LAB3:
A. RNA-seq
B. ChIP-seq
C. Differential expression

Welcome to the third computer exercise - RNA-seq, ChIP-seq and Differential expression. We will use Uppmax/Milou for the RNA-seq part and the differential expression part, and a web interface, Galaxy, for the ChIP-seq part. Detailed instructions below.

Report

You should write the answers to all the below questions in a lab report (one single file). Use the program OpenOffice to write the report ([1] Applications -> Accessories -> OpenOffice; [2] Archive->New->Text Document) . Start by saving the report as a file named lab2_<first_name>_<first_name>.odp (example: lab4_osqulda_osquar.odp). The file type .odp is the default file format of OpenOffice text documents, but you can also choose to save it as a .doc file if you wish. Check where the file is saved. Remember to save the file now and then as you go along. In the report, start all answers with the question number (Q1:, Q2:…).

Contents required: The report should contain answers to questions and the required figures (if any). You should also write the names of the two students that co-authored the submitted report (i.e., yourself and your lab partner).
Instructions for file format: Submit the report as **one single file**, use .doc, .pdf, .odt or .odp format. Please include all your figures (if any) in the .doc/.pdf/.odt/.odp file you submit. Do not submit figures as individual files, make sure to include them in the main report (in OpenOffice, use Infoga->Bildobjekt->Från fil). To save a document as .pdf, use (in OpenOffice) Arkiv->Exportera som PDF.
Instructions for submission: To submit the lab report, **email it to any of the lab assistants**. The lab report should be submitted **the latest one week after the lab takes place**. Again, note that it should all be in **one single file**. Don't forget to write your names on the first page of the report.

_____

You will find the documents and files related to this exercise at KTH Social (for BB2490) or on the Milou file system.

Linux commands and descriptions of commands can be found here:
http://www.computerhope.com/unix/overview.htm
Or you can look at a Linux introduction at the Uppnex web site:
https://www.uppmax.uu.se/linux-introduction-screencast

Use the LAB2 instructions to refresh how to log in to Milou/Uppmax and how to submit jobs using `sbatch`.

PART A.
RNA-seq

Note that the course has a number of reserved nodes for the lab today. The name of the reservation is `g2015056_05`. Please use this reservation when submitting your sbatch scripts, then your jobs will start more or less immediately.

A.1. GET THE DATA

Here we will use RNA-seq data generated by Illumina on samples taken from the mouth of healthy individuals and individuals with periodontitis. The full data set consists of **11 samples** of ~10 M reads each.

To save time, we will run the pipeline on only 2 files with drastically reduced size (100k reads in each). Those files are available here:
```
/proj/g2015056/BB2490/LAB3/rnaseq.healthTiss9F.fastq
/proj/g2015056/BB2490/LAB3/rnaseq.inflTiss9P.fastq
```

Copy these files to your own working directory (e.g., named Lab3) in your `/home/YOUR_USER_NAME` directory.

A.2. MAPPING/ASSEMBLING READS WITH TOPHAT

To run TopHat you need (i) an index of the reference genome, (ii) the fasta file of the reference genome, (iii) your set of RNA-seq reads. Optionally you can also specify (iv) a gene annotation file. There are also (v) certain options you might want to try.

The TopHat manual is available here: http://ccb.jhu.edu/software/tophat/manual.shtml
We will use TopHat version 1.2.0.

(i) Index
TopHat is from the same group that brought us the short read mapper Bowtie (Langmead et al., 2009; see http://bowtie-bio.sourceforge.net/index.shtml) and it uses the reference genome index built by Bowtie, and this index is available at http://bowtie-bio.sourceforge.net/tutorial.shtml. However, it has already been installed on Milou and is available (it comes in 6 files) at
```
/proj/g2015056/BB2490/LAB3/hg19.1.ebwt
/proj/g2015056/BB2490/LAB3/hg19.2.ebwt
/proj/g2015056/BB2490/LAB3/hg19.3.ebwt
/proj/g2015056/BB2490/LAB3/hg19.4.ebwt
/proj/g2015056/BB2490/LAB3/hg19.rev.1.ebwt
/proj/g2015056/BB2490/LAB3/hg19.rev.2.ebwt
```

(ii) Fasta file
And the fasta file is here
```
/proj/g2015056/BB2490/LAB3/hg19.fa
```

(iii) RNA-seq reads
This is one of the files mentioned under A.1.

(iv) Gene annotation file
TopHat takes RNA-seq data as input and finds splice junctions without using a reference gene annotation.
However, it can also accept a gene annotation file as input. The gene annotation file is expected to contain the known transcripts of the organism (in our case human). If an annotation file is provided as input to TopHat, the program will look for junctions between exons in the annotated transcripts.

(v) TopHat options
See full listing at the TopHat manual.
- - -

To start the mapping, you need to construct the `sbatch` script. A template is available:
`/proj/g2015056/BB2490/LAB3/Lab3_sbatch_template.sh`

The first command of the RNA-seq section of the lab has been entered already. Given time constraints, we will only run TopHat on **one** file (and as described above, it is also very reduced in size). Make your pick and enter its name in the sbatch script.

Q1. Easy question: which file did you choose?
Q2. What does the TopHat -o option mean? (You may change the value of the -o option).

Before submitting your sbatch script**, make sure you have added the following modules**: bioinfo-tools, samtools, BEDTools, python, bowtie, tophat/1.2.0, htseq, gcc/4.9.

Now**, use the `sbatch` command to submit your first RNA-seq job**!
The run will take ~30-45 minutes (for one file). Meanwhile proceed to section **A.3**. and/or **PART B (ChIP-seq)**.

When the job has finished, take a look in the results directory. There are two particularly interesting output files:
1. `accepted_hits.sam`. A list of read alignments in SAM format. SAM is a compact short read alignment format that is increasingly being adopted. The formal specification is here.
2. `junctions.bed`. A UCSC BED track of junctions reported by TopHat. Each junction consists of two connected BED blocks, where each block is as long as the maximal overhang of any read spanning the junction. The score is the number of alignments spanning the junction.
(Taken from http://ccb.jhu.edu/software/tophat/manual.shtml).

Actually, you get a `.bam` output file, not a `.sam`. But as you know you can use `samtools view` to convert to `.sam`.

Q3. What would the command line for converting your output `accepted_hits.bam` file to `accepted_hits.sam` look like? You want to keep all the headers (hint: look at the `samtools` documentation).

Perform this conversion (use the `sbatch` command if you deem it necessary).

Q4. Use the `junctions.bed` file to get the top scoring region, what is the score and on which chromosome is this junction? (hint: use the unix `sort` command to sort the file. As always, try `man sort` if you'd like to know more about the command). (Hint 2: the score is in column 5 of this bed file, as per the specifications available at http://genome.ucsc.edu/FAQ/FAQformat.html#format1).

Q5. What does the score mean?

A.3. USING AN ANNOTATION FILE WITH TOPHAT
You might want to use an annotation file with TopHat. Here, you will download your own annotation file from the table browser available at the UCSC Genome Browser.
1. Go to the UCSC Genome Browser web site, http://genome.ucsc.edu/ and click on Table Browser in the left menu.
2. Make sure you have the right genome (human) and genome version (GRCh37/hg19). Choose "group:Genes and Gene Predictions" and "track:RefSeq Genes".
3. Then choose "table:refFlat" and "output format:GTF - gene transfer format".
4. At the "region:" section, make sure the button "genome" is chosen (and not the "position").
5. Type in a suitable file name in the "output file" field and press "get output". The file will be saved.

What did we do here? Well, we extracted all RefSeq genes from the UCSC genome browser.
Q6. Use Google (or any other search engine) to find out more about RefSeq. What is it? Why are the RefSeq genes suitable to use as annotated genes?

Specifially, we extracted these RefSeq genes in the GTF format, http://genome.ucsc.edu/FAQ/FAQformat.html.

Q7. What is column 3 of the GTF format?
Q8. What is the name of the format whose 8 first columns are identical with GTF?

Use the TopHat manual to find out how to include a gene annotation file when running TopHat.

Q9. What would the sbatch line look like if you want to run TopHat with the GTF gene annotation file you downloaded?

You can run TopHat with this file if you feel like it, but it is not required.

Q10. There are other reliable aligners for RNA seq data available. Name one and specify how it differs from Tophat. If you were working with your own RNA seq data, which aligner would you prefer to use? Why?

A.4. USING HTseq

Run HTseq on your TopHat results. HTseq is a package that "provides infrastructure to process data from high-throughput sequencing assays". Pretty much what we want in this course. Specifically, for this computer exercise we will use the script `htseq-count`, which given a SAM file with alignments and a GTF (or GFF) file with genomic features, counts how many reads map to each feature.
You can learn more about HTseq here:
http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html.

The command for running the htseq-count script is:
```
htseq-count -m intersection-strict -t exon -i gene_id accepted_hits.sam
geneannotation.hg19.gtf > my_htseqcount_out.txt
```

Of course, when entering this into your `sbatch` script you should change this command so that you have the correct path to all the files (and the correct file names as well). The output `my_htseqcount_out.txt` can be named in any way you like, and make sure to put it in a suitable place in your directory structure. All options are described on the HTseq web page (see above).

The `geneannotation.hg19.gtf` file is the gene annotation file you downloaded in A.3. You can either use the file you downloaded yourself in A.3, or the file is actually also provided here:
```
/proj/g2015056/BB2490/LAB3/hg19.RefSeq.refFlat.gtf
```

Q11. What does the '-m `intersection-strict`' option mean? The answer is available here: http://www.huber.embl.de/users/anders/HTSeq/doc/count.html#count, but please provide the answer in your own words.

Run HTseq! Then look at the output file and answer the questions below.

Q12. What is the count of the gene named ECM1?
Q13. Use any Unix commands you deem suitable to find out how many genes that have a count larger than zero.

PART B
ChIP-seq

Here you will try the MACS peak caller for ChIP-seq data. As you might recall from the lecture, MACS is a program that takes a set of aligned reads as input, and outputs "peaks", i.e., regions that are enriched for reads in a statistically significant manner (according to the assumptions of the authors) and hence probably represent the transcription factor binding sites (or epigenetic modifications) that was probed for in the immunoprecipitation.

Optionally, MACS can also take a control sample data set as input, the most commonly used control sample is Input DNA, which we will use for the second part of this ChIP-seq exercise.

General information about MACS is available here: http://liulab.dfci.harvard.edu/MACS/; use the "Readme" link in top right corner to get information about how to run MACS and what options it has.

In this computer exercise, we have Illumina sequence reads of length 50 that come from a human cell line, and where the antibody used in the ChIP experiment targets the transcription factor **Sp1** (Specificity protein 1), a rather abundant transcription factor. See http://en.wikipedia.org/wiki/Sp1_transcription_factor for more information about this protein.

Since MACS has not been installed at Uppmax, we will use the MACS implementation available at **Galaxy** instead. According to itself, "Galaxy is an open, web-based platform for data intensive biomedical research. /.../ You can perform, reproduce, and share complete analyses."
In other words: *Galaxy is a web interface where you can upload and analyze your massively parallel sequencing data*.

Many different kinds of sequence data analysis can be performed at Galaxy, and many researchers use Galaxy to analyze their data. At times, the load is quite heavy on the web server, so using Uppmax is a better choice under most circumstances, and specifically for the more computing intense operations (it also provides a greater level of control). But since MACS is not available at Uppmax, this is an excellent opportunity to explore Galaxy.

Galaxy is available at: https://usegalaxy.org/ Go there now!

To learn more about how to use Galaxy, check out their introductory page http://wiki.galaxyproject.org/Learn, or their video tutorials available at http://wiki.galaxyproject.org/Learn/Screencasts.

B.1. RUNNING MACS ON A ChIP-SEQ DATA SET *WITHOUT* CONTROL SAMPLE

THE DATA FILE
The Illumina ChIP-seq reads have already been aligned to the human reference genome using BWA, and have been filtered (mapping quality >10, somewhat permissive perhaps; this is the column 5 of the file below) and converted to bed format, which is the standard input format of MACS. The file is available here and it contains 1670249 entries:
`/proj/g2015056/BB2490/LAB3/Sp1_sample.bed`

Copy this file to your local computer. From there, upload it to Galaxy:
Click on "Get Data" in the list to the **left** on the Galaxy web page. Then choose "Upload file" in the expanded list, and upload the Sp1_sample.bed file in the middle field of the web page (choose the file in the Browse field, then press Start). The name of the file will appear in the

History list to the **right**. Once the field surrounding the file name turns green, the upload has completed.

RUN MACS (version 1.0.1)
Once the file upload is complete, click "NGS: Peak Calling" in the list to the left, then choose "MACS" (you can ignore all the "MACS2 …" entries). You will be taken to a web page from which you can start the peak calling.
But before you start, take a look at the files and default values already entered on the page.

First check that your uploaded file is present in the list under "ChIP-Seq Tag File". Then check out the options.

Q14. The "Effective genome size" is by default set to be 2,700,000,000. What does the "Effective genome size" mean? (Use, e.g., the MACS documentation to find out).

The "tag" size corresponds to the length of your sequence reads. It is by default set to 25. You should change this to 50 since that is the read length in our sample. Also, change the p-value cutoff of 10^{-6} ("1e-6").

The other options should not be changed (except the "Experiment Name" which you may change).

To start the peak calling, simply press "Execute" at the bottom of the page.

Running MACS on this rather tiny data set will only take a minute or two. Soon, you'll see two novel files in the **History** list to the right: Investigate the resulting files.

If you, e.g., click on the newly created .bed file (with the MACS results), the field will expand slightly and you'll see a short summary of the file. You can also click various symbols so as to: **download** the file; see more details; **display** the data in the browser; delete the file; or rerun the analysis. It's safe to try any of these (of course, if you delete it, you'll need to run the analysis again).

Q15. How many peaks do you get that meet the p-value cutoff?

The bed file is suitable to upload to UCSC Genome Browser.
There is also an excel file: If you click on the "html report" file in the History list and then on "Display data in the browser" (the eye symbol), you can then download the excel file (the .xls file). This file is suitable to get detailed results for each peak (e.g., p-value, summit, number of tags in peak).

Q16. What is the maximum number of tags present in a peak?

B.2. RUNNING MACS ON A ChIP-SEQ DATA SET *WITH* CONTROL SAMPLE

MACS can also accept a control sample data set. We will use input DNA as control sample. The file is available here (just like the Sp1_sample.bed file, it contains the genomic positions of quality filtered and mapped reads):
`/proj/g2015056/BB2490/LAB3/inputDNA_control.bed`

Upload it to Galaxy.

Then use this file together with the Sp1_sample.bed file above to find the peaks with MACS.

Note that you have to specify the correct "ChIP-Seq Tag File" and the "ChIP-Seq Control File" before submitting your job.
Furthermore, use `pvalue=1e−6` and Tag size=50 again. Then press "Execute".

Investigate the new results.

Q17. The xls file contains an additional column compared with when MACS was run without a control sample, what is in this column?

Q18. How many peaks do you get that meet the p-value cutoff now that you've used Input DNA as a control sample?

THE FDR VALUES
If you look at the FDR values (given in %) you notice that they are quite high. By this inspection you conclude that maybe you need to investigate your results further.

When you click the "html report", the displayed text contains a section "Messages from MACS". This is the output from MACS that is directed to STDERR (in case you had run MACS from your own command line). Look in this section for clues - specifically warning messages - regarding this suspicious FDR behavior.

Q19. Do you find any warning message regarding the FDR? If so, what does it say? Use Google (or similar) to find out what it actually means (it is not readily available in the MACS documentation).

Look at the answer to question #10 at the MACS FAQ page, http://liulab.dfci.harvard.edu/MACS/FAQ.html, to learn a little more about the relationship between MACS p-values and FDR.

Q20. Use your perl/python/awk/bash/etc. programming skills to find out the average width of the peaks in the two peak sets (with and without input DNA, respectively) you have created. What are the widths? Note, that you can also use the "Statistics"-"Summary statistics" part in the list to the left on the Galaxy web page.

PART C
Differential expression with DESeq

In this part you will learn how to use DESeq, an R package, for differential expression analysis. DESeq takes as input count-data such as the output from htseq-count. The file

```
/proj/g2015056/BB2490/LAB3/RIN_BB2490.count
```

contains such results from an experiment on the degradation on RNA (this file is similar to the htseqcount file you produced in PART A, but this file is based on a larger read set and from different samples). The quality/degradation of RNA is often measured by a number called RIN (RNA integrity number). Here you will investigate the difference in expression between three samples of high quality (RIN 10) and three samples of medium/low quality (RIN 6). You will use the DESeq library for R. R, a program to perform statistical calculations and visualizations, is installed at the Milou computing cluster.

1. Add the necessary modules (if you haven't already done so)

```
module load bioinfo-tools
module load R
```

2. Start R and load the DESeq library
Start R (by simply typing "R" at the prompt) and load the DESeq library (see installation instructions at: http://www-huber.embl.de/users/anders/DESeq/, followed by the command "library(DESeq)") in the R console window.

3. Read in the data using the read.table function (type ?read.table for information of the function).

Q21. What do the last five rows in the data contain?

Q22. This gene list contains all the protein coding genes annotated in the human genome. How many are they?

Q23. Which gene has the highest expression in the data? In which sample is that?

Q24. How many genes have zero expression in all samples?

4. Perform the differential expression analysis.
Before you can send the data to DESeq it is good to 'tidy' it up a little bit. Start by renaming the rows of the table with the first column and then remove the first column from the data. Next remove the last five lines from the data. Here is one possible way to do this:

```
N = dim(data)[1]
rownames(data) = data[,1]
data = data[,-1]
data = data[1:(N-5),]
```

In an R script (recommended) or the R console window type the necessary commands, using the script below as a template:

```
conds = c("RIN10","RIN10","RIN10","RIN6","RIN6","RIN6")
cds = newCountDataSet(data,conds)
cds = estimateSizeFactors(cds)
cds = estimateDispersions(cds)
res = nbinomTest(cds,"RIN10","RIN6")
```

Note: The 'data' variable is what you get from the read.table function – you may, of course, have named it something different.

Q25. How many genes are differentially expressed at a FDR level of 5%? All the samples are from the same cell line, why do you think there is a difference in expression between samples of different qualities?

Q26. Plot the expression profile with log(foldChange) on the y-axis and log(baseMean) on the x-axis. Include the plot in your lab report. (Hint: if no plot window shows up, make sure that you logged in using the –X option, *i.e.* `ssh –X USER@milou.uppmax.uu.se`).

END OF LAB3.