Output description

Pipeline overview

- FastQC v0.11.8 - read quality control.
- Trimmomatic v.0.33 - adapter and low quality trimming.
- BWA v0.7.12 - mapping against reference genome.
- SAMtools v1.9 - Alignment result processing and unmapped reads selection.
- Flash v1.2.11 - Merge overlapping R1 and R2 reads.
- Seqtk v1.3 - Sequences conversions.
- Vsearch v2.15.0 - Amplicon clustering, searching and dereplication.

Depending on the analysis, we will have some ANALYSIS_IDs. This ANALYSIS_IDs are going to be composed of the date of the analysis and some tag that define the type of analysis done.

Preprocessing

FastQC

FastQC gives general quality metrics about your reads. It provides information about the quality score distribution across your reads, the per base sequence content (%T/A/G/C). You get information about adapter contamination and other overrepresented sequences.

For further reading and documentation see the FastQC help.

Output directory: `01-fastqc`

- `{sample_id}/{sample_id}_R[12]_fastqc.html`
  - html report. This file can be opened in your favourite web browser (Firefox/chrome preferable) and it contains the different graphs that fastqc calculates for QC.
- `{sample_id}/{sample_id}_R[12]_fastqc`
  - older with fastqc output in plain text.
- `{sample_id}/{sample_id}_R[12]_fastqc.zip`
  - zip file containing the FastQC report, tab-delimited data file and plot images

Trimming

Trimmomatic is used for removal of adapter contamination and trimming of low quality regions. Parameters included for trimming are:

- Nucleotides with phred quality < 10 in 3'end.
- Mean phred quality < 20 in a 4 nucleotide window.
- Read lenght < 50

Results directory: `02-preprocessing`

- Files:
  - `{sample_id}/{sample_id}_R[12]_filtered.fastq.gz`: contains high quality reads with both forward and reverse tags surviving.
  - `{sample_id}/{sample_id}_R[12]_unpaired.fastq.gz`: contains high quality reads with only forward or reverse tags surviving.

Note:To see how your reads look after trimming, look at the FastQC reports in the ANALYSIS/{ANALYSIS_ID}/03-preprocQC directory

Mapping

BWA

BWA or Burrows-Wheeler Aligner, is designed for mapping low-divergent sequence reads against reference genomes. The result alignment files are further processed with SAMtools, sam format is converted to bam, sorted and an index .bai is generated.

We mapped the fastq file agains both reference host genome and reference viral genome.

Output directory: `04-mapping`

- `{sample_id}_sorted.bam`
  - Sorted aligned bam file.
- `{sample_id}_sorted.bam.bai`
  - Index file for soreted aligned bam.
- `{sample_id}_flagstat.txt`

- Mapping stats summary.

Merge paired-end reads

Flash

Flash is a very fast and accurate software tool to merge paired-end reads from next-generation sequencing experiments.

Output directory: `05-flash`

- `{sample_id}/{sample_id}.extendedFrags.fastq`
  - Fastq file with R1 and R2 merged.
- `{sample_id].hist`
  - txt file with information for generating an histogram showing merged reads distribution.
- `{sample_id}.histogram`
  - txt file with information for generating an histogram showing merged reads distribution.
- `{sample_id}.notCombined.[R1|R2].fastq
  - Reads not able to be merged for R1 and R2.

Clustering and dereplication

Seqtk

Seqtk was used for converting fastq files into fasta files.

Output directory: `06-fastq2fasta`

- `{sample_id}.fasta`
  - Reads in fasta format.

Vsearch: search and clustering

Vsearch was used to search and cluster our fastq files according to the target amplicons. Only reads with 80% of sequences similarity with one of the amplicon references are kept for downstream analysis.

Output directory: `07-vsearch`

- `{sample_id}_matched.fasta`
  - Fasta file with sequences matching more than 80% of sequence identity with the target sequences.
- `{sample_id}_cluster.tab`
  - Tabular file with information about each read aligning with target files.

Vsearch: Dereplication

Vsearch was used to dereplicate our reads, collapsing identical reads into only one sequence indicating the count in the fasta header.

Output directory: `08-dereplication`

- `{sample_id}_{cluster}_seqs_derep.fasta`
  - Fasta file with identical reads collapsed and count added to the header.

QC stats

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.

Output directory: `99-stats`

- `multiqc/`
  - `multiqc_report.html`: a standalone HTML file that can be viewed in your web browser.
  - `multiqc_data/`: directory containing parsed statistics from the different tools used in the pipeline.
  - `multiqc_plots/`: directory containing static images from the report in various formats.

Custom Analysis

Output results generation: filtering and primer removal.

A custom python script is created for parsing matched reads in fasta format and tabular file with cluster information. We use cluster information for separating reads matching each amplicon into different files, and we use the cigar string for removing deletions in the reads ends corresponding with the primer sequences.

Output directory: `07-vsearch`

- `{sample_id}_RT_seqs.fasta`: fasta file with primer sequences trimmed and matching RT amplicon.
- `{sample_id}_V3_seqs.fasta`: fasta file with primer sequences trimmed and matching V3 amplicon