Project plan

* What is the aim of your project? What question(s) do you want to answer with your research?

The aim of the project is to assemble the genome of *Leptospirillum ferriphilum* to gain more understanding of it. Furthermore, the aim is to analyse the differential expression between *Leptospirillum ferriphilum* cultured in continuous cultures (with Fe2+) and bioleaching cultures (containing chalcopyrite). The gained knowledge will determine if there are possible new advantageous biomining processes.   
  
I want to answer the questions:  
- What genes are included in the *Leptospirillum ferriphilum* genome?  
- How does the differential expression differ between *Leptospirillum ferriphilum* growing incontinuous cultures (with Fe2+) and bioleaching cultures (containing chalcopyrite)?

* What type of analyses will you perform in order to answer these questions? And in which order? Which softwares will you use? Are there any time bottlenecks? If so, can you identify any analyses that will require longer times?

The analysis I will perform can be seen in fig. 1. In the figure it is possible to see that analyses taking a long time are the genome assembly using canu which will take approximately 11h, aligning reads using BWA which will take approximately 5h and feature counting using HTSeq which will take around 8h. Furthermore, one can see that both the annotation of the genome and the aligning of the clean RNA-reads need to be done before the feature counting can be done.

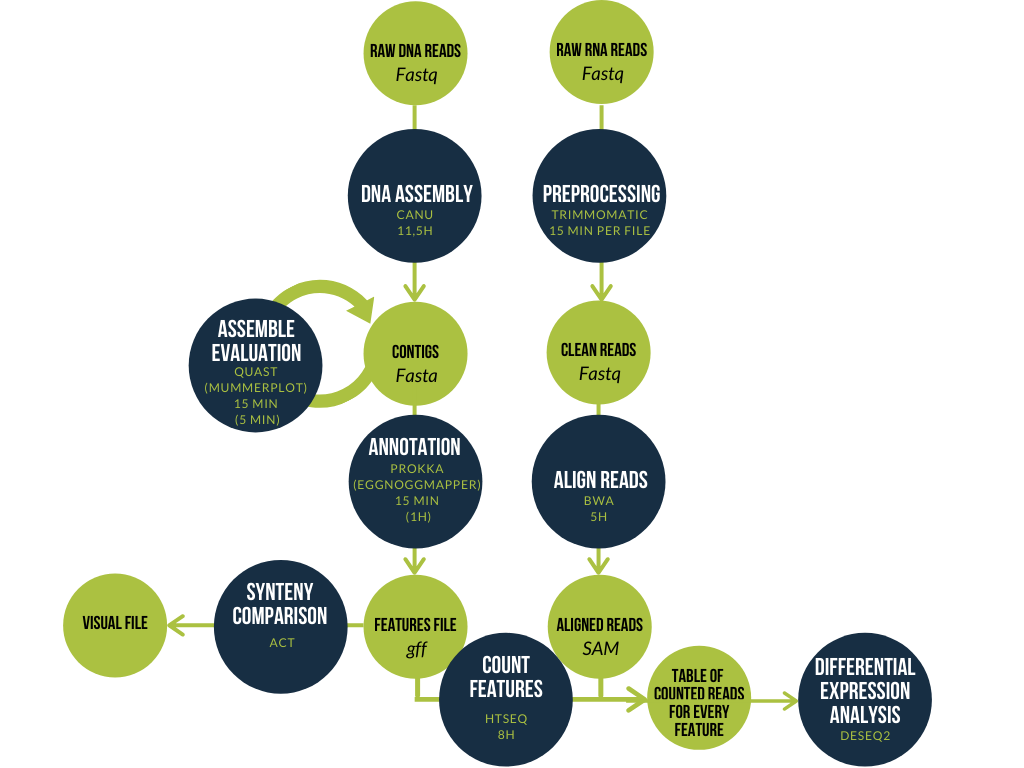


Figure 1. Workflow of the project. Showing different analysis steps with its corresponding software, approximate running time, input and output.

* What is the time frame for your project? Can you define some time checkpoints for when you should have finished certain analyses? When do you need to have finished running all the softwares so you can start to analyze the data?

As it is stated in appendix 3 in the student manual, I should have completed the DNA assembly and annotation before 15/4 (having three 4-hour computer lab sessions), the read alignment and feature counting should be done before 4/5 (having six 4-hour computer lab sessions between the two checkpoints). This means that all of the software-running should be finished before 4/5.

* What types of data will you be handling? How much space do you need in order to store the data? (You might not know that in advance, but pay attention to this as you work and manage your available space!)

The types of data I will be handling can be seen in fig. 1. Exactly how much space I need is something that I do not know right now. However, the 10 raw RNA read-files are approximately 2GB each and the 10 raw DNA read-files are approximately the same size. Meaning that the total space of the raw data is approximately 40 GB (more than what I am possible to store in UPPMAX).

* How will you organize your data? (see “Project Organization”)

I will organize my working directory thinking of the five useful tips from the student manual:

* + Keep the metadata
  + Data and code should be separated
  + Keep data files with unique and **informative** names
  + It is handy to generate folder or file numbers that start with a number
  + (Big) data files should be compressed