Irilenia Nobeli
i.nobeli@bbk.ac.uk

MSc Bioinformatics with Systems Biology
Birkeck, University of London
Sequence Analysis and Genomics, Autumn 2017
Assignment 2

DEADLINE : Friday 19th January 2018 (midnight)
Please submit your answer as a single PDF/Word document on Moodle.

Assignment weight
This assignment counts as 30% of your total mark for the module Sequence Analysis and Genomics. It is thus important that you put enough effort into it to achieve at least a pass mark (15 marks, equivalent to 50%).

Please pay attention to the following points:
- I expect a *maximum* of 2-4 pages of text. You can attach a number of figures/tables in addition to this text but I urge you to be selective and only show things that are interesting and worth including.
- You are free to use whatever tools you think are appropriate for your analysis. There is a variety of web servers and free software available and several of them we have already explored in the practicals.
- When using a software tool or server, you must give all information that will make your work reproducible: name and web address of the tool, version, reference, parameters or any other relevant information.
- Always back up your statements with appropriate references, output from programs or screen snapshots but again be selective about what you include.

Description of the assignment
This assignment is largely exploratory and is meant to give you a chance to carry out a mini research project, using your newly acquired skills.

The aim of the assignment is to provide information on *rpfB*, a gene in the bacterium *Mycobacterium tuberculosis (Mtb).* This gene is also known by its id Rv1009. To facilitate comparison of answers and marking, please use the latest *Mycobacterium tuberculosis H37Rv* when quoting coordinates (preferably genome assembly ASM19595v2 or GCA_000195955.2 from Ensembl).

In your report, I want to see **six** (or more, if you wish!) attempts at obtaining information using skills relevant to the SAG module. Each of these attempts will be awarded a maximum of 5 marks, for a maximum total of 30 marks. I give you below a few ideas for things to try out as well as some additional information that could help you with this task (each one of the points below would count as one piece of evidence from the 6 I'm asking you to collect):

<u>Questions to explore relating to the rpfB gene/protein:</u>
- Conservation of this gene/protein across bacteria (both species that are phylogenetically close to Mtb and others).

Irilenia Nobeli
i.nobeli@bbk.ac.uk

- Information on the function of this protein based on sequence analysis (e.g. using HHblits to identify homologues in other species).
- Information on the likely cellular location of this protein, as evidenced by motifs in the sequence.
- Information on any operon that this gene is part of (including conservation of the operon across closely related species). What is the genomic context of this gene (what other genes surround it and are they likely to be part of the same network?)
- Exploring how good is the current annotation of the start codon (could the protein start earlier or later?) – in bacteria this annotation is not always perfect.

<u>Questions to explore relating to the region immediately upstream of rpfB</u>
The intergenic region between *tatD* (the gene preceding *rpfB*) and *rpfB* is unusually long for a bacterium and this length suggests the presence of at least one and possibly more functional elements in this region.

- Can you identify likely regulatory elements in the region between *tatD* (the gene preceding *rpfB*) and *rpfB*? Examples are the Shine Dalgarno sequence, the -10 and -35 elements etc.
- Can you identify binding sites of transcription factors in the promoter region of this gene?
- Additional information: Experiments probing the transcription start site (TSS) suggest the presence of at least two such sites, one at 1127876 and the other at 1127955 (note that the first TSS overlaps the end of the *tatD* gene – this is not a mistake!). This suggests that an RNA transcript may be expressed ahead of the coding region of the gene. Explore its sequence and structure, assuming that it starts at the first TSS and ends approximately at coordinate 1128002.
- Transcriptomic data suggests the expression of an antisense RNA with coordinates: 1127876:1128036 (on the negative or reverse strand). Explore this transcript and its potential function.

You are encouraged to use the literature to help you in your exploration of the gene and its upstream region, but you will be given marks mostly for work you carry out yourselves rather than information you have pulled out from the literature. The exception is where you have combined information from the literature to come to a novel conclusion (e.g. information obtained for another bacterium helps you come to a conclusion about Mtb).

Of course, you do not need to stick to the list above and you are welcome to try out anything else you may think would give you some additional interesting answers. There are several freely available resources for exploring mycobacteria.
.
Not that much is known about this protein and the regulation of its expression. Hence, you will be awarded as much for the effort you put into it, as for the content of your answer, given that the "correct" answer is not fully known.