Exam NGS

Backgroud

You work as a bioinformatician in a small hospital. The in-house lab recently acquired an Illumina MiSeq sequencer. Lab is trying to switch all diagnostic processes to next generation sequencing (NGS).
One of the first processes to address is the prediction of antibiotic resistance alternatively to antibiogram. Antibiograms are complex and time-consuming, so NGS analysis has obvious advantages.
Hospital board made a decision and your task now is to develop the antibiotic resistance prediction pipeline.

Technical requirements

The laboratory collects the samples and performs sequencing (single-end) twice a day: at noon and in the evening. The FASTQ files for resistance prediction are placed together in one folder and named "patient**x**.fastq", where **x** is the ascending patient number (patient data are pseudonymised for data protection reasons).
The resistance prediction is to be carried out using the srst2 tool. On the server, on which the pipeline has to run, nextflow and singularity are available – other tools must not be installed (so any dependencies must go through singularity-images).

Task description

Develop a nextflow pipeline that includes the following analysis steps:

- Trimming the data using fastp (with default parameters)
- Generation of a quality report on the trimmed data using fastqc
- Resistance prediction using srst2
- The pipeline should meet the following requirements:
- All external tools (fastp, fastqc, srst2) should be run in singularity images
- The pipeline should get a whole folder with fastq files as an input
- Analysis should be performed on all fastq files (parallelization where possible)
- Output results should be summarized in a single file
- The FASTA-file containing the resistance prediction reference database should be supplied as a parameter on command line
- To test and apply the pipeline please use CARD_v3.0.8_SRST2 database from SRST2 repository (https://github.com/katholt/srst2/blob/master/data/CARD_v3.0.8_SRST2.fasta)

CAUTION!
Use the following Singularity Image for SRST2, and NOT the newest one:
https://depot.galaxyproject.org/singularity/srst2%3A0.2.0--py27_2

TIPPS:

- It is sufficient if you include the __results.txt files for all srst2 runs merged into one final result file. Do NOT try to cut the headers: those are different in every __results.txt file, it would be way too much effort to combine the headers
- Remember that channel.fromPath() can output multiple files if used wisely. Pointing to directory "mydir" (channel.fromPath("mydir")) will create channel with single path to this directory. Using wildcard * will get the files in the directory "mydir" (channel.fromPath("mydir/*")) as an array: ["a.txt", "b.txt"]. If these files should be processed separately, the channel can be flattened (channel.fromPath("mydir/*").flatten())
- Also remember that a process only runs as long as all of its input channels are filled. If srst2 process uses both a channel with fastq files (potentially many) and a reference file (only one) you'll probably need to .combine() the channels

Application

Test your pipeline with the fastq files in rawdata.tar.gz on the BSCW server. Here is a brief listing of relevant classes of antibiotics with corresponding antibiotic resistance-bearing genes:

Aminoglycosides	APH, StrA, StrB, SAT
Betalactames	AmpC, TEM, OXA*
Trimethoprimes	Dfr*
Sulfonamides	Sul*
Macrolides	Mph*
Tetracyclines	Tet*
Glycopeptides	ANT*
Phenicol	Cat_Phe

Example: if gene "aph6-Id.v1.171" is detected, it confers resistance to antibiotics from the class of Aminoglycosides (since "APH*" is listed under "Aminoglycosides" in the table)

Your hospital stocks the following antibiotics:

Name	Class	Priority
Doxycycline	Tetracyclines	1
Gentamycine	Aminoglycosides	2
Chloramphenicol	Phenicol	3
Trimetoprim	Diaminopyrimidine	4

Based on the results of your pipeline run, suggest a suitable antibiotic treatment for each of the three patients. Always suggest the antibiotic with the lowest priority value (e.g. doxycycline rather than gentamycin), but the antibiotic is only given if no resistance genes against it are found in the sample.

Example: Patient25 has "tetA.v1_2317" in srst2 result. He/she should not be prescribed tetracyclines - Tet* mediates resistance to tetracyclines - and Dexocycline would fail. Patient25 should get antibiotic with next priority: gentamycin.

Write your recommendations in a text file "Recommendations.md" in your git repository

The task submission is only possible as a link to a git repository (if it is private, remember to invite me to the repository on github!)