mpox isolates. It performs reference-based assemblies, identifies variants, designates clades, and generates a detailed quality control (QC) report.](https://github.com/CDCgov/ONT-Seq-analysis)

[CDCgov/Mpox\_AmpSeq: Mpox\_AmpSeq is a custom nextflow-style pipeline for mpox F13L amplicon sequencing. It generates reference-based consensus sequences, variant reports, NextClade outputs including clade designation, and multiple quality control metrics.](https://github.com/CDCgov/Mpox_AmpSeq)

NextFlow Installation and Configuration

1. In location of choice, clone pipeline from github.

git clone https://github.com/CDCgov/Mpox\_AmpSeq.git

git clone https://github.com/CDCgov/ONT-Seq-analysis.git

1. In location of choice, change permissions of directory to ensure executability.

chmod +x ~/path/to/cdcgov-ont-seq-analysis\_main/

chmod +x ~/path/to/cdcgov-Mpox\_AmpSeq/

1. Ensure that additional python script is executable (additional asset python scripts should also be handled like this if used (see bin))

chmod +x ~/path/to/NextFlow\_main/main/bin/check\_samplesheet.py

* 1. Note if you plan on using any additional post-processing script, you will need to make them executable in the same way

Creating a samplesheet.csv

Data will either be:

Already in fastq format, in which case you can:

1. Run the script in /assets/ont\_fastq\_concat\_and\_samplesheet\_create.sh
   1. /path/to/ONT-Seq-analysis/assets/ont\_fastq\_concat\_and\_samplesheet\_create.sh
   2. Press enter and read message,
      1. if you have characters other than the message specifies write y.
         1. You will need to manually update the naming of your fastq data to meet standards described in message.
      2. If you have no other characters that specified in the message write n.
         1. Input the path to your fastq data.
            1. NOTE YOU MUST WRITE FULL FILE PATH

Ex. /scicomp/instruments-pure/23-4-631\_Nanopore-MinION-NP234947/2024/PRB-2024-MPXV-10Kb-15kb-Amplicons-JiushengDeng/PRB-MPXV-10K-15K-barcodes-cladeIIb-TPOXX-4tissues-3-4-2025/4tissues/20250304\_1249\_MC-113388\_AXD563\_57180a83/fastq\_pass/

* + 1. The same script exists in /Mpox\_AmpSeq/
  1. Congrats, you just made a samplesheet, move on to running your NextFlow pipeline!

Data is in raw fast5 output, in which case you need to:

1. Convert to pod5
   1. pod5 convert fast5 /full/path/to/fast5 -o /path/to/merged.out.pod5
      1. Example path to fast5: /scicomp/instruments-pure/23-4-631\_Nanopore-MinION-NP234947/2023/PRB02232023-EC-RABV-MPOX/seq/20230223\_1655\_MN18626\_ANV188\_26d73d90/fast5/
2. Basecall and demultiplex with dorado
   1. qsub /path/to/daisy\_dorado\_pod5\_new\_minqscore\_automodel.sh -f /path/to/merged.out.pod5
   2. This will take a long time, sometimes overnight. Output will be in /path/to/dorado/
      1. Inside the directory dorado will be folders per barcodes of compressed fastq reads
3. Create a samplesheet
   1. /path/to/ONT-Seq-analysis/assets/create\_samplesheet\_only.sh
      1. Enter the path to the data output directory with contatenated fastq files: /path/to/dorado/
      2. A similar script exists in /Mpox\_AmpSeq/samplesheet.sh
4. Congrats, you just made a samplesheet, move on to running your NextFlow pipeline!

Running your NextFlow Pipeline

**For ONT-Seq-analysis (10kb and 15kb amplicon data)**

General Usage:

nextflow run ~/path/to/ONT-Seq-analysis \

-profile <docker/singularity/…/institute> \

--input sampleshee.csv \

--outdir <OUTDIR> \

-resume <#if applicable> \

--fasta <fasta\_path\*> \

--bed\_file <bed\_path\*> \

--fai\_file <fai\_path\*> \

--gff\_file <gff\_path\*> \

--mmi\_file <mmi\_path\*> \

--nextclade\_dataset\_name ‘nextstrain/mpox/all-clades’

\*The biggest issues will be with file paths, to avoid this write out full file path, not relative to working directory; all reference information is available in ONT-Seq-analysis/assets/references/NC063383/ you can point to these files

**Recommended Steps**

1. Once fastq data is available, create a samplesheet.csv
   1. We have written scripts to help with this
      1. If fastq files are already concated by barcode
         1. Script 1: /assets/create\_samplesheet.sh
      2. If fastq files are not concated by barcode
         1. Script 2: /assets/ont\_fastq\_concat\_and\_samplesheet\_create.sh
   2. Run the nextflow pipeline using the general usage example from above.
   3. Additional filtering of NextClade output data available.

The module Ivar\_Variants is turned off by default. To include it on the run, simply add --run\_ivar\_variants true to the nextflow run command (shown above).

Nextclade output files can be modified to remove unnecesary columns. To do so, run the filtering script (assets/nextclade\_tsv\_column\_filter.sh) inside the Nextclade output directory. The resulting TSV file will contain the following colums: 'index', 'seqName', 'clade', 'lineage', 'outbreak', 'qc.overallScore', 'qc.overallStatus', 'totalSubstitutions', 'totalDeletions', 'totalInsertions', 'totalFrameShifts', 'totalMissing', 'totalNonACGTNs', 'failedCdses', 'warnings', 'errors'. Feel free to modify as needed.

We have also included a python script to extract mutations from each nextclade output file and for each specimen, returning a table highlighting key mutations that could indicate a designation of Clade I or Clade II, including the different lineages. Run the python script (assets/pythonX\_nextclade\_parser.py) inside the Nextclade output directory. It will generate a directory named parser that will contain a tsv file for each specimen with the following format:

**For MPOX\_AmpSeq (F13L amplicon data)**

General Usage

nextflow run ~/path/to/Mpox\_AmpSeq \

-profile <docker/singularity/…/institute> \

--input sampleshee.csv \

--outdir <OUTDIR> \

-resume <#if applicable> \

--fasta <fasta\_path\*> \

--bed\_file <bed\_path\*> \

--fai\_file <fai\_path\*> \

--gff\_file <gff\_path\*> \

--mmi\_file <mmi\_path\*> \

--nextclade\_dataset\_name ‘nextstrain/mpox/all-clades’

\*The biggest issues will be with file paths, to avoid this write out full file path, not relative to working directory; all reference information is available in Mpox\_AmpSeq/assets/genome/ you can point to these files

**Recommended Steps**

1. Once fastq data is available, create a samplesheet.csv
   1. We have written scripts to help with this
      1. If fastq files are already concated by barcode
         1. Script 1: /assets/create\_samplesheet.sh
      2. If fastq files are not concated by barcode
         1. Script 2: /assets/ont\_fastq\_concat\_and\_samplesheet\_create.sh
   2. Run the NextFlow pipeline using the general usage example from above.
   3. Additional filtering of NextClade output data available (see Read.me)

Post-processing Scripts:

Two additional scripts are available in the /assets/ directory to help summarize and analyze your results after the Nextflow pipeline completes. These should be run from within your Nextflow output directory.

1. /assets/table\_summary.sh: This script generates a comprehensive summary table of sequencing metrics and Nextclade mutation information. It integrates outputs from SAMTOOLS and 'NextClade' to output a table with read statistics (total reads, mapped reads, average coverage, and average read length), and Nextclade-specific details including nucleotide and amino acid substitutions, as well as nucleotide and amino acid deletions, clade assignment, and Nextclade's overall coverage. The final table is output in the ./summary\_stats directory.

Usage:

/path/to/Mpox\_AmpSeq/assets/table\_summary.sh

1. /assets/match\_mutations.sh: This script performs a targeted analysis by cross-referencing the amino acid substitutions identified by Nextclade against a custom mutation database. It requires your mutation database to be provided as a tab-separated file named mutation\_database.tsv within the /assets/ directory. For each sample, if a Nextclade amino acid substitution matches an entry in your database, the script will report the sample name, the Nextclade amino acid substitution, and all corresponding details from your database (Mutation, AminoAcid, TotalCount, APOBEC3\_Context). The report is saved to ./summary\_stats/matched\_mutations\_report.tsv.

Usage:

/path/to/Mpox\_AmpSeq/assets/match\_mutations.sh