The goal of this homework is to familiarise you with some of the aspects of rna-seq, including:

1. Processing of RNA-seq
2. Quality Control in RNA-seq
3. Analysis of RNA-seq

Any of these steps you can do using your computers or a computational cluster provided by us. If you decide to use your own laptops/computers make sure to install the packages that are used in the pipelines from the scripts.

Even if you decide to do this homework using your own computer, you can use all the scripts provided in the folder as a reference on how to use the tools.

If you decide to do this homework using your own computer, please, see file `download_reference.bash` to get the fasta/gtf and other files for mouse reference genome. If you are using our cluster, reference is already downloaded to folder /opt/data/reference.

Using our cluster you might see error/warning like
`/opt/conda/lib/libtinfo.so.6: no version information available`
Just ignore it.

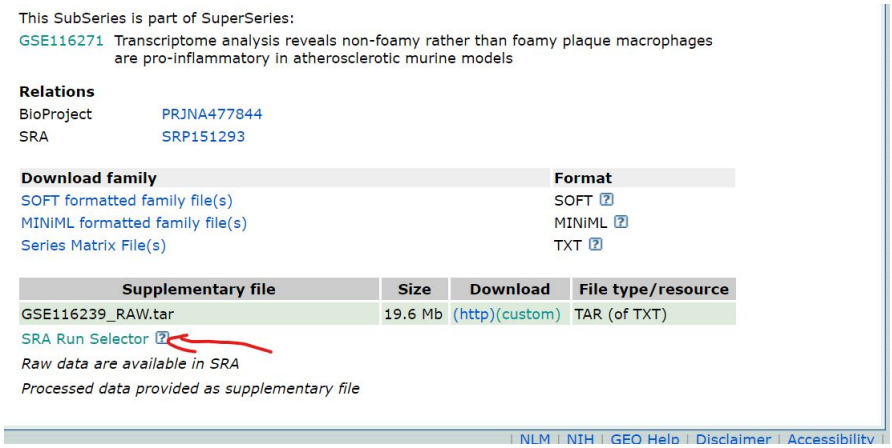
First thing to start, make sure you download the reference genome and create genome indexes for kallisto and Hisat2:

1. Download **index.bash**
2. Put it in the folder "**index**" in your home folder (create the folder, if it's not there)
3. Run `bash index.bash` (this command might take some time to run, please be patient)

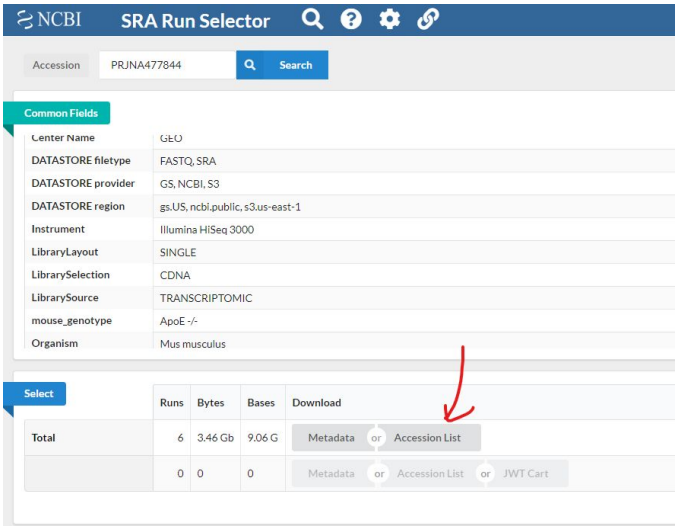
Now you are ready to start running the pipeline, however, we have to know which datasets to process.

1. To do that, find your dataset at GEO omnibus (I will use GSE116239 as an example)

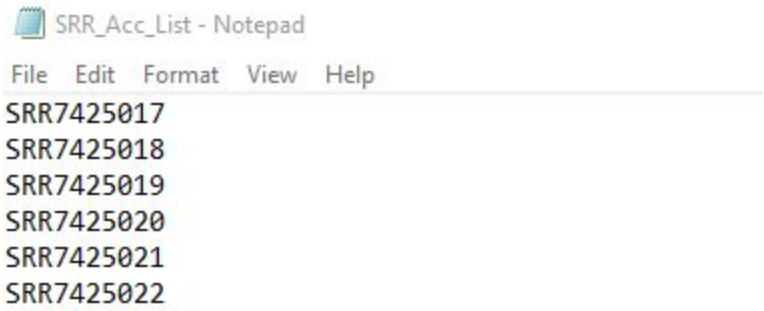
2. press Run SRA Selector link at the bottom of the page.



3. On the newly opened page press accession list button



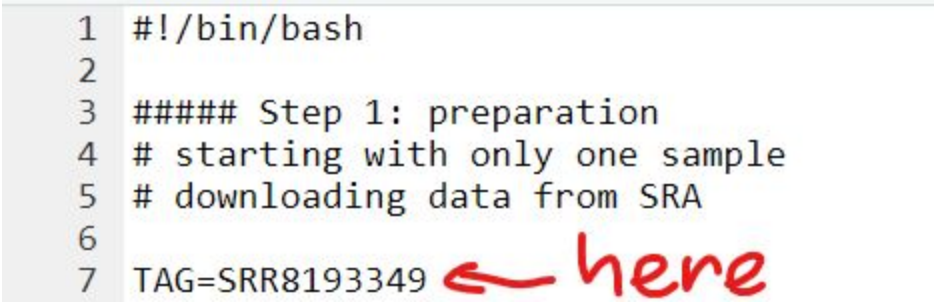
4. These are IDs for your datasets



You might not be familiar with any of the tools needed for analysis of RNA-seq data, but fear not. You might use file **pipeline.bash** which has all of the commands required to run the homework. You might use this file as a reference on how to run any individual tool, or you might use right away as a whole. You might also write your own pipeline, if you are willing to. This pipeline was tested with 4cpus and 8gb RAM, seems to work.

If you want to use this pipeline.bash file, please make sure that you change USERNAME (it appears several times in the script) to your actual username on the cluster (pipeline will resolve absolute paths to the files).

Once you have those IDs you can put them in the file called **pipeline.bash** after TAG=



Once you put your ID, you can save the pipeline.bash and in terminal run
`bash pipeline.bash`

This will take some time to download and analyze the sample
You have to do this with all the samples.

After you think, you have processed the samples, contact me, and we will double-check that by looking at your QC report.

Once you processed all the samples, proceed to DESeq2 analysis:

1. Combine counts into count matrix
2. Run VST for PCA, run PCA
3. Run differential expression
4. Find pathways for DE