Output Bacterial Characterization: resistances, plasmids, virulence and mlst.

Table of Contents

Pipeline overview

- FastQC v0.11.8 - read quality control.
- Trimmomatic v.0.33 - adapter and low quality trimming.
- Ariba v.2.14.6 - a gene finder for resistance, plasmids, etc. identification.
- Srst2 v.0.2.0 - a gene finder for mlst identification.

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 de 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` and `03-preprocQC

- `{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 (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 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 03-preprocQC directory

MultiQC

MultiQC aggregates results from bioinformatics analyses across many samples into a single report.

Results directory: `99-stats`

- Files:
  - `multiqc_report.html`: report in html with all qc metrics aggregated.

Strain characterization

Ariba

Ariba (7) is a tool that identifies antibiotic resistance genes by running local assemblies. It can also be used for MLST calling.

Output directory: `04-ariba`

- `get_prep_ref`: databases references for resistance, virulence, etc.
- `db_*_mlst`: mlst reference database.
- `run_mlst`: mlst results.
  - `{sample_id}/{database_run_folder}`
    - `{sample_id}report.tsv`: report with detected genes (field explanation in ANEXI)
    - `assembled_genes.fa.gz`: sequence for the assembled resistance/virulence/etc genes.
- `run_dbs`: databases results.
  - `{sample_id}/{database_run_folder}`
    - `{sample_id}report.tsv`: report with detected genes (field explanation in ANEXI)
    - `assembled_genes.fa.gz`: sequence for the assembled resistance/virulence/etc genes.
- `summary`: summary files for all samples.

SRST2

SRST2 is designed to take Illumina sequence data, a MLST database and/or a database of gene sequences (e.g. resistance genes, virulence genes, etc) and report the presence of STs and/or reference genes.

Output directory: `05-mlst`

- {sample_id}/ : contains results per sample.
- summary: contains a table with mlst determination for all samples.

ANEXES

ANEX I

Column	Description
1. ariba_ref_name	ariba name of reference sequence chosen from cluster (needs to rename to stop some tools breaking)
2. ref_name	original name of reference sequence chosen from cluster, before renaming
3. gene	1=gene, 0=non-coding (same as metadata column 2)
4. var_only	1=variant only, 0=presence/absence (same as metadata column 3)
5. flag	cluster flag
6. reads	number of reads in this cluster
7. cluster	name of cluster
8. ref_len	length of reference sequence
9. ref_base_assembled	number of reference nucleotides assembled by this contig
10. pc_ident	%identity between reference sequence and contig
11. ctg	name of contig matching reference
12. ctg_len	length of contig

Column	Description
13. ctg_cov	mean mapped read depth of this contig
14. known_var	is this a known SNP from reference metadata? 1 or 0
15. var_type	The type of variant. Currently only SNP supported
16. var_seq_type	Variant sequence type. if known_var=1, n or p for nucleotide or protein
17. known_var_change	if known_var=1, the wild/variant change, eg I42L
18. has_known_var	if known_var=1, 1 or 0 for whether or not the assembly has the variant
19. ref_ctg_change	amino acid or nucleotide change between reference and contig, eg I42L
20. ref_ctg_effect	effect of change between reference and contig, eg SYS, NONSYN (amino acid changes only)
21. ref_start	start position of variant in reference
22. ref_end	end position of variant in reference
23. ref_nt	nucleotide(s) in reference at variant position
24. ctg_start	start position of variant in contig
25. ctg_end	end position of variant in contig
26. ctg_nt	nucleotide(s) in contig at variant position
27. smtls_total_depth	total read depth at variant start position in contig, reported by mpileup
28. smtls_nts	nucleotides on contig, as reported by mpileup. The first is the contig nucleotide
29. smtls_nts_depth	depths on contig, as reported by mpileup. One number per nucleotide in the previous column
30. var_description	description of variant from reference metdata
31. free_text	other free text about reference sequence, from reference metadata