nf-core/viralrecon

This document describes the output produced by the pipeline. Most of the plots are taken from the MultiQC report, which summarises results at the end of the pipeline. Please click here to see an example MultiQC report generated using the parameters defined in this configuration file to run the pipeline on samples which were prepared from the ncov-2019 ARTIC Network V1 amplicon set and sequenced on the Illumina MiSeq platform in 301bp paired-end format.

The directories listed below will be created in the results directory after the pipeline has finished. All paths are relative to the top-level results directory.

Pipeline overview

The pipeline is built using Nextflow and processes data using the following steps:

- Preprocessing
  - parallel-fastq-dump - Download samples from SRA
  - cat - Merge re-sequenced FastQ files
  - FastQC - Raw read QC
  - fastp - Adapter and quality trimming
- Variant calling
  - Bowtie 2 - Read alignment relative to reference genome
  - SAMtools - Sort, index and generate metrics for alignments
  - iVar trim - Primer sequence removal for amplicon data
  - picard MarkDuplicates - Duplicate read marking and removal
  - picard CollectMultipleMetrics - Whole genome coverage and alignment metrics
  - mosdepth - Whole-genome and amplicon coverage metrics
  - VarScan 2, BCFTools, BEDTools || iVar variants and iVar consensus || BCFTools and BEDTools - Variant calling and consensus sequence generation
    - SnpEff and SnpSift - Genetic variant annotation and functional effect prediction
    - QUAST - Consensus assessment report
  - BCFTools isec - Intersect variants across all callers
- De novo assembly
  - Cutadapt - Primer trimming for amplicon data
  - Kraken 2 - Removal of host reads
  - SPAdes || metaSPAdes || Unicycler || minia - Viral genome assembly
    - BLAST - Blast to reference assembly
    - ABACAS - Order contigs according to reference genome
    - PlasmidID - Assembly report and visualisation
    - Assembly QUAST - Assembly quality assessment
    - Minimap2, seqwish, vg - Call variants from induced genome variation graph
    - Assembly SnpEff and SnpSift - Genetic variant annotation and functional effect prediction
- Workflow reporting and genomes
  - MultiQC - Present QC for raw reads, alignment, assembly and variant calling
  - Reference genome files - Saving reference genome indices/files
  - Pipeline information - Report metrics generated during the workflow execution

Preprocessing

parallel-fastq-dump

Please see the usage docs for a list of supported public repository identifiers and how to provide them to the pipeline. The final sample information for all identifiers is obtained from the ENA which provides direct download links for FastQ files as well as their associated md5sums. If a download link exists, the files will be downloaded by FTP otherwise they will be downloaded using parallel-fastq-dump.

▶ Output files

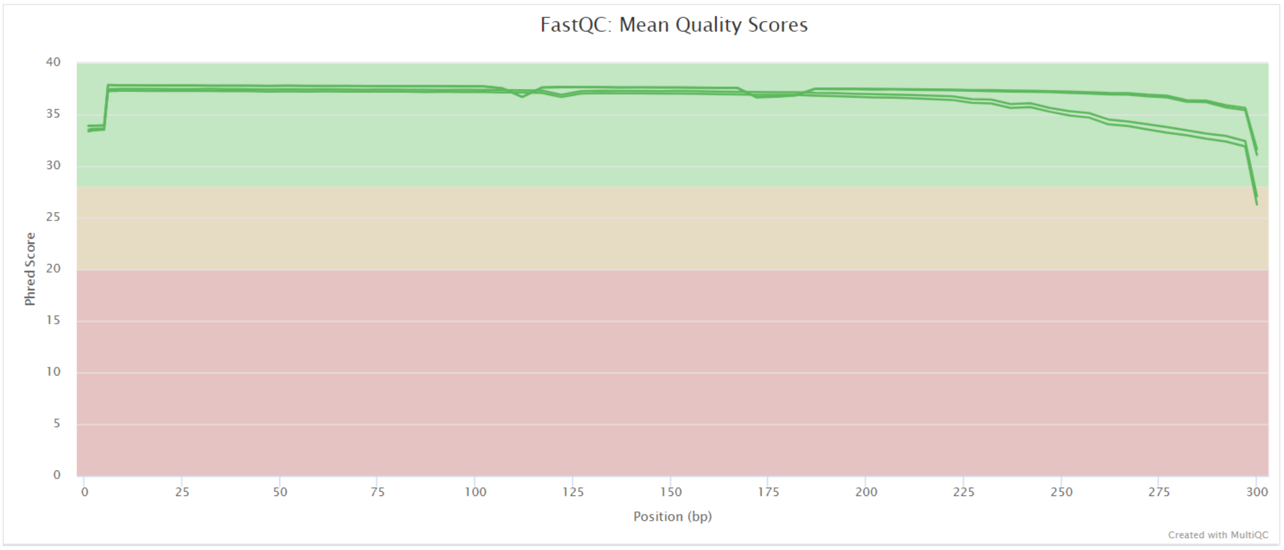
cat

If multiple libraries/runs have been provided for the same sample in the input samplesheet (e.g. to increase sequencing depth) then these will be merged at the very beginning of the pipeline in order to have consistent sample naming throughout the pipeline. Please refer to the usage docs to see how to specify these samples in the input samplesheet.

FastQC

FastQC gives general quality metrics about your sequenced reads. It provides information about the quality score distribution across your reads, per base sequence content (%A/T/G/C), adapter contamination and overrepresented sequences. For further reading and documentation see the FastQC help pages.

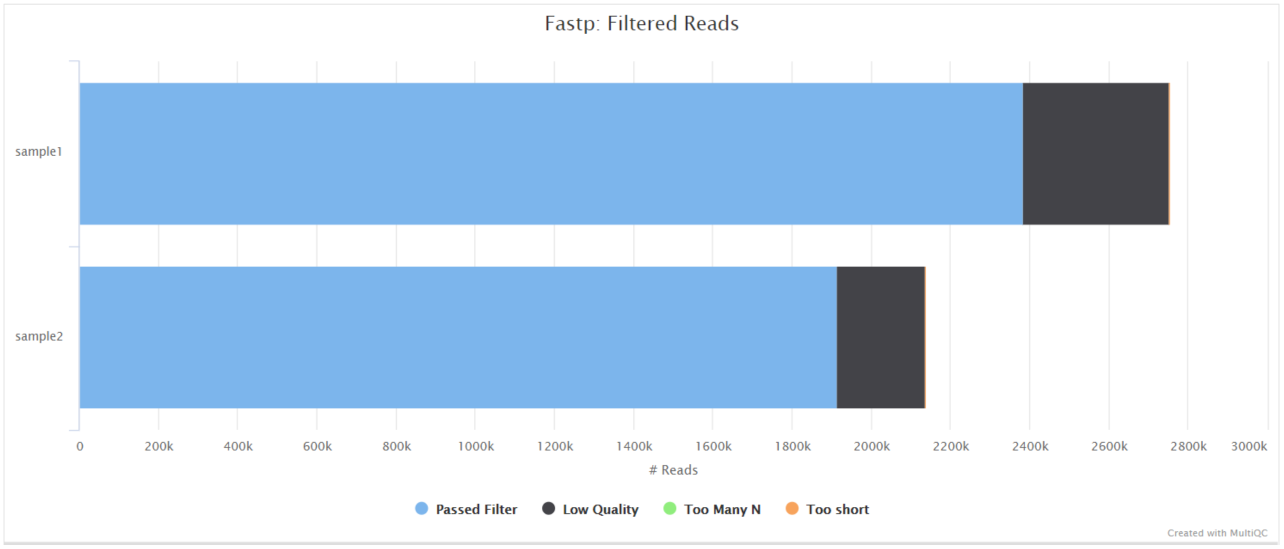
▶ Output files



fastp

fastp is a tool designed to provide fast, all-in-one preprocessing for FastQ files. It has been developed in C++ with multithreading support to achieve higher performance. fastp is used in this pipeline for standard adapter trimming and quality filtering.

▶ Output files



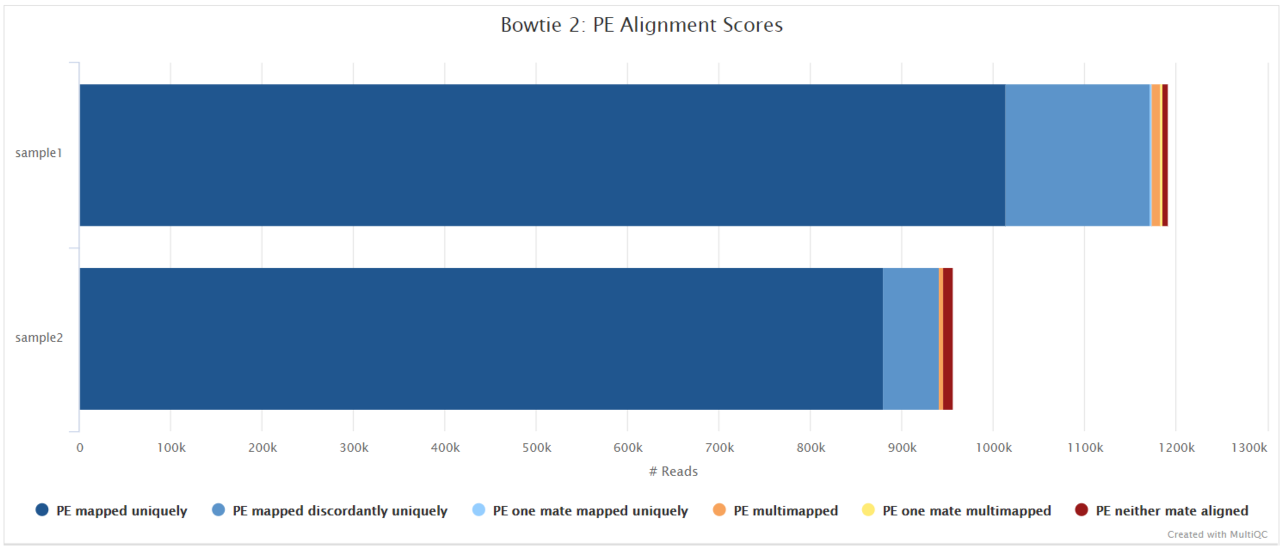
Variant calling

A file called `summary_variants_metrics_mqc.tsv` containing a selection of read and variant calling metrics will be saved in the `variants/` results directory. The same metrics have also been added to the top of the MultiQC report.

Bowtie 2

Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. Bowtie 2 supports gapped, local, and paired-end alignment modes.

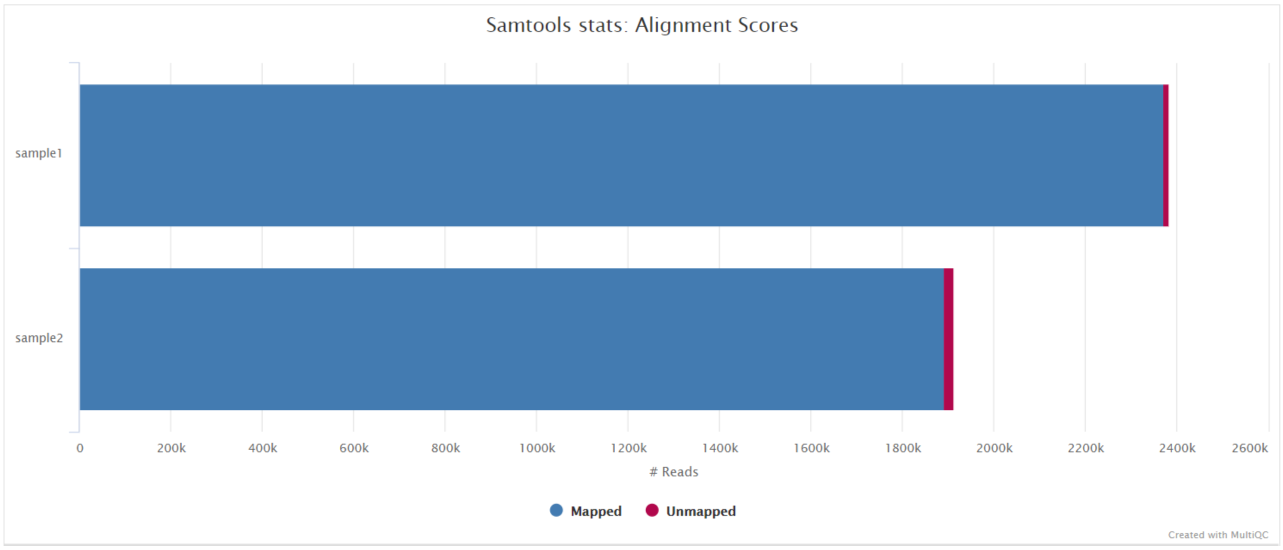
▶ Output files



SAMtools

Bowtie 2 BAM files are further processed with SAMtools to sort them by coordinate, for indexing, as well as to generate read mapping statistics.

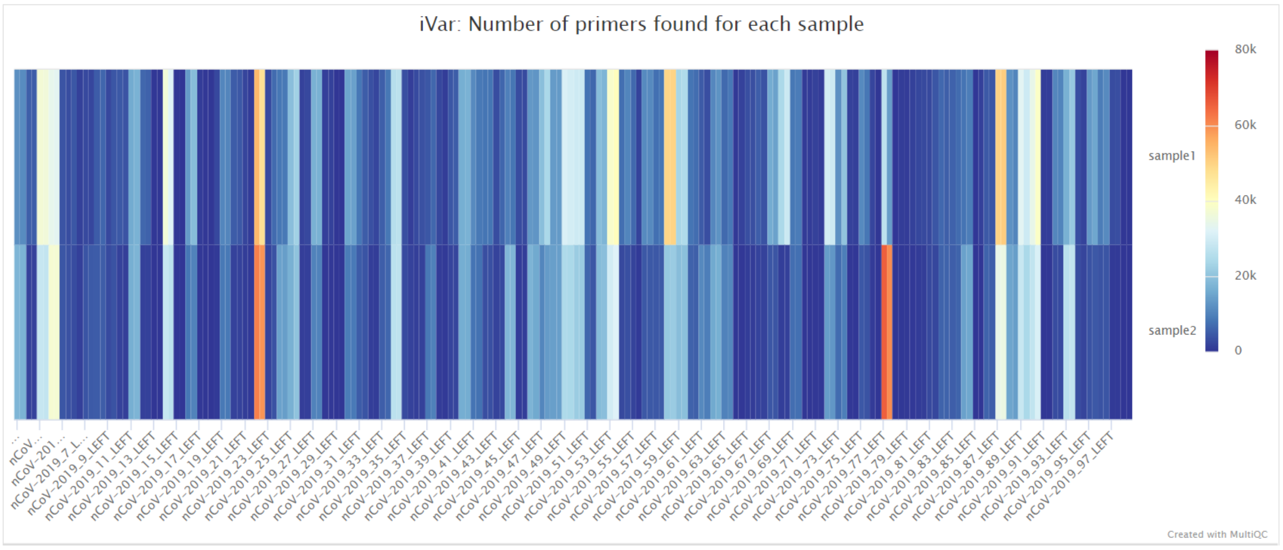
▶ Output files



iVar trim

If the `--protocol amplicon` parameter is provided then iVar is used to trim amplicon primer sequences from the aligned reads. iVar uses the primer positions supplied in `--amplicon_bed` to soft clip primer sequences from a coordinate sorted BAM file.

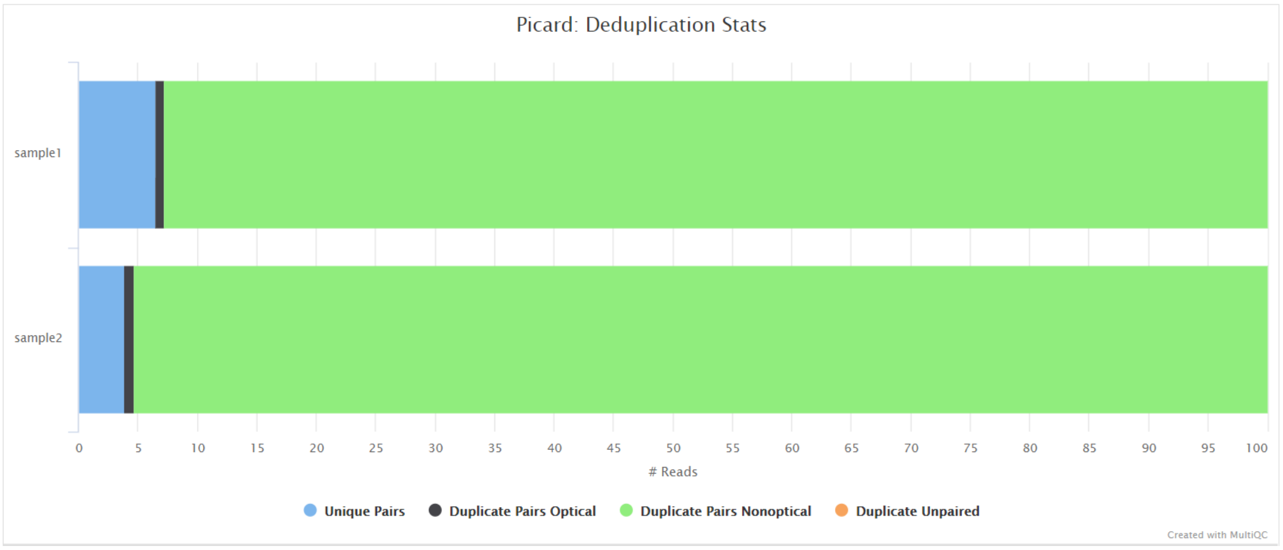
▶ Output files



picard MarkDuplicates

Unless you are using UMIs it is not possible to establish whether the fragments you have sequenced from your sample were derived via true biological duplication (i.e. sequencing independent template fragments) or as a result of PCR biases introduced during the library preparation. By default, the pipeline uses picard MarkDuplicates to *mark* the duplicate reads identified amongst the alignments to allow you to guage the overall level of duplication in your samples. However, you can also choose to remove any reads identified as duplicates via the `--filter_dups` parameter.

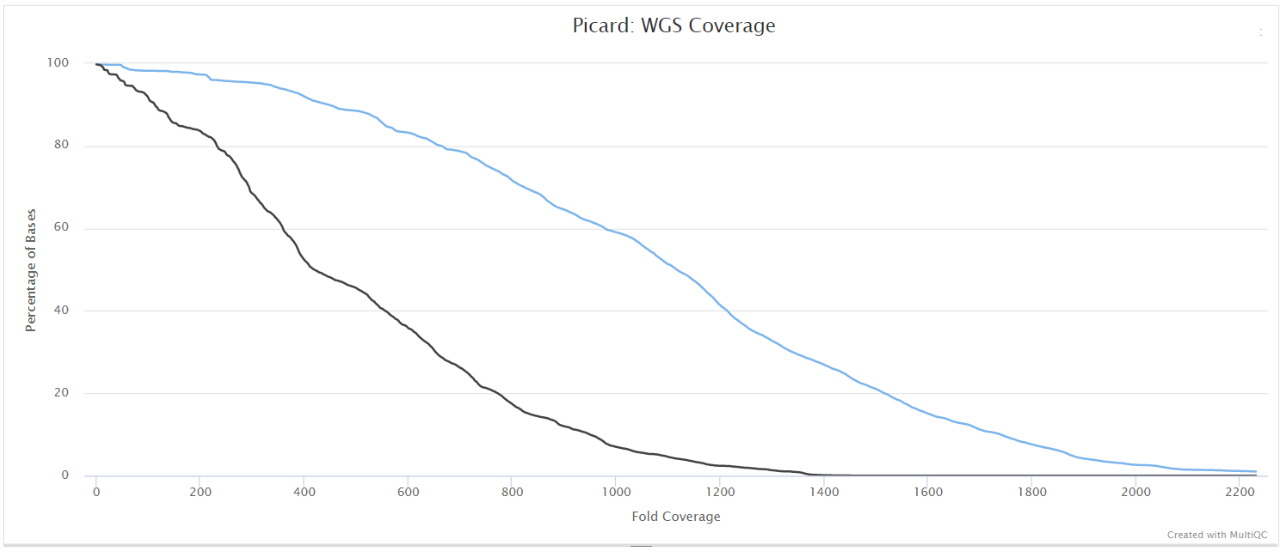
▶ Output files

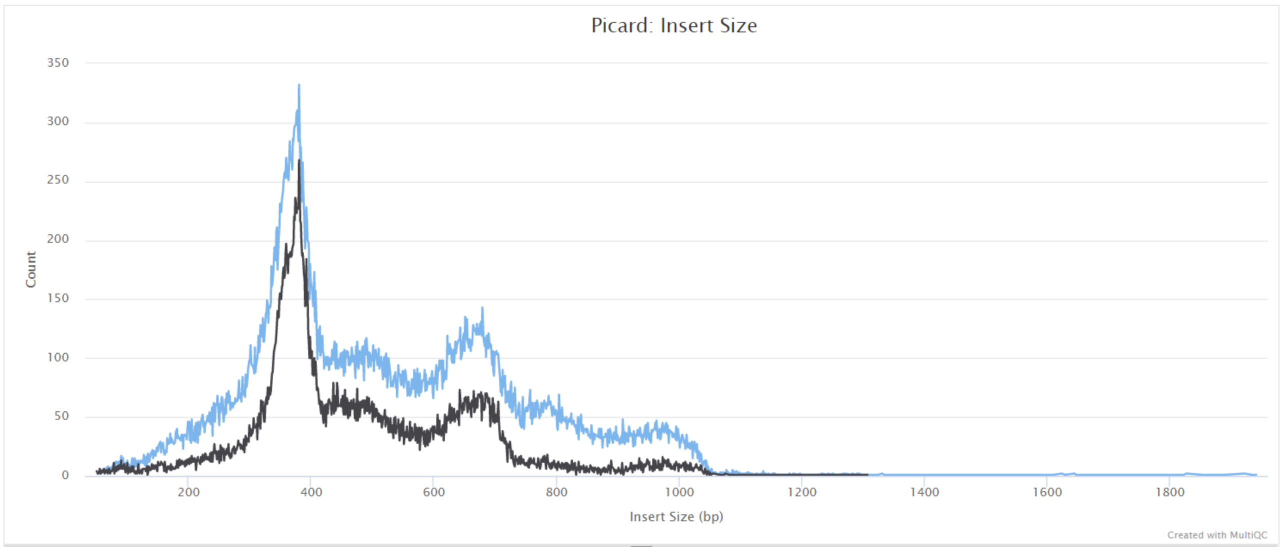


picard CollectMultipleMetrics

picard-tools is a set of command-line tools for manipulating high-throughput sequencing data. We use picard-tools in this pipeline to obtain mapping and coverage metrics.

▶ Output files

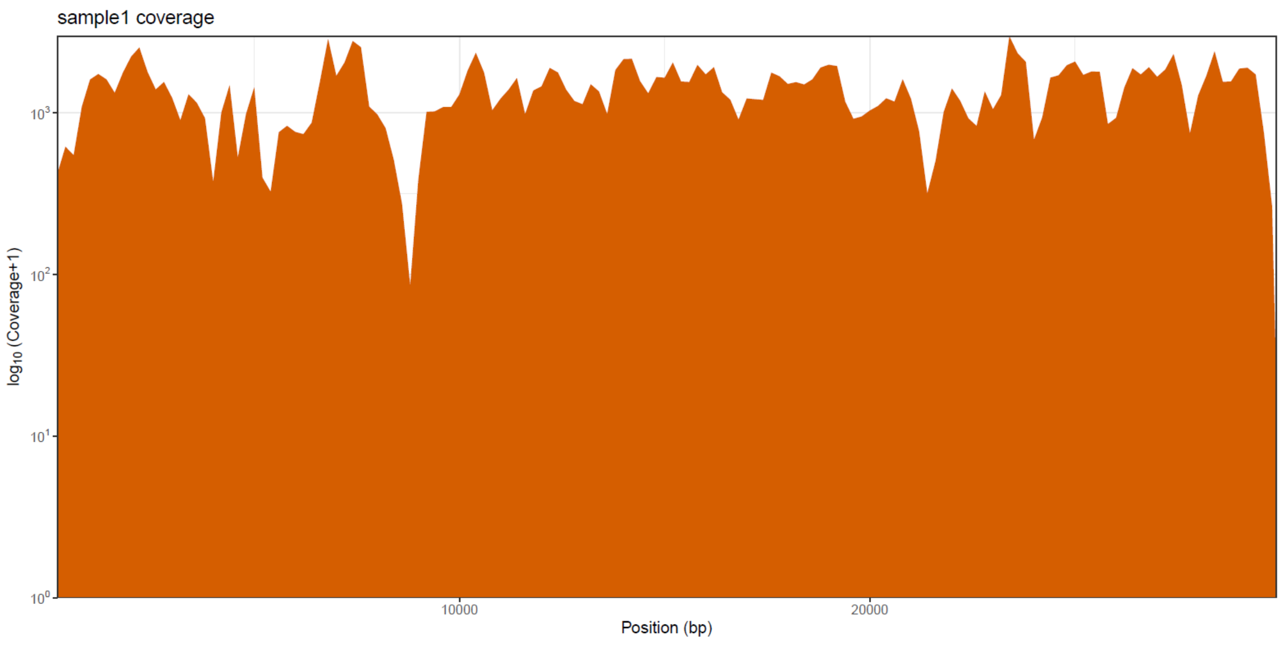


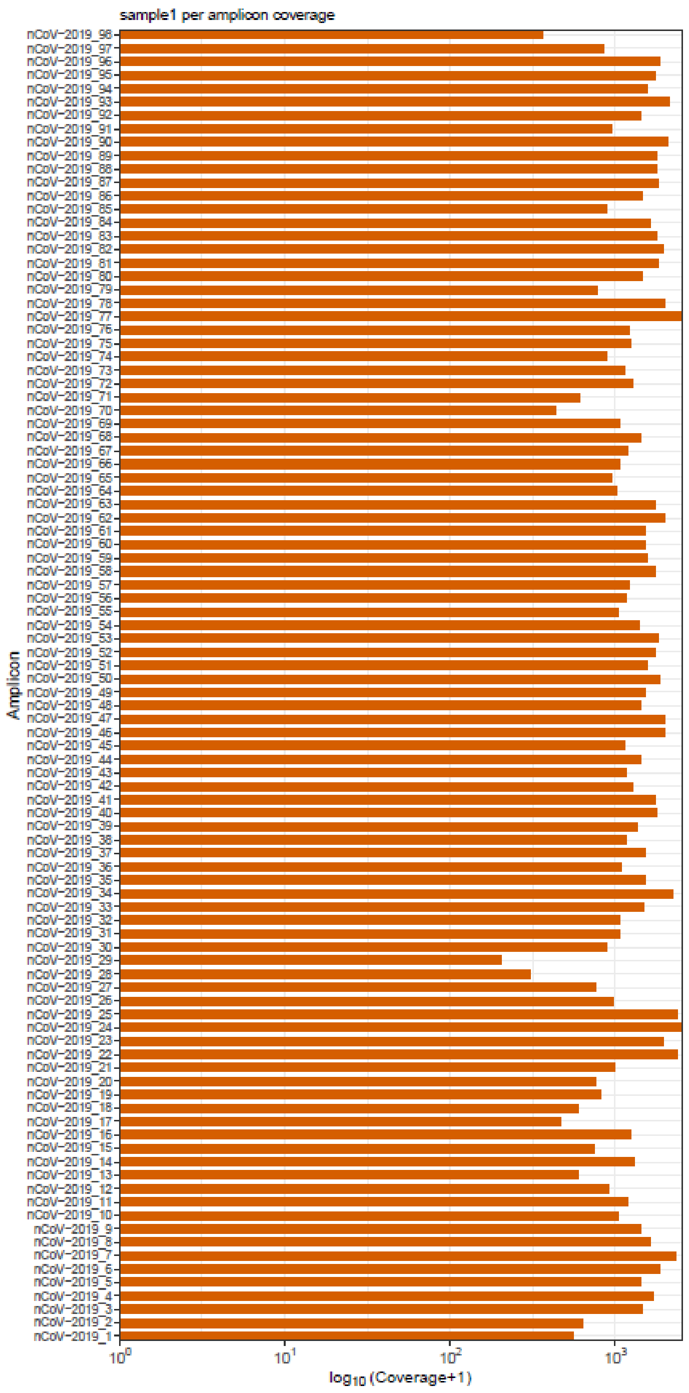


mosdepth

mosdepth is a fast BAM/CRAM depth calculation for WGS, exome, or targeted sequencing. mosdepth is used in this pipeline to obtain genome-wide coverage values in 200bp windows and for `--protocol amplicon` to obtain amplicon/region-specific coverage metrics. The results are then either rendered in MultiQC (genome-wide coverage) or are plotted using custom `R` scripts.

▶ Output files





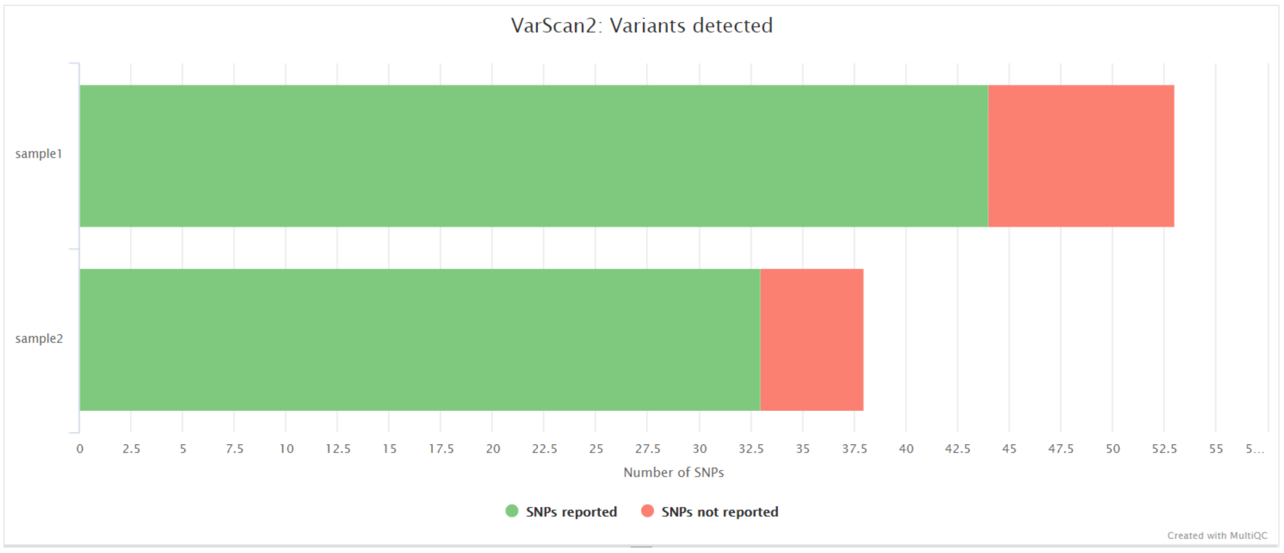
VarScan 2, BCFTools, BEDTools

VarScan 2 is a platform-independent software tool to detect variants in NGS data. In this pipeline, VarScan 2 is used in conjunction with SAMtools in order to call both high and low frequency variants.

BCFtools is a set of utilities that manipulate variant calls in VCF and its binary counterpart BCF format. BCFTools is used in the variant calling and *de novo* assembly steps of this pipeline to obtain basic statistics from the VCF output. It is also used in the VarScan 2 variant calling branch of the pipeline to generate a consensus sequence by integrating high frequency variant calls into the reference genome.

BEDTools is a swiss-army knife of tools for a wide-range of genomics analysis tasks. In this pipeline we use `bedtools genomecov` to compute the per-base mapped read coverage in bedGraph format, and `bedtools maskfasta` to mask sequences in a Fasta file based on intervals defined in a feature file. This may be useful for creating your own masked genome file based on custom annotations or for masking all but your target regions when aligning sequence data from a targeted capture experiment.

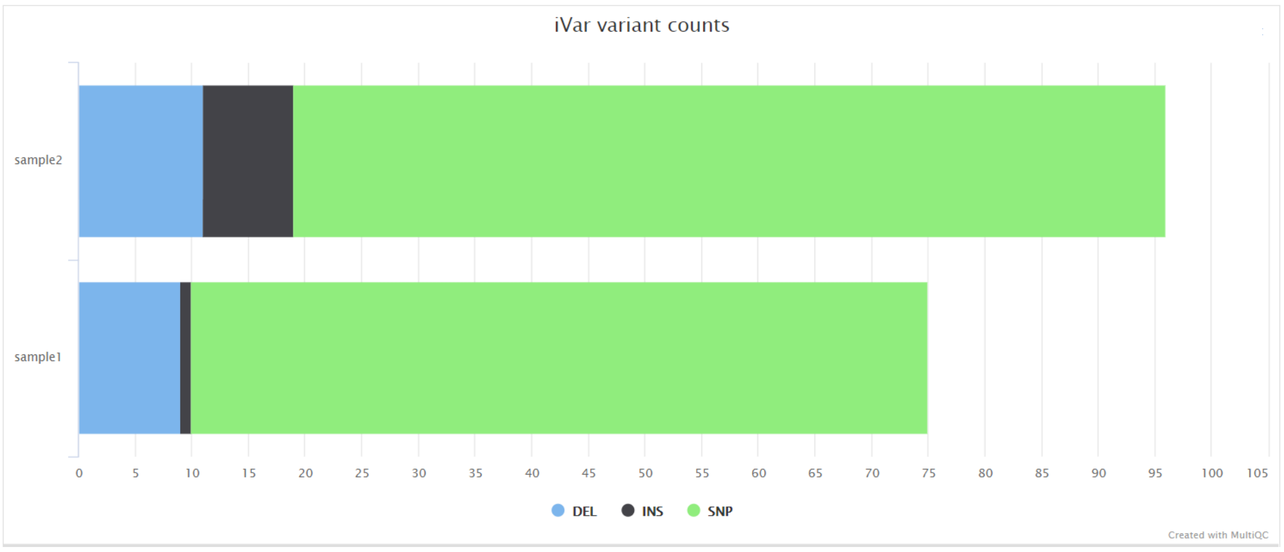
▶ Output files



iVar variants and iVar consensus

iVar is a computational package that contains functions broadly useful for viral amplicon-based sequencing. We use iVar in this pipeline to trim primer sequences for amplicon input data as well as to call variants and for consensus sequence generation.

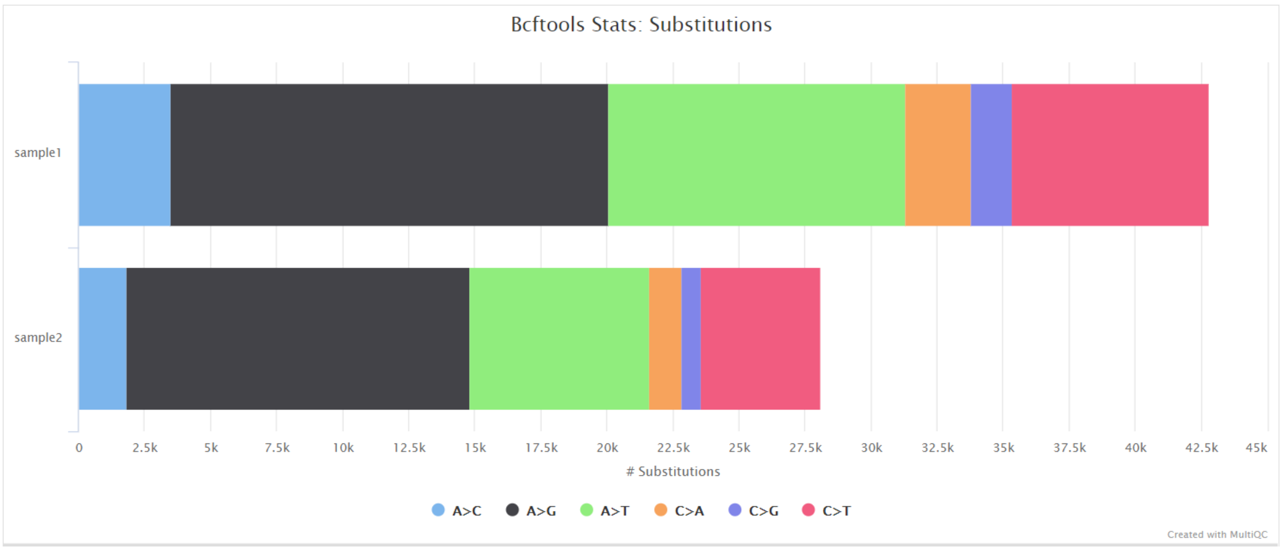
▶ Output files



BCFTools and BEDTools

BCFtools can be used to call variants directly from BAM alignment files. The functionality to call variants with BCFTools in this pipeline was inspired by work carried out by Conor Walker. In contrast to VarScan 2 and iVar, the original variant calls obtained by BCFTools are not filtered further by a higher allele frequency. It seems that the default calls obtained by BCFTools appear to be comparable with the high frequency variants generated by VarScan 2 and iVar.

▶ Output files

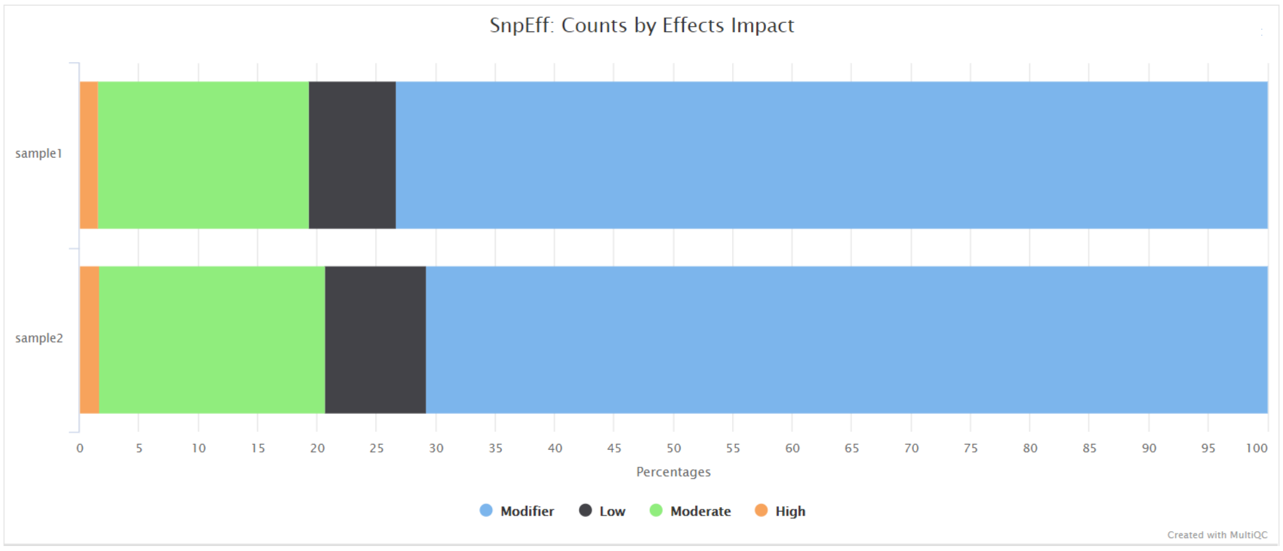


SnpEff and SnpSift

SnpEff is a genetic variant annotation and functional effect prediction toolbox. It annotates and predicts the effects of genetic variants on genes and proteins (such as amino acid changes).

SnpSift annotates genomic variants using databases, filters, and manipulates genomic annotated variants. After annotation with SnpEff, you can use SnpSift to help filter large genomic datasets in order to find the most significant variants.

▶ Output files



QUAST

QUAST is used to generate a single report with which to evaluate the quality of the consensus sequence across all of the samples provided to the pipeline. The HTML results can be opened within any browser (we recommend using Google Chrome). Please see the QUAST output docs for more detailed information regarding the output files.

▶ Output files

BCFTools isec

BCFTools isec can be used to intersect the variant calls generated by the 3 different callers used in the pipeline. This permits a quick assessment of how consistently a particular variant is being called using different algorithms and to prioritise the investigation of the variants.

▶ Output files

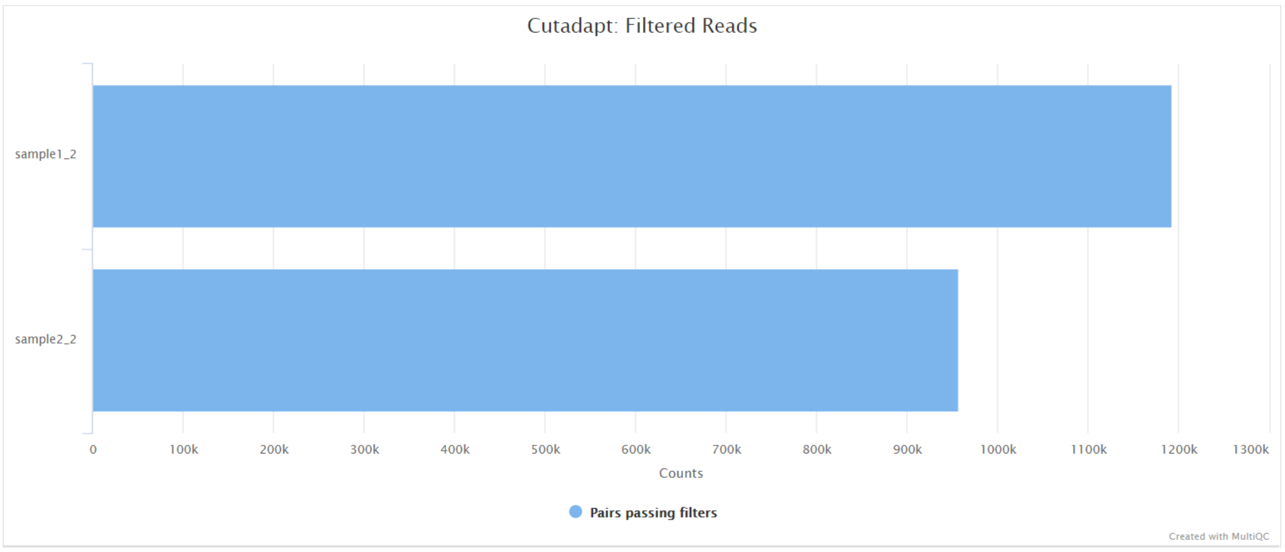
De novo assembly

A file called `summary_assembly_metrics_mqc.tsv` containing a selection of read and *de novo* assembly related metrics will be saved in the `assembly/` results directory. The same metrics have also been added to the top of the MultiQC report.

Cutadapt

In the variant calling branch of the pipeline we are using iVar trim to remove primer sequences from the aligned BAM files for amplicon data. Since in the *de novo* assembly branch we don't align the reads, we use Cutadapt as an alternative option to remove and clean the primer sequences directly from FastQ files.

▶ Output files

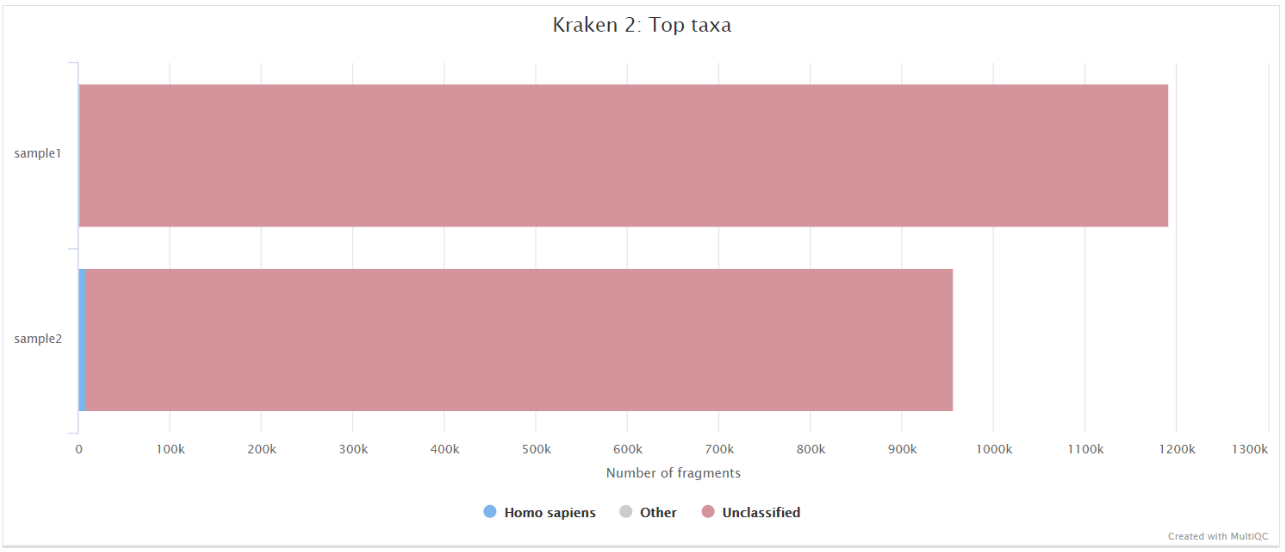


Kraken 2

Kraken 2 is a sequence classifier that assigns taxonomic labels to DNA sequences. Kraken 2 examines the k-mers within a query sequence and uses the information within those k-mers to query a database. That database maps k-mers to the lowest common ancestor (LCA) of all genomes known to contain a given k-mer.

We used a Kraken 2 database in this workflow to filter out reads specific to the host genome. The remainder of the reads are then passed to numerous *de novo* assembly algorithms in order to reconstruct the viral genome.

▶ Output files



SPAdes

SPAdes is an assembly toolkit containing various assembly pipelines. Generically speaking, SPAdes is one of the most popular de Bruijn graph-based assembly algorithms used for bacterial/viral genome reconstruction.

Bandage is a program for visualising *de novo* assembly graphs. By displaying connections which are not present in the contigs file, Bandage opens up new possibilities for analysing *de novo* assemblies.

▶ Output files

metaSPAdes

metaSPAdes is a de Bruijn graph-based assembler that is distributed with SPAdes and executed via the `--meta` option. It can be used for the simultaneous reconstruction of multiple genomes as observed in metagenomics data.

▶ Output files

Unicycler

Unicycler is an assembly pipeline for bacterial genomes. It can assemble Illumina-only read sets where it functions as a SPAdes-optimiser.

▶ Output files

minia

Minia is a short-read assembler based on a de Bruijn graph, capable of assembling a human genome on a desktop computer in a day. The output of Minia is a set of contigs. Minia produces results of similar contiguity and accuracy to other de Bruijn assemblers.

▶ Output files

BLAST

blastn is used to align the assembled contigs against the virus reference genome.

▶ Output files

ABACAS

ABACAS was developed to rapidly contiguate (align, order, orientate), visualize and design primers to close gaps on shotgun assembled contigs based on a reference sequence.

▶ Output files

PlasmidID

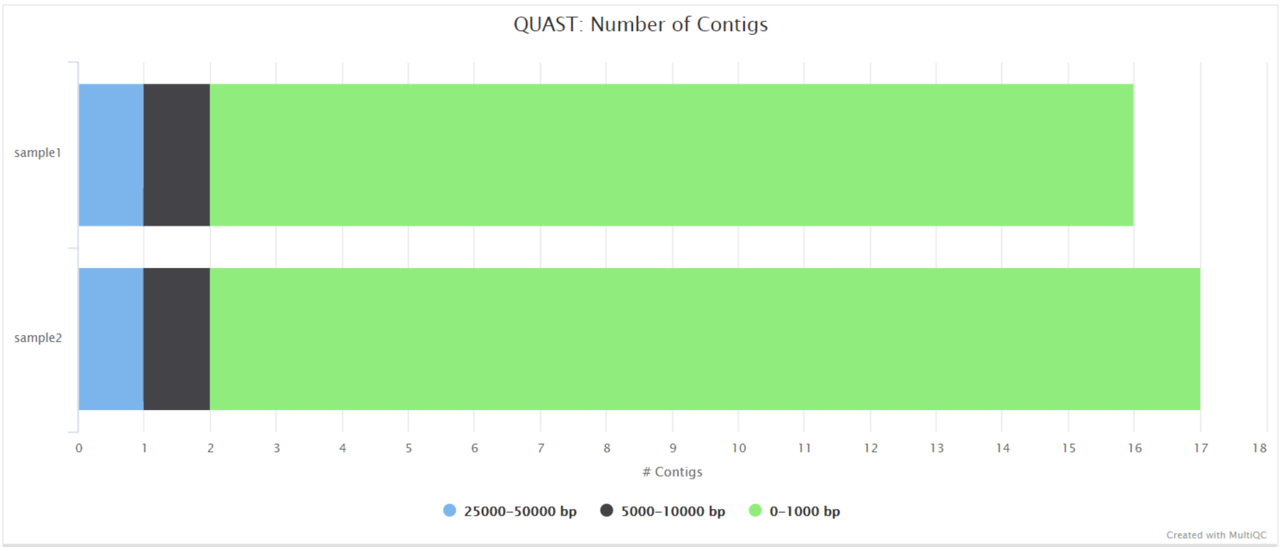
PlasmidID was used to graphically represent the alignment of the reference genome relative to a given assembly. This helps to visualize the coverage of the reference genome in the assembly. To find more information about the output files refer to the documentation.

▶ Output files

Assembly QUAST

QUAST is used to generate a single report with which to evaluate the quality of the *de novo* assemblies across all of the samples provided to the pipeline. The HTML results can be opened within any browser (we recommend using Google Chrome). Please see the QUAST output docs for more detailed information regarding the output files.

▶ Output files



Minimap2, seqwish, vg

Minimap2 is a versatile sequence alignment program that aligns DNA or mRNA sequences against a large reference database. Minimap2 was used to generate all-versus-all alignments between scaffold assembly contigs and the reference genome.

seqwish implements a lossless conversion from pairwise alignments between sequences to a variation graph encoding the sequences and their alignments. seqwish was used to induce a genome variation graph from the all-versus-all alignment generated by Minimap2.

vg is a collection of tools for working with genome variation graphs. vg was used to call variants from the genome variation graph generated by seqwish.

Bandage, a Bioinformatics Application for Navigating De novo Assembly Graphs Easily, is a GUI program that allows users to interact with the assembly graphs made by de novo assemblers and other graphs in GFA format. Bandage was used to render induced genome variation graphs as static PNG and SVG images.

▶ Output files

Assembly SnpEff and SnpSift

SnpEff is a genetic variant annotation and functional effect prediction toolbox. It annotates and predicts the effects of genetic variants on genes and proteins (such as amino acid changes).

SnpSift annotates genomic variants using databases, filters, and manipulates genomic annotated variants. After annotation with SnpEff, you can use SnpSift to help filter large genomic datasets in order to find the most significant variants.

▶ Output files

Workflow reporting and genomes

MultiQC

MultiQC is a visualization tool that generates a single HTML report summarizing all samples in your project. Most of the pipeline QC results are visualised in the report and further statistics are available in the report data directory.

Results generated by MultiQC collate pipeline QC from FastQC, fastp, Cutadapt, Bowtie 2, Kraken 2, VarScan 2, iVar, samtools flagstat, samtools idxstats, samtools stats, picard CollectMultipleMetrics and CollectWgsMetrics, BCFTools, SnpEff and QUAST.

The default `multiqc config file` has been written in a way in which to structure these QC metrics to make them more interpretable in the final report.

The pipeline has special steps which also allow the software versions to be reported in the MultiQC output for future traceability. For more information about how to use MultiQC reports, see http://multiqc.info.

Please click here to see an example MultiQC report generated using the parameters defined in this configuration file to run the pipeline on samples which were prepared from the ncov-2019 ARTIC Network V1 amplicon set and sequenced on the Illumina MiSeq platform in 301bp paired-end format.

▶ Output files

Reference genome files

A number of genome-specific files are generated by the pipeline because they are required for the downstream processing of the results. If the `--save_reference` parameter is provided then the Bowtie 2 alignment indices, BLAST and Kraken 2 databases downloaded/generated by the pipeline will be saved in the `genome/` directory. It is recommended to use the `--save_reference` parameter if you are using the pipeline to build a Kraken 2 database for the host genome. This can be quite a time-consuming process and it permits their reuse for future runs of the pipeline or for other purposes.

▶ Output files

Pipeline information

Nextflow provides excellent functionality for generating various reports relevant to the running and execution of the pipeline. This will allow you to troubleshoot errors with the running of the pipeline, and also provide you with other information such as launch commands, run times and resource usage.

▶ Output files