Standard Operating Procedure (SOP)
Illumina Viral Surveillance Panel Analysis Pipeline

Dieter Best, Ph.D.

February 2, 2026

1 Purpose

This document describes the Standard Operating Procedure (SOP) for running the Illumina Viral Surveillance Panel analysis pipeline on Linux systems. The pipeline performs:

- Sequencing quality control
- Host read removal
- Viral detection
- Variant calling
- Phylogenetic-ready consensus generation
- Cohort-level quality control reporting

The pipeline is implemented in Workflow Description Language (WDL) and is fully containerized using Docker images.

2 Scope

This SOP applies to:

- Illumina Viral Surveillance Panel libraries
- Paired-end Illumina sequencing data

- Linux systems with Docker and a WDL engine (Cromwell recommended)

The pipeline is suitable for clinical surveillance, public-health monitoring, and research use.

3 Pipeline Overview

The pipeline processes samples independently using a scatter–gather model. Each sample undergoes the following stages:

1. Raw read quality control (FastQC)
2. Host alignment and contamination estimation
3. Host read removal (depletion)
4. Viral classification (Kraken2)
5. Viral alignment (BWA)
6. Variant calling
7. Variant quality control (bcftools stats)
8. Global cohort-level reporting (MultiQC)

4 Software Requirements

4.1 Operating System

- Linux (Ubuntu 20.04 or newer recommended)

4.2 Required Software

- Docker
- Java 11+
- Cromwell (WDL execution engine)

Example installation (Ubuntu):

```
sudo apt update
sudo apt install -y docker.io openjdk-11-jre
```

5 Reference Databases

The pipeline requires the following reference resources:

- Host reference genome (FASTA + full BWA index)
- Viral reference genomes (FASTA)
- Kraken2 viral database

All BWA index files must be provided explicitly:

- `.amb`
- `.ann`
- `.bwt`
- `.pac`
- `.sa`

6 Pipeline Inputs

Per sample:

- Paired FASTQ files (`R1`, `R2`)
- Sample identifier

Global inputs:

- Host reference and index files

- Viral reference database
- Kraken2 database directory

Inputs are supplied via a JSON file.

7 Running the Pipeline

7.1 Example Command

```
java -jar cromwell.jar run viral_pipeline.wdl \
  --inputs inputs.json
```

8 Pipeline Stages (Detailed)

8.1 Raw Read Quality Control

FastQC is run on all input FASTQ files to assess:

- Base quality
- GC content
- Adapter contamination

8.2 Host Alignment

Reads are aligned to the host genome using BWA-MEM. The resulting BAM is coordinate-sorted and indexed.

This BAM is used exclusively for QC and host contamination estimation.

8.3 Host Contamination and Viral Fraction

Host contamination is defined as:

$$\text{Host contamination (\%)} = \frac{\text{host-mapped reads}}{\text{total reads}} \times 100$$

Viral fraction is calculated as:

$$\text{Viral fraction (\%)} = 100 - \text{host contamination (\%)}$$

An automatic QC flag is assigned:

Viral Fraction	QC Flag	Interpretation
$\geq 10\%$	PASS	Adequate viral signal
1–10%	WARN	Low viral signal
$< 1\%$	FAIL	Likely negative or failed enrichment

8.4 Host Read Removal

Reads where **both mates are unmapped to host** are extracted using `samtools` and converted back to paired FASTQ files.

Duplicate removal is intentionally **not performed**, as targeted viral panels frequently rely on PCR duplication for sensitivity at low viral load.

8.5 Viral Detection (Kraken2)

Host-depleted FASTQs are classified using Kraken2 with a viral reference database. Summary reports are generated for downstream QC aggregation.

8.6 Viral Alignment

Reads are aligned to viral reference genomes using BWA. The pipeline separates:

- Alignment
- SAM to BAM conversion
- Sorting (coordinate and name)
- Indexing

This modular structure improves reproducibility and debugging.

8.7 Variant Calling

Variants are called from the viral BAM using bcftools. The resulting VCF is bgzipped and indexed.

8.8 Variant Quality Control

`bcftools stats` is run on each VCF to generate:

- SNP and indel counts
- Ts/Tv ratios
- Depth and quality distributions

These metrics are aggregated automatically by MultiQC.

8.9 Global MultiQC Report

A single MultiQC report is generated across all samples, including:

- FastQC summaries
- Host contamination and viral fraction
- Automatic QC flags
- Samtools alignment statistics
- Coverage metrics
- Kraken2 summaries
- bcftools variant statistics

This report is the primary QC deliverable.

9 Pipeline Outputs

Per sample:

- Host-depleted FASTQ files
- Viral BAM and index
- Variant VCF and index
- Host contamination and QC metrics (TSV)

Cohort-level:

- MultiQC HTML report
- MultiQC data directory

10 Quality Control and Interpretation

- Samples flagged `FAIL` should not be used for downstream analysis
- `WARN` samples may require repeat sequencing or review
- Coverage and variant QC should be interpreted in the context of viral load

11 Notes and Limitations

- Duplicate reads are not removed due to lack of UMIs
- Viral fraction is an approximation for panel-based data
- This pipeline is not designed for de novo viral discovery

12 Versioning and Validation

All software versions are container-pinned. Pipeline changes must be documented and revalidated prior to production use.

13 Contact

For pipeline questions, maintenance, or validation documentation, contact the pipeline maintainer.