Bioinformatics and Data Centre
Core Facilities
University of Gothenburg

Technical evaluation part 1

You have 1.5 hrs to complete this part of the technical evaluation. The script is to be written in python and submitted via a public GitHub repo. Emphasis should be put on making the script reliable, robust, and reproducible. Email the web address of the repo to marcela.davila@gu.se at the end of the allocated session together with part 2 (see below).

***Problem**:* Every week, virus samples are being sequenced and analyzed using a commonly found pipeline. This pipeline aligns all reads generated against a reference genome, and from this alignment it produces a consensus sequence. Besides the consensus fasta sequences, the pipeline also produces a file with quality metrics for all the samples analyzed. You have been provided such a file with quality metrics for each sample (samples.txt). It consists of 373 samples (samples.txt), and the head of the file looks like this:

```
sample,pct_N_bases,pct_covered_bases,longest_no_N_run,num_aligned_reads,qc_pass
DN-64554,3.91,96.07,7055,489499,TRUE
DC-31756,4.14,95.93,7055,527966,TRUE
DD-28879,4.32,95.68,9033,444775,TRUE
DD-22466,3.63,96.37,7055,621979,TRUE
DC-95171,4.13,95.86,8855,510658,TRUE
DT-54370,3.80,96.18,12196,612425,TRUE
DT-97532,0.24,99.74,29786,644779,TRUE
DD-48974,0.12,99.86,29822,719234,TRUE
DC-90361,0.24,99.74,29786,519893,TRUE
```

As you can see this is a csv file with 6 columns. The columns show:

1. Name of the sample
2. Percentage of the bases in the consensus is denoted as ‘N’
3. Percentage of the bases of the reference genome has been covered
4. How long is the longest sequence in the consensus consisting of no ‘N’s
5. How many reads aligned to the reference genome
6. Does this consensus sequence pass the quality filter?

One thing to know is that the second letter of the sample name denotes from where this sample came from. In this example dataset, there are 4 possible origins, 'C', 'T', 'D' and 'N'. Consider that in the future there may be more origins.

Task: Write a script that can be run automatically each time such a quality file is generated, and look at how many samples from each origin fail the set quality cut-off (< 95 % covered bases of the reference genome or ‘FALSE’ in column 6).

As this script should run automatically once a week, the script should also serve as a warning system, sending warnings if there are certain origins producing more than 10% failed samples. Therefore, you need to implement a system that notifies its user in some way, telling them the latest results.

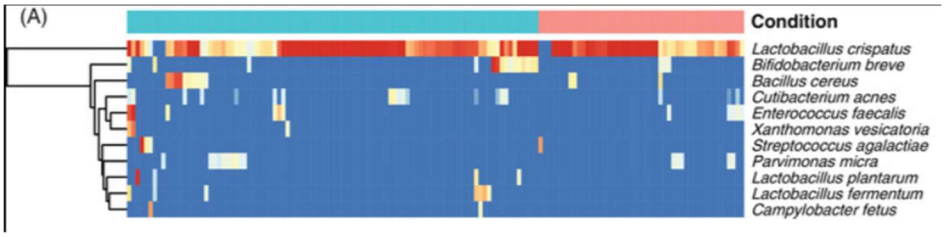
Bioinformatics and Data Centre
Core Facilities
University of Gothenburg

Technical evaluation part 2

You have 30 mins to complete this part of technical evaluation. Provide the answers as a text file and email to marcela.davila@gu.se at the end of the allocated session. Good luck!

1. In case the time given for the practical task was not enough for you to finish, explain as detailed as possible, what is left to do and why was not possible to complete a working version of this script.

2. A research group has sequenced case-control vaginal swabs for both human and microbial RNA using Illumina sequencing.
    a. Describe a workflow to perform a taxonomic analysis including species abundance, microbiome-functional pathways and correlation between human and microbial gene expression. Justify any programs, databases and statistical tests you would use.
    b. Discuss any limitations we should be aware to perform this kind of analysis
    c. This is a heatmap you obtain showing the differentially expressed bacteria (FDR<0.05), what are your comments to the researcher?



3. You are working with 3 consulting projects. One is a differential expression analysis from microarray data. The second is the analysis of TMT data that you got from the Proteomics facility while the third is the identification of a virus integration in some clinical samples. You will be delivering the results sometime during next week, as agreed with the different users. A fourth user got the reviewer's comments and they need to answer back in a week, so they need your help during this week. What would you do?

4. During a first meeting consultation, the analysis that the group is asking you to do is quite new and you haven't performed it before. As usual, they need results in as soon as possible for a grant application. What would you do?

5. After having delivered the final results to a user, you are contacted by the PI saying that the results you delivered do not make sense. You check them and you realize you made a mistake in one of the calculations. What would you do?