LAB3:

1. **RNA-seq**
2. **ChIP-seq**
3. **DD2399 only: Advanced assignment**

2012-02-08, 08:00-12:00 in *Karmosin*.
Welcome to the third computer exercise – RNA-seq and ChIP-seq. We will use Uppmax/Kalkyl for the RNA-seq part, and a web interface, Galaxy, for the ChIP-seq part. Detailed instructions below. There is also an advanced assignment (C.) for students taking the DD2399 course.

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 kristoffer.sahlin@scilifelab.se. The lab report should be submitted **before Thursday 16 february, 2012, at 17:00**. 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.

February 8, 2012

LAB3, BB2490 Analysis of data from high-throughput molecular biology experiments

You will find the documents related to this exercise at KTH Social (for DD2399) or on the Kalkyl 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.uppnex.uu.se/uppnex-book/getting-started/screencast

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

PART A.
RNA-seq

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 (yes, the prefix of these files is "Lab4" although this is lab 3):

```
/proj/g2012009/INBOX/BB2490_Lab3/Lab4_rnaseq.healthTiss9F.fastq
/proj/g2012009/INBOX/BB2490_Lab3/Lab4_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://tophat.cbcb.umd.edu/manual.html
We will use TopHat version 1.2.0.

<u>(i) Index</u>

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 Boletus and is available (it comes in 6 files) at

```
/proj/g2012009/INBOX/BB2490_Lab3/hg19.1.ebwt
/proj/g2012009/INBOX/BB2490_Lab3/hg19.2.ebwt
/proj/g2012009/INBOX/BB2490_Lab3/hg19.3.ebwt
/proj/g2012009/INBOX/BB2490_Lab3/hg19.4.ebwt
/proj/g2012009/INBOX/BB2490_Lab3/hg19.rev.1.ebwt
/proj/g2012009/INBOX/BB2490_Lab3/hg19.rev.2.ebwt
```

(ii) Fasta file

And the fasta file is here

```
/proj/g2012009/INBOX/BB2490_Lab3/hg19.fa
```

(iii) RNA-seq reads

This is the file mentioned under A.1.

(iv) Gene annotation file

TopHat finds splice junctions without 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/g2012009/INBOX/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.

Now**, use the** `sbatch` **command to submit your first RNA-seq job**!
The run will take ~20 minutes (for one file). Meanwhile proceed to section A.3.

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://tophat.cbcb.umd.edu/manual.html).

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).

Enter this command into your `sbatch` script and perform the conversion.

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 Tables in the top list (white text on blue).
2. Make sure you have the right genome (human) and genome version (GRCh37/hg19). Choose "group:genes and gene predictions track" 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 choosen (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?

Specifically, 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.

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/g2012009/INBOX/BB2490_Lab3/hg19.RefSeq.refFlat.gtf`

Q10. 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.

Q11. What is the count of the gene named ECM1?
Q12. 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 sequencing data analysis can be performed at Galaxy, and a lot of researchers use Galaxy to analyze their data. At times, the load is quite heavy on the web server, so using the 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 today is an excellent opportunity to explore Galaxy.

Galaxy is available at: http://main.g2.bx.psu.edu/. Go there now!

To learn more about how to use Galaxy, check out their introductory page http://wiki.g2.bx.psu.edu/Learn, or their video tutorials available at http://wiki.g2.bx.psu.edu/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/g2012009/INBOX/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 middle panel of the page, and upload the Sp1_sample.bed file. 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

Once the file upload is complete, click "NGS: Peak Calling" in the left list, then choose "MACS". 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.

Q13. 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 read. 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 .bed file, 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).

Q14. 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).

Q15. 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/g2012009/INBOX/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.

Q16. The xls file contains an additional column compared with when MACS was run without a control sample, what is in this column?

Q17. 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.

Q18. 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.

Q19. Use your perl/python/awk/bash/etc. programming skills to find out the average width of the peaks in the two peak sets (with or without input DNA) you have created. What are the widths? Note, that you can also use the "Statistics"-"Summary statistics" part in the list to the left.

END OF LAB.

PART C: Advanced assignments for DD2399

For those of you taking DD2399, the grade is set based on the number of "advanced assignments" you solve. For the transcriptomics part of the course, there is one assignment: **transcript assembly**.

You find the assignment on the web page

http://www.csc.kth.se/dd2399/omsys12/labs

Due date: Thursday Feb 16