

Pipeline description

Pipeline for bacterial assembly and annotation using nextflow.

Pipeline overview

- [FastQC](#) v0.11.8 - read quality control.
- [Trimmomatic](#) v0.33 - adapter and low quality trimming.
- [Unicycler](#) v0.4.6 - prokaryote assembler.
- [Quast](#) v4.1 - assemblies quality control
- [Kmerfinder](#) v.3.1 - species and contamination determination.

Note:

Depending on the analysis, we could have some ANALYSIS_IDs. This ANALYSIS_IDs are going to be composed of the date of the analysis, and an analysis identification. You can find a README in the ANALYSIS folder with a brief description of the different analysis.

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`
 - folder 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](#) (1) 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 length < 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 03-preprocQC directory

Kmerfinder

[Kmerfinder](#) (2) is a software used for species identification and the determination of possible contamination in the sample. We use this software using the bacterial database provided by the developers, and with the "winner takes it all" algorithm. You can check [here](#) for a description of the columns in the output.

Output directory: `04-kmerfinder/{sample_id}/`

- `data.json`
 - results in json format.
- `results.spa`

- results in spa format.
- `results.txt`
 - results in txt format. **This is the format you have to use if you are going to open it with excel.**

NOTE: You can also find in `99-stats` a summary of all samples results (`kmerfinder.csv`).

Assembly

Unicycler

[Unicycler](#) (3) is an assembly pipeline for bacterial genomes. It can assemble Illumina-only read sets where it functions as a SPAdes-optimiser.

Output directory: `08-unicycler/{sample_id}`

- `{sample_id}.fasta`: fasta file containing the assembled reads in form of contigs and scaffolds. This is the file we use for annotation and upstream analysis.
- `{sample_id}.gfa`: Graph files for the different assembly optimizations. This files can be used by advanced users with software like [Bandage](#)

Annotation and quality control

QUAST

[QUAST](#) (4) evaluates genome assemblies. We compared the reference genome with the contigs and scaffold assemblies. The html results can be opened with any browser (we recommend using Google Chrome).

Output directory: `00-assemblies/quast_results`

- `quast_results/date/report.html`
 - Compressed format of the indexed variants file.
 - The meaning of the different metrics:
 - Contigs ($\geq x$ bp): is total number of contigs of length $\geq x$ bp.
 - Total length ($\geq x$ bp): is the total number of bases in contigs of length $\geq x$ bp.
 - Contigs: is the total number of contigs in the assembly.
 - Largest contig: is the length of the longest contig in the assembly.
 - Total length: is the total number of bases in the assembly.
 - Reference length: is the total number of bases in the reference genome.
 - GC (%): is the total number of G and C nucleotides in the assembly, divided by the total length of the assembly.
 - Reference GC (%): is the percentage of G and C nucleotides in the reference genome.
 - N50: is the length for which the collection of all contigs of that length or longer covers at least half an assembly.
 - NG50: is the length for which the collection of all contigs of that length or longer covers at least half the reference genome. This metric is computed only if the reference genome is provided.
 - N75 and NG75: are defined similarly to N50 but with 75 % instead of 50 %.
 - L50 (L75, LG50, LG75) is the number of contigs equal to or longer than N50 (N75, NG50, NG75). In other words, L50, for example, is the minimal number of contigs that cover half the assembly.

Bibliography

1. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.
2. Rapid and precise alignment of raw reads against redundant databases with KMA Philip T.L.C. Clausen, Frank M. Aarestrup, Ole Lund.
3. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017.
4. Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi and Glenn Tesler. QUAST: quality assessment tool for genome assemblies, *Bioinformatics* (2013) 29 (8): 1072-1075.